Published online 20 January 2020

Nucleic Acids Research, 2020, Vol. 48, No. 4 ***e23***
doi: 10.1093/nar/gkz1209

MetaOmGraph: a workbench for interactive exploratory data analysis of large expression datasets

Urminder Singh [1,2,3], Manhoi Hur [2], Karin Dorman[1,3,4] and Eve Syrkin Wurtele[1,2,3,*]

[1]Bioinformatics and Computational Biology Program, Iowa State University, Ames, IA 50011, USA, [2]Center for Metabolic Biology, Iowa State University, Ames, IA 50011, USA, [3]Department of Genetics Development and Cell Biology, Iowa State University, Ames, IA 50011, USA and [4]Department of Statistics, Iowa State University, Ames, IA 50011, USA

Received October 28, 2019; Revised December 05, 2019; Editorial Decision December 15, 2019; Accepted December 17, 2019

ABSTRACT

The diverse and growing omics data in public domains provide researchers with tremendous opportunity to extract hidden, yet undiscovered, knowledge. However, the vast majority of archived data remain unused. Here, we present MetaOmGraph (MOG), a free, open-source, standalone software for exploratory analysis of massive datasets. Researchers, without coding, can interactively visualize and evaluate data in the context of its metadata, honing-in on groups of samples or genes based on attributes such as expression values, statistical associations, metadata terms and ontology annotations. Interaction with data is easy via interactive visualizations such as line charts, box plots, scatter plots, histograms and volcano plots. Statistical analyses include co-expression analysis, differential expression analysis and differential correlation analysis, with significance tests. Researchers can send data subsets to R for additional analyses. Multithreading and indexing enable efficient big data analysis. A researcher can create new MOG projects from any numerical data; or explore an existing MOG project. MOG projects, with history of explorations, can be saved and shared. We illustrate MOG by case studies of large curated datasets from human cancer RNA-Seq, where we identify novel putative biomarker genes in different tumors, and microarray and metabolomics data from *Arabidopsis thaliana*. MOG executable and code: http://metnetweb.gdcb.iastate.edu/ and https://github.com/urmi-21/MetaOmGraph/.

INTRODUCTION

Petabytes of raw and processed data generated with microarray, RNA-Seq (bulk and single-cell) and mass spectrometry for small molecules and proteins are available through public data repositories (1–4). These data represent multiple species, organs, genotypes and conditions; some are the results of groundbreaking research. Buried in these data are biological relationships among molecules that have not yet been explored. Integrative analysis of data from the multiple studies representing diverse biological conditions is the key to fully exploit these vast data resources for scientific discovery (5,6). Such analysis allows efficient reuse and recycling of these available data and metadata (1,5,7). Higher statistical power can be attained with bigger datasets, and the wide variety of biological conditions can reveal the complex regulatory structure of genes. Yet, despite the availability of such vast data resources, most bioinformatic studies use only a limited amount of the available data.

A common goal of analyzing omics data is to infer functional roles of particular features (genes, proteins, metabolites or other biomolecules) by investigating co-expression and differential expression patterns. A wide variety of R-based (8) tools can provide specific analyses (9–12). Such tools are based upon rigorous statistical frameworks and produce accurate results when the model assumptions hold. Several tools avoid the need to code by providing 'shiny' interfaces (13) to various subsets of R's functionalities (14–20). Such R tools based on the 'shiny' interface have the general limitations that they are not well suited for very large datasets and can have limited interactivity.

Increasing the usability of the vast data resources by enabling efficient exploratory analysis would provide a tremendous opportunity to probe the expression of transcripts, genes, proteins, metabolites and other features across a variety of different conditions. Such exploration can generate novel hypotheses for experimentation, and hence improving the fundamental understanding of the function of genes, proteins and their roles in complex biological networks (6–7,21–25).

At present, there are very limited options for researchers to interact with expression datasets using the fundamental principles of exploratory data analysis (26). Exploratory

*To whom correspondence should be addressed. Tel: +1 515 708 3232; Email: mash@iastate.edu

data analysis is a technique to gain insight into a dataset, often using graphical methods which can reveal complex associations, patterns or anomalies within data at different resolutions. By adding interactivity for visualizations and statistical analyses, researchers with little or no programming experience are able to directly explore the underlying, often complex and multidimensional, data themselves. Researchers in diverse domains (e.g. experts in Parkinson's disease, malaria or nitrogen metabolism) can mine and re-mine the same data, extracting information and deriving testable hypotheses pertinent to their particular areas of expertise. These hypotheses can inform the design of new laboratory experiments. Being able to explore and interact with data becomes even more critical as datasets become larger. The information content inherent in the vast stores of public data is enormous. Due to the sheer size and complexity of such big data, there is a pressing requirement for effective interactive analysis and visualization tools (27,28).

In this paper, we present MetaOmGraph (MOG), a Java software, to interactively explore and visualize large expression datasets. MOG overcomes the challenges posed by the size and complexity of big datasets by efficient handling of the data files. Further, by incorporating metadata, MOG adds extra dimensions to the analyses and provides flexibility in data exploration. At any stage of the analysis, a researcher can save her/his progress. Saved MOG projects can be shared, reused and included in publications. MOG is user-centered software, designed for exploring diverse types of numerical data and their metadata, but specialized for expression data.

MATERIALS AND METHODS

Overview

MOG is an interactive software that can run on any operating system capable of running Java (Linux, Mac and Windows). MOG's Graphical User Interface (GUI) is the central component through which all the functionality is accessed (Figure 1). Access to MOG is easy. MOG is a standalone program and runs on the researcher's computer; thus, the researcher does not need to rely on internet accessibility for computations, and is not slowed down by the data transfer latency. Furthermore, the data in a researcher's project is secure, remaining on the researcher's computer, particularly important for confidential data such as human RNA-Seq.

Interactive data exploration. MOG displays all the data in interactive tables and trees, providing a flexible and structured view of the data. The user can interactively filter or select data for analysis. This ability is particularly important for aggregated datasets, as users may wish to split data into groups of studies, treatments or organs. A novel aspect of MOG is its capability of producing *interactive* visualizations. The researcher can visualize data via line charts, histograms, box plots, volcano plots, scatter plots and bar charts, each of which is programmed to allow real-time interaction with the data and the metadata. Users can group, sort, filter, change colors and shapes, zoom and pan interactively, via the GUI. At any point in the exploration, the researcher can look-up external databases: GeneCards (29), Ensembl (30), EnsemblPlants (31), RefSeq (32), TAIR (33) and ATGeneSearch (http://metnetweb.gdcb.iastate.edu/MetNet_atGeneSearch.htm) for additional information about the genomic features in the dataset. Researchers can also easily access SRA and GEO databases using the accessions present in the study metadata.

Efficient, multithreaded and robust. A key advantage of MOG is its minimal memory usage, enabling datasets to be analyzed that are too large for other available tools. Researchers with a laptop/desktop computer can easily run MOG with data files containing thousands of samples and fifty thousands of transcripts. MOG achieves computational efficiency via two complementary approaches. First, MOG indexes the data file, rather than storing the whole data in main memory. This enables MOG to work with very large files using a minimal amount of memory. Second, MOG speeds up the computations using multithreading, optimizing the use of multi-core processors. MOG is robust and can cope with most of the errors and exceptions (such as missing values or forbidden characters) that can occur when handling diverse data types. Bug reports can be submitted with a single click, if encountered.

Data-type agnostic. Although specifically created for the analysis of omics data, which is the focus of this paper, MOG is designed to be flexible enough to generally handle numerical data. A user can supplement a MOG project with any type of metadata about the features, and about the studies. Thus, a MOG user can interactively analyze and visualize voluminous data on any topic. For example, a user could create a project on: transmission of mosquito-borne infectious diseases world-wide; public tax return data for world leaders over the past 40 years; daily sales at Dimo's Pizza over 5 years; player statistics across all Women's National Basketball Association (WNBA) teams; climate history and projections since 1900.

Leverage of third party Java libraries. In addition to the functionality we have programmed into MOG, MOG borrows some functionality from freely available and extensively tested third-party Java libraries (JFreeChart, Apache Commons Math, Nitrite and JDOM). We have combined these to create a highly modular system that is amendable to changes and extensions and developers can easily implement new statistical analyses and visualizations in the future. MOG is an open source project and we plan to expand and develop it further through community driven efforts. Information on how to contribute to MOG, and who to contact with further questions, is provided at https://github.com/urmi-21/MetaOmGraph/blob/master/CONTRIBUTING.md.

Interface to R. Based on the utility and popularity of R for data analysis, we have implemented a GUI to facilitate execution of R scripts through MOG. MOG's GUI enables a user to interactively select or filter data using MOG; these data are then passed to R. This avoids the need to constantly write new R code to specify different genes and samples for analyses. For example, a user can write an R script for hierarchical clustering of genes based

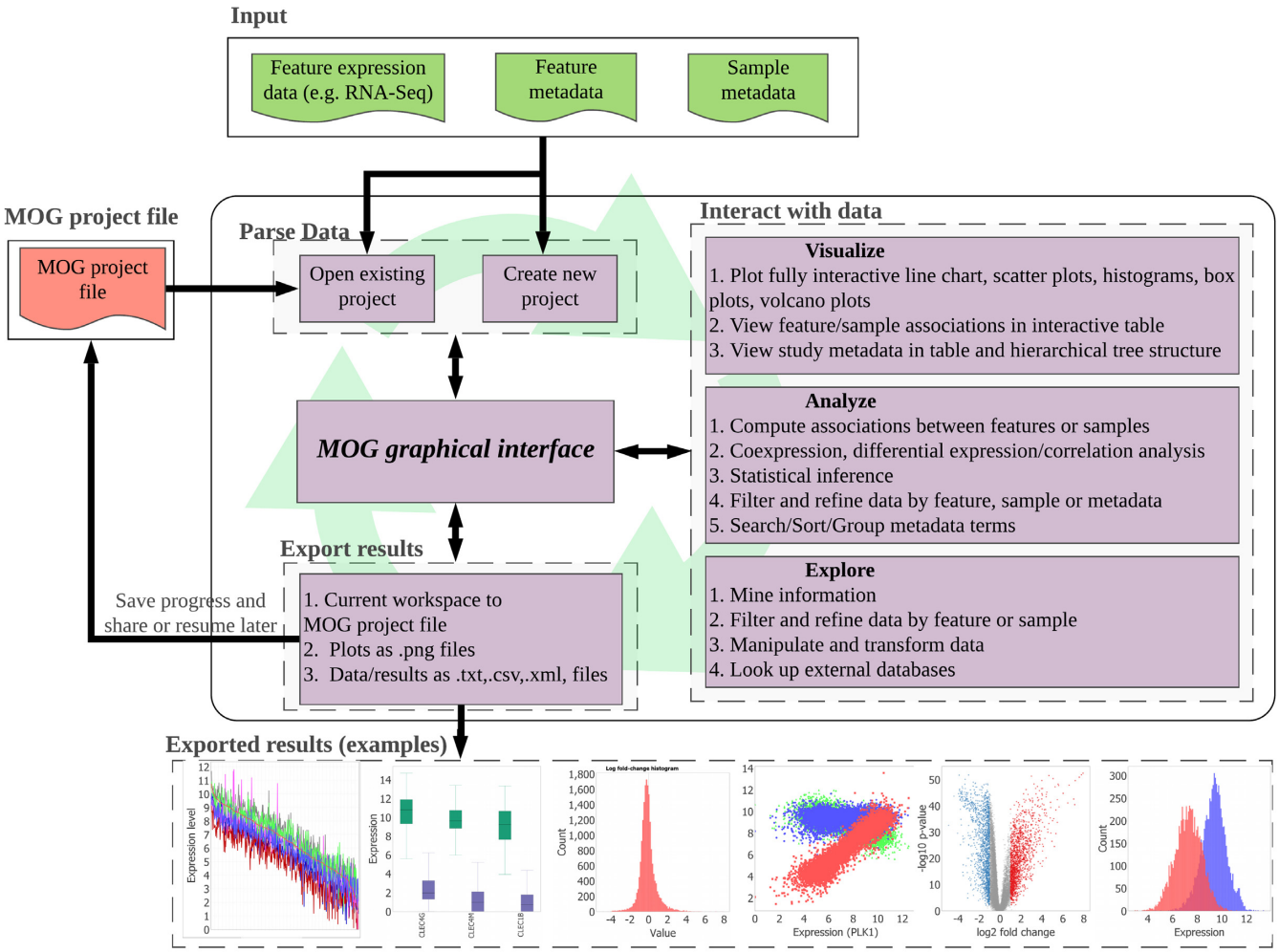

Figure 1. An overview of MOG's modules. All functionality is accessed through MOG's GUI. First, the researcher selects an existing MOG project or creates a new MOG project (.mog) with input data files. Once the project is open in MOG, the workflow is non-linear. The GUI enables interactive exploration of data through a choice of statistical analyses and data visualizations. The researcher can export visualizations and results throughout the analysis, and can save her/his feature lists and statistical analyses in the MOG project file for future exploration. Saved MOG projects can be shared and further analyzed by new researchers.

on the expression levels, interactively select or filter data using MOG, and execute the R script. More details on how to use MOG for executing R scripts are provided in the user manual available from (https://github.com/urmi-21/MetaOmGraph/tree/master/manual).

Creating a new MOG project or using an existing one

A user can quickly create a new MOG project using two delimited files: (i) a file with unique identifiers (IDs) for each feature (e.g. gene), metadata about that feature and numerical data quantifying each feature across multiple conditions (e.g. multiple samples and studies), and (ii) a file containing unique identifiers for each sample and metadata about the samples and studies in the datafile. These are virtually combined by MOG, using the unique identifier in each file (Supplementary Figure S1). Selecting appropriate methods for data normalization, batch correction and vetting are important considerations for a user when creating a new project (Supplementary File 1).

New MOG projects, as well as those from well-vetted datasets, including the human and *Arabidopsis thaliana* datasets described herein, can be re-opened, analyzed, modified or shared. Ongoing exploration results, such as correlations, lists and other interactive analyses, can be saved in any MOG project, regardless of whether it was obtained from our website or created from custom data.

Detecting statistical association within data

Measures of statistical association between a pair of features in a dataset quantify the similarity in their expression patterns across the samples that comprise that dataset (34–37). Genes with significant statistical association may participate in common biological processes and pathways (23,34,38). Genes with significant association only under specific conditions may reveal their functional significance under those conditions (39,40).

MOG provides the researcher with several statistical measures to estimate associations/co-expression among the

features. It can also compute association between samples, which reflects similarity between the samples. Choosing appropriate statistical measures and interpretations for each dataset is left to the user.

Correlation, mutual information and relatedness. We have incorporated four key methods that measure association among pairs of features. Each has its own advantages and disadvantages, depending on the types of relationships the researcher wishes to detect, and the characteristics of the dataset being explored.

MOG can compute pairwise Pearson and Spearman correlation for pairs of selected features across all samples or conversely, between selected samples across all features. The Pearson correlation coefficient measures the extent of a linear relationship between two random variables, X and Y, whereas, the Spearman correlation coefficient measures monotonic relationships between the two variables. Both excel at detecting linear relationships, however, Spearman is less sensitive to outliers (41). Pearson and Spearman correlations are often used to find co-expressed genes and generate matrices used for inferring gene expression networks (23,35,39).

MOG also computes pairwise mutual information (MI) between selected features across samples. MI quantifies the amount of information shared between two random variables. Let (X, Y) be a pair of discrete random variables over the space $\mathcal{X} \times \mathcal{Y}$. Then, the MI for X and Y is defined as:

$$I(X; Y) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right),$$

where, $p(x, y)$ is the joint probability mass function of X and Y, and $p(x)$ and $p(y)$ are the marginal probability mass functions for X and Y, respectively. Compared to correlation measures, MI is a more general approach that can detect complex, non-linear associations. The interpretation of the MI value is different than that of correlation values: an MI value of zero, $I(X; Y) = 0$, implies statistical independence of X and Y, whereas a correlation value of zero need not imply statistical independence (41). MI has been applied to detect non-linear associations in gene expression datasets (25,42–44). MOG computes MI using B-splines density estimation, as described in Daub *et al.* (42).

MOG can also determine the context likelihood of relatedness (CLR) (37). CLR determinations aim to identify biologically relevant associations by discounting features (e.g. genes) that have promiscuous associations. Specifically, the CLR compares the MI value between each pair of features to the background distribution of MI values that include either of these features (37).

Meta-analysis of correlation coefficients. MOG can perform meta-analysis of Pearson correlations. Studies using microarray data showed that meta-analysis and analysis of pooled normalized samples each bring out meaningful, but different, relationships among genes (24). For meta-analysis of correlation coefficients, MOG calculates a weighted average of the individual Pearson correlation coefficients computed from each study. The weights are proportional to the sample size, i.e. correlations estimated from larger studies are more trusted (45,46). Meta-analysis can be useful when multiple studies run a similar experiment (e.g. effect of heat-stress on *A. thaliana*), but may not control ancillary sources of variation (e.g. coverage variation in RNA-Seq data). MOG provides a choice between a fixed effects model (FEM) or a random effects model (REM) (45,46) for the meta-analysis. The FEM combines the estimated effects by assuming that all studies probe the same correlation in the same population, i.e. studies are homogeneous. In contrast, the REM allows studies to be heterogeneous, with additional, uncontrolled sources of variation (45,46). The FEM does not account for all heterogeneities, thus the researcher should choose a model and interpret the results with appropriate caution.

Differential expression between groups

Determining differentially expressed features from aggregated datasets provides direction for further data exploration. In MOG, we have incorporated several popular statistical methods to evaluate differential expression between two groups of samples. For analysis of groups with independent samples, we have implemented: Mann–Whitney U test (a non-parametric test that makes no assumptions about data distribution); Student's *t*-test (assumes equal variance and normally distributed data); Welch's *t*-test (does not assume equal variance, assumes a normal distribution of data); and a permutation test (makes no assumptions about data distribution; computes null distribution empirically using the data). For analysis of groups with paired samples, we have implemented: a Paired *t*-test (assumes normal distribution of data); a Wilcoxon signed-rank test (a non-parametric test; no assumption of data distribution); and a permutation test for paired samples (makes no assumptions about data distribution but computes null distribution empirically using the data).

MOG's methods to identify differentially expressed genes are general statistical methods which are designed for large sample sizes (30 or more samples for gene expression data). Computation of these methods via MOG permits interactivity, which promotes data exploration. A limitation of the interactive differential expression analysis methods implemented in MOG is that they are designed for large sample sizes and use normalized data as input. For smaller sample sizes, a user can apply specialized model-based methods, accessible through R, to infer differentially expressed genes in RNA-Seq or microarray datasets. For example, methods like edgeR (12), DESeq2 (11) and limma (10) require raw counts as input and can provide more reliable differential expression analysis (47) for smaller sample sizes. Tools like ideal (20) and DEBrowser (19) provide interactive interface for accessing these popular differential expression analysis methods (10–12).

Differential correlation between groups

Features whose correlation with other features is significantly different only under particular environmental, genetic or developmental conditions are designated as differentially correlated. Such *shifting* biological interactions among these genes or their regulators (40,48) reflect the context-dependency of gene expression.

MOG can find the features whose Pearson correlation to a user-selected feature differs significantly between two groups of samples. To do this, MOG applies a Fisher transformation (49) and performs a hypothesis test for equality of Pearson correlation coefficients from the two groups. (The difference of the two Fisher transformed Pearson correlation coefficients follows a normal distribution (40)). The researcher can choose to conduct a test for statistical significance on the Fisher transformed Pearson correlation coefficients or on the raw Pearson correlation coefficients.

Statistical significance determinations

For each statistical test, MOG provides a non-parametric option (a permutation test) and parametric options (calculations under distributional assumptions) to estimate *P*-values.

Empirical *P*-values are calculated by a permutation test that estimates the null distribution of a test statistic by randomly permuting the labels of the observed data points (assuming that the labels are exchangeable under the null hypothesis) (50). Because permutation tests do not rely on any data distribution, they are applicable even if parametric assumptions are not met. More permutations yield more precise estimates of the null distribution and *P*-values, but at the cost of longer computation times. MOG accelerates computation of permutation tests by multithreading, and processing the permuted datasets in parallel (Supplementary File 1).

MOG provides three popular parametric methods to adjust the *P*-values for multiple comparisons: the Bonferroni method (51), the Holm method (52) and the Benjamini–Hochberg (BH) method (53). Bonferroni and Holm methods are applied to control the family-wise error rate (FWER), whereas the BH method controls the false discovery rate (FDR). Controlling the FWER limits the total number of false positives; the Holm method is less conservative as compared to the Bonferroni method. In contrast, controlling the FDR controls the proportion of false positives among the significant tests.

Datasets

To create case-studies with MOG, we assembled MOG projects based on three technical platforms.

Human cancer RNA-Seq dataset (7142 samples). We created a new MOG project based on the well-vetted dataset from Wang *et al.*, (21). This dataset combines RNA-Seq data from The Cancer Genome Atlas (TCGA, tumor and non-tumor samples) (https://cancergenome.nih.gov/) and Genotype Tissue Expression (GTEX, non-tumor samples) (54).

To create the MOG project, we excluded from the dataset any organ types in which the number of tumor or non-tumor samples was <30. To ensure statistical independence among the samples, we removed all non-tumor samples from TCGA and included only one TCGA tumor-sample per patient (Supplementary File 2). We also excluded an outlier sample with very low expression values, based on a preliminary exploration of sample replicates using MOG (See 'Results' section).

We then compiled metadata for the studies/samples and for the genes and integrated this metadata into the dataset. We downloaded the study and sample metadata (TCGA metadata from TCGAbiolinks (55); GTEX metadata from GTEX's website (https://gtexportal.org/home)). We were unable to locate metadata for 17 of the TCGA samples and excluded these samples from the dataset (Supplementary File 2). We extracted metadata about the genes from the HGNC (https://www.genenames.org/), NCBI Gene (https://www.ncbi.nlm.nih.gov/gene), Ensembl (30), Cancer Gene Census (56) and OMIM (57) databases and added these information to the gene metadata in our dataset. We also eliminated the 1870 genes that were not reported for all studies resulting in a dataset, called herein, '*Hu-cancer-RNASeq-dataset*'.

We generated the MOG project (*Hu-Cancer-18212-7412-RNASeq.mog*) from the *Hu-cancer-RNASeq-dataset* and its metadata. The MOG project contains expression values for 18 212 genes, 30 fields of metadata detailing each gene, across 7142 samples representing 14 different cancer types and associated non-tumor tissues (Table 1); it also has 23 fields of metadata describing each study and sample in the dataset. We used MOG to log_2 transform the data for subsequent analyses.

A. thaliana *microarray dataset (424 samples).* We created a new MOG project, *AT-Affy-22746-424-microarray.mog*, based on the *A. thaliana* curated microarray dataset (*'AT-microarray-dataset'*) from Mentzen and Wurtele (23). This dataset had been compiled using 963 Affymetrix ATH1 chips with 22 746 probes from 70 diverse studies encompassing different conditions of development, stress, genotype and environment. All chips in the dataset were individually normalized and scaled to a common mean using MAS 5.0 algorithm. Only chips with good quality biological replicates were kept and all the biological replicates were averaged to yield 424 samples. At last, median absolute deviation (MAD)-based normalization (58) was applied to the data. We compiled new metadata for the genes from TAIR gene annotations (33) and added phylostratal inferences (59). The sample metadata were obtained from Mentzen and Wurtele (23).

A. thaliana *metabolomics GC-MS dataset (656 samples).* The small molecule composition (metabolomics) data that we used to create a MOG project were from 656 GC-MS samples describing the effect of 50 knock out (or knock down) mutations of genes of mostly unknown functions on the accumulation of metabolites in *A. thaliana* (60) (called herein, *'AT-metab-dataset'*). We downloaded these data from the Plant/Eukaryotic and Microbial Resource (PMR) (61). We created the MOG project *AT-Mutation-242-656-metab.mog* with this dataset.

RESULTS

We illustrate MOG's usability and flexibility by exploring three diverse datasets from different perspectives. The statistical analyses and visualizations shown were generated exclusively using MOG. Often, our exploration led us to conclusions consistent with prior experimental or *in silico*

Table 1. Tumor and non-tumor samples in the *Hu-cancer-RNASeq-dataset* and the number of *up*regulated and *down*regulated genes in each tumor type with respect to the corresponding normal samples, as calculated by MOG

TCGA disease	GTEX organ	#TCGA samples	#GTEX samples	Total	#Up	#Down
Breast invasive carcinoma (BRCA)	Breast	965	89	1054	1093	2827
Colon adenocarcinoma (COAD)	Colon	277	339	616	1401	3036
Esophageal carcinoma (ESCA)	Esophagus	182	659	841	1989	2229
Kidney Chromophobe (KICH)	Kidney	60	32	92	986	4214
Kidney renal clear cell carcinoma (KIRC)	Kidney	470	32	502	1877	2263
Kidney renal papillary cell carcinoma (KIRP)	Kidney	236	32	268	1152	2737
Liver hepatocellular carcinoma (LIHC)	Liver	295	115	410	1527	1485
Lung adenocarcinoma (LUAD)	Lung	491	313	804	1361	2753
Lung squamous cell carcinoma (LUSC)	Lung	486	313	799	2210	3734
Prostate adenocarcinoma (PRAD)	Prostate	426	106	532	577	1633
Stomach adenocarcinoma (STAD)	Stomach	380	192	572	1527	1631
Thyroid carcinoma (THCA)	Thyroid	441	318	759	993	1525
Uterine Corpus Endometrial Carcinoma (UCEC)	Uterus	141	82	223	2135	3250
Uterine Carcinosarcoma (UCS)	Uterus	47	82	129	2419	2491

results. In other cases, the exploration led us to completely novel predictions that could be tested experimentally.

Preliminary exploration of the *Hu-cancer-RNASeq-dataset*

Determining that a dataset is valid, properly normalized and free of batch effects is a critical preliminary step in the analysis. To verify that samples from similar biological conditions exhibit similar expression patterns for all the genes, we used MOG to compute pairwise Pearson correlations among samples from the same biological condition (tumor/non-tumor and organ type). All the samples had high Pearson correlations (>0.70) with others taken from the same organ and tumor status, except one sample from lung adenocarcinoma (LUAD), which we removed from the dataset (Additional File 1).

We visualized the distribution of Pearson correlation values for non-tumor samples. For homogeneous samples, such distributions should appear unimodal. However, several organs show multimodal distributions (Supplementary Figure S2). This finding led us to conjecture that samples might have been taken from different anatomical sites within these organs. By exploring further with MOG, we were able to identify additional metadata on sub-locations in the colon and esophagus that support this conjecture (Supplementary Figure S2). However, the stomach sample metadata does not further specify location (or any other obvious factor, such as gender, race or age) that might distinguish subgroups of samples. Because the stomach samples are of several distinct types, a researcher might want to consider analyzing them as such.

Using MOG to identify a catalog of differentially expressed genes in cancers

We wanted to identify *key genes* that are regulated by, or implicated in, the molecular and cellular processes driving cancer, and to further explore the processes in which these genes are involved. For this task, we used MOG first to identify the differentially expressed genes in samples from each tumor type versus corresponding non-tumor samples, and then to examine the expression patterns of these genes. We define a gene as differentially expressed between two groups if it meets each of the following criteria:

Table 2. MOG identifies 35 genes as differentially expressed in all of the 14 tumor types

Upregulated in each cancer	Downregulated in each cancer
BIRC5, BUB1, CDC45	ADH1B, C7, CHRDL1
CDKN2A, CENPF, DLGAP5	CMTM5, DCN, DES
FAM111B, KIF4A, KIF20A	DPT, GPM6A, GSTM5
MELK, MKI67, PBK	HPD, HSPB6, MRGPRF
PKMYT1, TOP2A, TPX2	NKAPL, PEG3, PI16
UBE2C	PTGDS, SCN7A, TCEAL2
	TGFBR3

(Mann–Whitney U test, $|FC| \geq 2$, BH corrected P-value $<10^{-3}$).

(i) Estimated fold change in expression of 2-fold or more (log fold change, $|logFC| \geq 1$ where $logFC$ is calculated as in limma (10).)
(ii) Mann–Whitney U test, on the $\log_2$ transformed data, is significant between the two groups (BH corrected P-value $<10^{-3}$)

In each type of cancer in the *Hu-cancer-RNASeq-dataset*, between 2000–5000 of the 18 212 genes are differentially expressed (Table 1 and Supplementary File 3). Thirty-five of these genes are consistently differentially expressed in all of the cancers (Table 2).

Several genes that are deeply implicated in cancer are not differentially expressed in any of the tumor types we analyzed. One example is tumor suppressor protein 53 (TP53) (Figure 2 A and B). (TP53 is differentially expressed in colorectal tumors (62); colorectal tumors are not included in the *Hu-cancer-RNAseq-dataset*).

Fifteen of the 16 genes upregulated across all tumor types are co-expressed across the tumor samples, across the non-tumor samples and across the combined tumor plus non-tumor samples (Figure 2 C and Supplementary File 4). Cyclin dependent kinase inhibitor (CDKN2A) is an outlier (Spearman correlation < 0.50) (Figure 2 D and Supplementary File 4). This co-expression might imply that these 15 genes function together as a module in both tumor and non-tumor cells.

In contrast, there is no co-expression cluster among the 19 genes that are downregulated across all cancer types; 62 individual gene pairs are correlated across all the samples

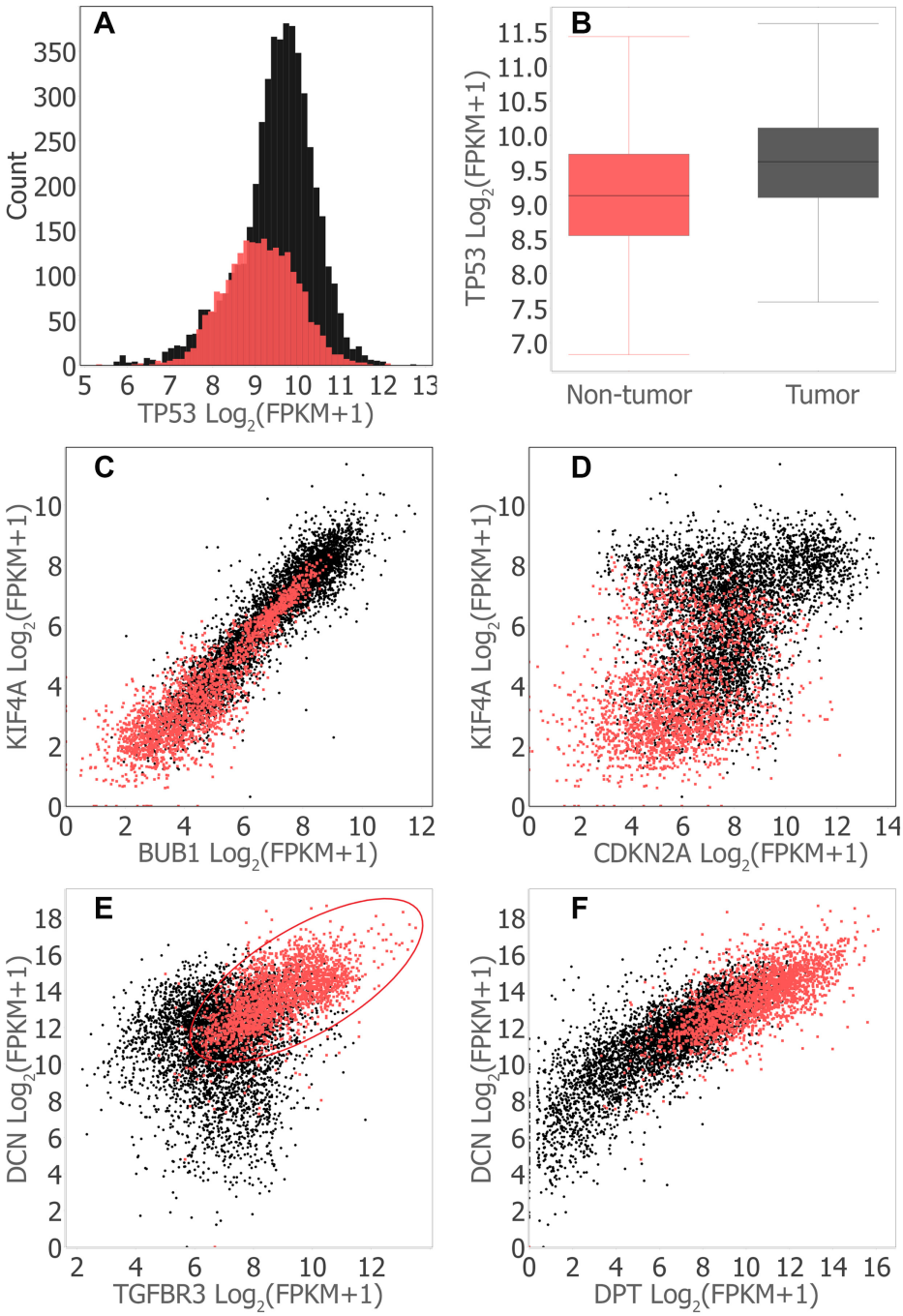

Figure 2. MOG visualizations of expression of selected genes across all tumor types and non-tumor samples. Tumor samples, black dots; non-tumor samples, red dots. Correlations and differential expression analyses were performed using MOG (Mann–Whitney U test, $|FC| \geq 2$, BH corrected P-value $<10^{-3}$). Plots were generated in MOG by interactively splitting the gene expression data into the categories 'tumor' and 'non-tumor' using the sample metadata. (**A**) Histogram showing the distribution of Tumor Protein P53 (TP53) expression (number of bins set to 50). (**B**) Box plot summarizing the expression of TP53 over all tumor versus all non-tumor samples. The horizontal line inside the box represents the median log expression, which is 9.1 for non-tumor samples and 9.6 for tumor samples. (**C**) Scatter plot visualizing co-expression of mitotic checkpoint serine/threonine kinase (BUB1) and kinesin family member 4A (KIF4A) (both are upregulated across all tumor types). (Spearman correlation=0.92 in non-tumor samples; Spearman correlation=0.84 in tumor samples; Spearman correlation=0.94 over both tumor and non-tumor samples). (**D**) Scatter plot visualizing co-expression of genes cyclin dependent kinase inhibitor 2A (CDKN2A) and KIF4A. Both are upregulated across all tumor types, but they are are not co-expressed. (**E**) Scatter plots visualizing transforming growth factor beta receptor3 (TGFBR3), which has a complex role as regulator of angiogenesis (117), decorin (DCN), autophagy, mitophagy and embryonic cell development including endovascular differentiation (118). TGFBR3 and DCN are downregulated across all tumor types and are co-expressed in non-tumor samples (Spearman correlation=0.64) but not in tumor cells (Spearman correlation=0.14). The co-expression of TGFBR3 and DCN in only the non-tumor samples suggests that the processes in which each gene participates are associated under normal conditions; the loss of this association in tumors is consistent with a hypothesis that an imbalance, or factors that cause that imbalance, may further contribute to the etiology of cancer. (**F**) Scatter plot visualizing co-expression of genes dermatopontin (DPT) and DCN Spearman correlations are 0.82 (tumor samples), 0.69 (non-tumor samples) and 0.84 (combined samples) (Both gene are downregulated across all tumor types.)

(Spearman correlation ≥ 0.60) (Supplementary File 4). Expression of seven of these gene pairs is strongly correlated only among tumor samples but is not correlated among non-tumor samples; conversely, 18 gene pairs are strongly correlated among non-tumor samples but not among the tumor samples (e.g. Figure 2E)—this finding indicates a context-dependent coordination of these gene pairs. Four gene pairs are strongly correlated among both tumor and in non-tumor samples (e.g Figure 2F).

Functional analysis of differentially expressed genes. To determine whether the genes that are differentially expressed in cancers are involved in known biological processes, we performed gene ontology (GO) enrichment analysis using GO::TermFinder (63) and Revigo (64) on the genes that are upregulated, downregulated or not significantly changed across all the cancer types. Consistent with the behavior of cancer cells, upregulated genes are significantly enriched in GO terms related to cell proliferation: cell cycle, cell division, organelle organization, regulation of cellular component organization and regulation of cell cycle (Supplementary File 4 and Figure S3). The 5784 genes that did not change expression were enriched in GO terms RNA processing, mRNA metabolic process, nucleic acid metabolic process and gene expression (Supplementary Figure S4 and File 4). The downregulated genes show no significant GO term enrichment.

Using MOG for gene-level exploration

With the aim to use MOG from the vantage point of an individual gene, we selected the glypican 3 (GPC3) gene as an interesting candidate for a case study. GPC3, encoding a glycosylphosphatidylinositol-linked heparan sulfate proteoglycan, is located on the X chromosome and has been implicated as a critical regulator of tissue growth and morphogenesis (65). GPC3 inhibits cell proliferation and hedgehog signaling during embryonic development (66). In tumors, GPC3's role is complex and not well understood. It can promote or inhibit cell growth depending on the cancer type (67,68). Mutations in GPC3 have been linked to Wilms tumor as well as Simpson-Golabi-Behmel syndrome (SGBS) (69,70).

GPC3 Expression patterns. We explored expression patterns of GPC3 with regards to the 14 tumor types. Differential expression of GPC3 in non-tumor versus tumor samples varies by organ. GPC3 expression is 30-fold higher in the LIHC samples than in the non-tumor liver samples, and 8-fold higher in the UCS samples compared to the non-tumor uterus samples (Supplementary File 5). In contrast, GPC3 is downregulated in nine tumor types (BRCA, COAD, ESCA, KIRC, KIRP, LUAD, LUSC, THCA and UCEC) and unchanged in three tumor types (KICH, STAD and PRAD) (Figure 3A and B; Supplementary File 5).

These results are consistent with targeted studies of liver, breast and lung tumors. GPC3 expression is upregulated in liver cancer (67,71–72), and has been suggested as a diagnostic biomarker and as a potential target for cancer immunotherapy in hepatocellular carcinoma (71–73). GPC3 is downregulated in breast (74), lung (75) and ovarian cancers (76), and it may act as a tumor suppressor in lung and renal cancer (76,77).

GPC3 Co-expression patterns. We then investigated co-expression patterns of GPC3 in the tumor and non-tumor tissues from different organs (Additional File 3). GPC3 co-expression patterns differ between tumor and non-tumor samples according to the organ sampled (Figure 3C), moreover, the genes whose expression is correlated with GPC3 are distinct according to organ types, all reflecting the complex role of this gene (Supplementary File 5). For example, 4219 genes are co-expressed with GPC3 in non-tumor esophagus samples, whereas no gene is co-expressed with GPC3 in non-tumor samples from prostate and stomach (Supplementary File 5). Co-expressed genes also differed according to whether disease was present. For seven organs, fewer genes were co-expressed with GPC3 in tumor samples than in non-tumor samples (Supplementary File 5). For example, 192 genes were co-expressed with GPC3 in non-tumor liver samples, whereas no genes were significantly co-expressed with GPC3 in LIHC tumor samples (Figure 3D and E).

We analyzed GO term enrichment for those organs with more than 10 GPC3-co-expressed genes: colon, esophagus, kidney and liver. The term cell adhesion is enriched in GPC3-co-expressed genes from colon, esophagus, kidney and liver. The terms cell development, extracellular matrix organization and multicellular organism development are enriched among GPC3-co-expressed genes in colon, esophagus and kidney. Other GO terms are over-represented in a organ-specific manner (Supplementary File 5).

GPC3-associated clusters in tumor versus non-tumor samples from liver. To further explore potential interactions of GPC3 with other genes, we used MOG to build two gene co-expression networks from the 3012 genes that are differentially expressed in LIHC—one network from non-tumor liver samples, and a second from LIHC samples (Additional File 3). Then, we imported each network into Cytoscape (78) and identified the tightly connected modules using MCODE (79).

In the network built from non-tumor liver samples, MCODE ranked the GPC3-containing cluster second most significant (73 nodes (genes); MCODE score 30.7). GPC3 was directly connected with 21 genes in this cluster (Supplementary Figure S5), which is most enriched in GO terms: sulfur compound catabolic process, aminoglycan catabolic process and extracellular matrix organization (Supplementary Figure S6 and File 5).

In contrast, in the LIHC samples, GPC3 was not significantly co-expressed with any other genes, and thus was absent from the entire LIHC network. However, the LIHC network does contain a module with 114 genes (MCODE score 94.3), 33 of which are in the GPC3-containing cluster identified from the non-tumor network (17 of these genes are directly connected with GPC3 in the non-tumor network) (Supplementary Figure S5). This cluster is enriched in GO terms: extracellular matrix organization, blood vessel development and vasculature development (Supplementary File 5 and Figure S7).

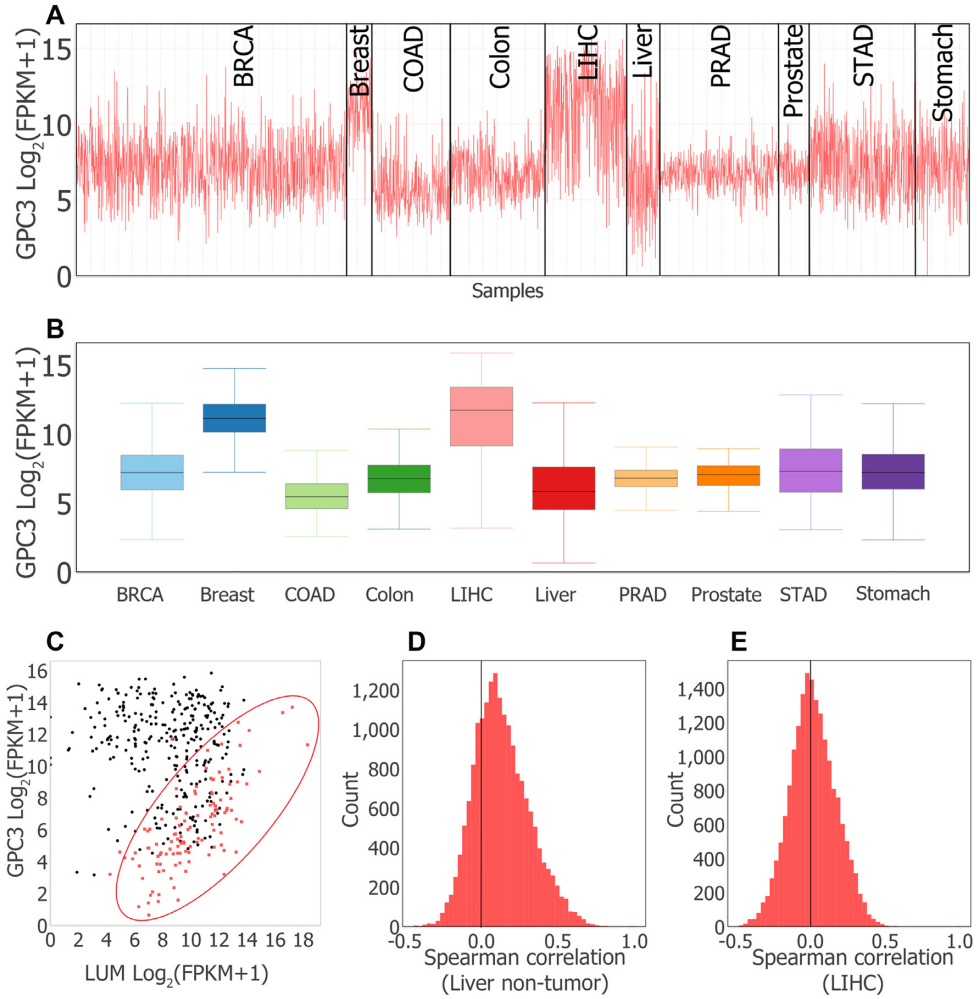

Figure 3. MOG visualizations of glypican 3 (GPC3) expression pattern in tumor and non-tumor organs. (**A**) Line chart generated by interactively filtering by study metadata to retain 3184 samples from 5 tumor types and corresponding non-tumor organs, and grouping the chart by organ/tumor type. (**B**) Box plot summary of data in (A). Generated by interactively splitting box plot according to organ/tumor type. (**C**) Scatter plot showing co-expression of GPC3 and Lumican (LUM) in liver non-tumor and LIHC samples. In non-tumor liver (red), GPC3 and LUM expression are strongly correlated (Spearman correlation ≤ 0.7). In LIHC samples (black), GPC3 and LUM expression show no association (Spearman correlation $= -0.1$). (**D** and **E**) Histograms of distribution of Spearman correlation coefficients of expression of GPC3 with all other genes. Non-tumor liver samples (D), LIHC samples (E). The longer right tail of non-tumor liver samples indicates Spearman correlation coefficients of GPC3 expression with selected genes are higher in non-tumor than LIHC samples.

Stage-wise analysis of *Hu-cancer-RNASeq-dataset*

Identifying new candidate biomarkers for cancers. To identify potential biomarkers for tumors, we used MOG to distinguish genes whose expression is associated with the disease progression. We used MOG to separate samples by organ, and then by early stage (stage I or stage II) and late stage (stage III and later), based on the study metadata. At last, we performed a Mann–Whitney U test on those genes that are upregulated in tumor versus non-tumor samples (Supplementary File 3) to reveal the genes that are upregulated in late stage compared to early stage (expression increase 2-fold or more, and BH corrected P-value < 0.05). These genes have increasing expression with cancer progression. We similarly identified the genes that have a decreasing pattern of expression with cancer progression.

ESCA, KIRP, KIRC THCA all included metadata and had sufficient numbers per stage to detect differentially expressed genes. (Full results in Additional file 4.) MOG reveals 221 genes that increase expression during tumor progression (gene numbers for each tumor type are: ESCA:91, KIRP:89, THCA:25, KIRC:24), and 227 genes that decrease expression (gene numbers for each tumor type: ESCA:89, KIRP:68, LIHC:64, KIRC:13) (Supplementary File 6 and Additional File 4). Of these 448 genes, 122 are flagged as prognostic markers by The Human Protein Atlas (THPA), which identifies prognostic markers by survival analysis (80). For example, Figure 4B and C shows the expression pattern of two such genes, Phosphoenolpyruvate Carboxykinase 1 (PCK1, known to be downregulated in KIRC (81) and general marker of renal failure (82)) and Chromosome 10 Open Reading Frame 99 (C10orf99, a known colon cancer inhibitor (83), and positive marker of KIRC (84)), in KIRC and KIRP.

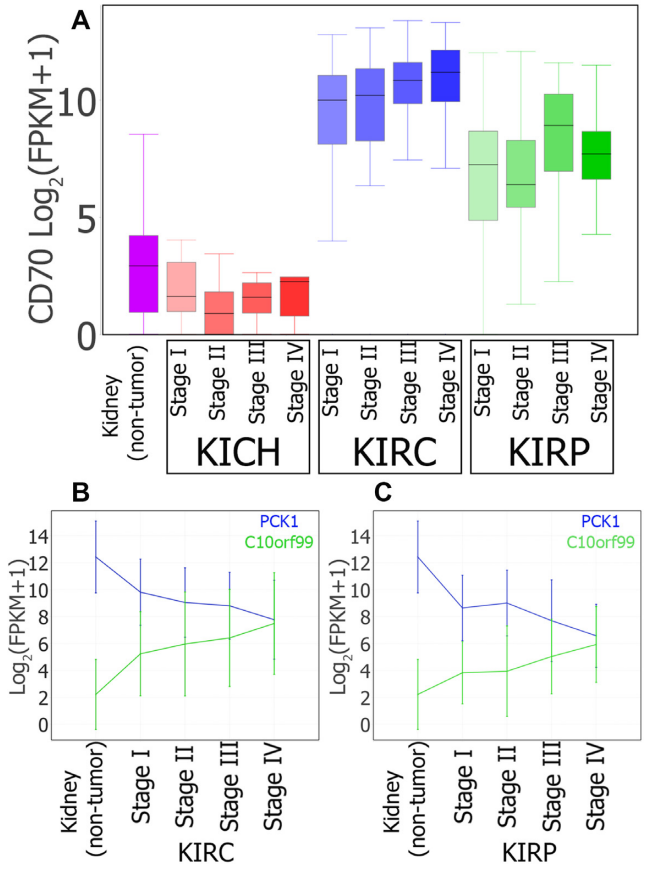

Figure 4. MOG visualization of expression of selected genes during progression of three types of renal cancer. (**A**) Box plots summarizing CD70 expression in non-tumor kidney and in different stages of KICH, KIRC and KIRP cancer progression. CD70 is designated as prognostic unfavorable for renal cancer by THPA (80). However, although CD70 levels in tumor samples increase 93-fold in KIRC and 14-fold in KIRP, CD70 levels *decrease* in KICH by 3-fold ($logFC = -1.56$; B-H corrected P-value $= 0.004$). (**B** and **C**) Line charts showing average expression of PCK1 (blue) and C10orf99 (green) over different stages of KIRC (B) and KIRP (C). The vertical lines are error bars. THPA designates PCK1 as prognostic favorable and C10orf99 as prognostic unfavorable for renal cancer.

Three hundred and twenty-seven genes that were identified in our study as differentially expressed in at least one tumor type were *not* labeled as prognostic in THPA (Supplementary File 6). For example, out of the 111 genes that increase during progression of KIRC or KIRP, only 56 were flagged as unfavorable prognostic for renal cancer by THPA. Of the 79 genes MOG identifies as decreasing with cancer progression in KIRC or KIRP, 39 were labeled as prognostic favorable for renal cancer by THPA. Twenty-seven genes out of the 64 that we identified by MOG as decreasing with cancer progression in LIHC were labeled by THPA as prognostic favorable for liver cancer. Out of 25 genes identified as having increasing pattern in THCA, none were labeled as prognostic by THPA. We propose that these genes may provide new candidates as biomarkers for prognosis of these tumor types (Supplementary File 6).

A number of the 327 genes identified as differentially expressed in MOG but not listed in THPA have been experimentally evaluated for their potential as prognostic markers (Table 3). For example, ARG1, CYP2C8, CYP3A4, CYP3A7 and CYP4A11, which we identified using MOG as decreasing expression with LIHC progression, have each been recently studied as prognostic markers for hepatocellular carcinoma (85–88). MOG analysis provides additional support for use of these genes as biomarkers.

Using MOG to analyze and visualize the results by tumor type can reveal more nuanced information. For example, the Cluster of Differentiation 70 (CD70) gene is flagged by THPA and high CD70 expression is prognostic unfavorable for renal cancer. MOG analysis shows CD70 expression is higher in two types of renal tumors, KIRC and KIRP, and increases with disease progression (Figure 4A), but CD70 levels in another renal tumor type, KICH, have slightly *lower* expression than in non-tumor samples; thus, specifically in the case of KICH, *low* CD70 levels might be an unfavorable prognosis.

For prognosis and personalized medicine (89,90) *exceptions* can be extremely important, because specific tumor sub-types might respond differently to a particular treatment. By using MOG to explore RNA-Seq from large numbers of conditions and organs, a researcher can visualize data for individual samples or groups of that show changed expression of a prognostic marker or sets of markers, and compare these to those that do not.

Such exploration could suggest statistical analyses to try out in other, independent datasets to determine whether subsets of non-canonical samples might have a biologically distinct signature, revealing a different modality for a particular cancer. This in turn could be followed up by targeted experimental approaches or clinical studies.

Exploring genes of unknown functions in *AT-microarray-dataset*

Our aim in the case study of *AT-microarray-dataset* was to explore patterns of expression of genes with little or nothing known about them. The well-vetted dataset we used (23), encompasses expression values for 22 746 genes across 424 *A. thaliana* samples, representing 71 diverse studies and a wide variety of environmental, genetic and developmental conditions (23). We updated the gene metadata to the current TAIR annotations (33) and added phylostrata designations (obtained from phylostratr (59)).

We sought to identify genes of unknown or partially known function that might be involved in photosynthesis, the process that gave rise to the earth's oxygenated atmosphere and the associated evolution of extant complex eukaryotic species. We focused particularly on regulation of the assembly and disassembly of the photosystem I and II light harvesting complexes; these dynamic processes respond sensitively to shifts in light and other environmental factors (91–94). In particular, Met1 (AT1G55480) is a 36 Kda protein that regulates the assembly of the photosystem II (PSII) complex (94). To explore genes that might be involved in PSII assembly, we calculated Spearman correlation of Met1 expression with that of the 22 746 genes represented on the Affymetrix chip (Figure 5). This analysis finds 104 genes whose expression is highly correlated to Met1 (Spearman's coefficient > 0.9) across all conditions (Supplementary File 7).

Table 3. Genes identified by MOG as showing changing expression with cancer progression (B-H corrected *P*-value < 0.05) that had been identified in experimental studies as potential prognostic biomarkers but were not marked as prognostic for the given cancer type in The Human Protein Atlas (THPA) (80)

Disease	Gene	Gene name	Pattern	Ref.
LIHC	ARG1	arginase 1	Decreasing	(85)
LIHC	CYP2C8	cytochrome P450 family 2 subfamily C member 8	Decreasing	(86)
LIHC	CYP3A4	cytochrome P450 family 3 subfamily A member 4	Decreasing	(87)
LIHC	CYP3A7	cytochrome P450 family 3 subfamily A member 7	Decreasing	(87)
LIHC	CYP4A11	cytochrome P450 family 4 subfamily A member 11	Decreasing	(88)
THCA	CHI3L1	chitinase 3 like 1	Increasing	(119)
THCA	SFTPB	surfactant protein B	Increasing	(120)
THCA	CD207	CD207 molecule	Increasing	(121)
THCA	MUC21	mucin 21, cell surface associated	Increasing	(121)
THCA	MMP7	matrix metallopeptidase 7	Increasing	(122,123)
THCA	IGFL2	IGF like family member 2	Increasing	(121)
THCA	KLK7	kallikrein related peptidase 7	Increasing	(124,125)
THCA	FN1	fibronectin 1	Increasing	(126)

We examined whether genes of photosynthesis were over-represented in this Met1-co-expressed cohort. Among the Met1 co-expressed genes, the Gene Ontology (GO) Biological Functional terms most highly over-represented (*P*-value $< 10^{-5}$) are integral to the light reactions of photosynthesis: generation of precursor metabolites and energy; photosynthetic electron transport in photosystem I (PSI); reductive pentose-phosphate cycle; response to cytokinin; and PS2 assembly (Supplementary File 7). For example, the gene most highly correlated with Met1 is At2g04039, a gene encoding the NdhV protein, which is thought to stabilize the nicotinamide dehydrogenase (NDH) complex of PS1 (95); phylostratal analysis (59) indicates that NdhV has homologs across the photosynthetic organisms, streptophyta (land plants and most green algae). Eighteen of the Met1 co-expressed genes are designated as 'unknown function' or 'uncharacterized'; six are restricted to Viridiplantae. These genes would be good candidates to evaluate experimentally for a possible function in photosynthetic light reaction.

Our next aim was to use MOG to directly explore an orphan gene (a gene encoding a protein unrecognizable by homology to those of other species) (59,96–97), and to determine potential processes that it might be involved in. First, we filtered each gene's target description to retain 'unknown'. From these, we filtered to retain only the phylostratigraphic designation '*A. thaliana*'. From this gene list, we identified genes that had an expression value greater than 100 in at least five samples. We selected the orphan gene of unknown function, At2G04675, for exploratory analysis. At2G04675 encodes a predicted protein of 67 aa with no known sequence domains (domains searched using CDD (98)). A Pearson correlation analysis of the expression pattern of At2G04675 with the other genes represented on the Affymetrix chip showed 48 genes had a Pearson correlation of higher than 0.95 (Supplementary File 7); these genes are expressed almost exclusively in pollen (the male gametophyes of flowering plants) (Figure 6). The exploration implicates At2G04675 as a candidate for involvement in some aspect of pollen biology.

Using MOG to further explore genes that are associated with pollen, we identified sets of leaf and pollen samples (Supplementary Figure S8 and Additional File 5), and then calculated genes that are differentially expressed in the leaf samples versus the pollen samples using a Mann–Whitney U test (fold change of 2-fold or more; BH corrected *P*-value $< 10^{-3}$) (Additional File 5). The GO terms most highly enriched (*P*-value $< 10^{-20}$) among genes upregulated in pollen are processes of cell cycle, mitosis, organellar fission, chromosome organization and DNA repair (Additional File 5). This reflects and emphasizes the critical role of these processes in male gametophyte development, particularly sperm biogenesis. Each angiosperm pollen grain must produce two viable sperm each used in the double fertilization of the ovule. Above all else, proper mitogenesis is essential to the function of a pollen grain. We visualized the *leaf versus pollen* differential analysis by volcano plot (Figure 7 and Supplementary Figure S9), this time to explore genes upregulated in leaf. Among these is At1G67860, an Arabidopsis specific gene encoding a protein of 'unknown function'. We used MOG to correlate expression of this gene versus all genes across all samples. One hundred sixteen genes, dispersed across all five chromosomes, are co-expressed with At1G67860 (Spearman correlation ≥ 0.65) (Supplementary File 7). The genes are expressed almost exclusively in mature leaf (Supplementary Figure S10). Most have no known function; a GO enrich-

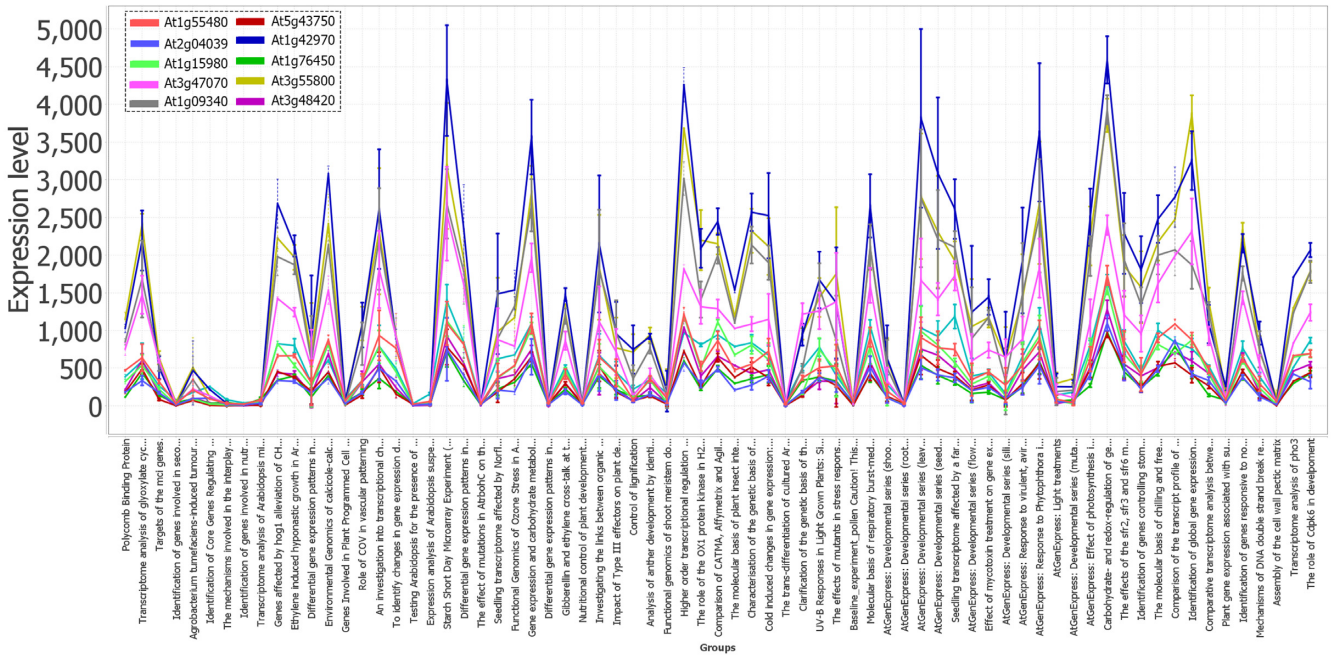

Figure 5. Spearman correlation followed by line-plot visualization, using MOG, shows that Met1 (At1G55480) is highly expressed in photosynthetic organs and highly correlated with several genes of unknown function. The 'peaks' of expression are all leaf samples; the 'troughs' of expression are predominantly root and cell culture samples. *AT-microarray-dataset* representing 71 diverse studies and a wide variety of environmental, genetic and developmental conditions (23). Several genes of unknown function are closely co-expressed in this cluster.

ment test indicates that GO biological processes overrepresented (P-value $< 10^{-3}$) among the genes are: defense response, response to stress, response to external biotic stimulus and response to other organism (Supplementary File 7).

Identifying co-expressed metabolites in *AT-metab-dataset*

Metabolomics is providing a growing resource for understanding metabolic pathways and identifying the structural and regulatory genes that shape these pathways and their interconnected lattice (61,99–100). Here, we use the *AT-metab-dataset* metabolomics dataset that represents a comprehensive study of 50 mutants with a normal morphological phenotype but altered metabolite levels, and 19 wild-type control lines (60). There are 8–16 biological replicates for each genetic line; data is corrected for batch effects. Data and metadata were retrieved from PMR (61). The aim of this case study was to tease out co-expressed metabolites that are affected by genetic perturbations. We identified a group of four highly co-expressed metabolites (Pearson correlation > 0.8): the amino acid arginine, its precursor L-ornithine, cyclic ornithine (3-amino-piperidine-2-one), and one unidentified metabolite. Plots across the means of the biological replicates of each sample (Supplementary Figure S11), shows accumulation of these metabolites is upregulated over 4-fold in four mutant lines: *mur9*, mutants have altered cell wall constituents; *vtc1*, encodes GDP-mannose pyrophosphorylase, required for synthesis of manose, major constituent of cell walls, upregulated upon bacterial infection; *cim13*, gene of unknown function associated with disease resistance, *eto1*, negative regulator of biosynthesis of the plant hormone ethylene. An arginine-derived metabolite, nitrous oxide, has been widely implicated in signaling pathways in plants (101). MOG analysis might suggest to a researcher a potential relationship between arginine and the cell wall defense response, providing a suggestion for future experimentation.

Comparison to other software

Few tools that do not require coding are available for on-the-fly exploration of expression data. Most are 'shiny' (13) apps (15–18,102) providing a web interface to a limited number of R packages for data visualization, batch correction, differential expression analysis, PCA analysis (among samples) and gene enrichment analysis. Although shiny (13) is constantly improving, existing tools written in R (15–17) must rely on R's present capabilities for interactive applications (103). In contrast to R, Java, MOG's platform, has been used to develop numerous software with interfaces that are interactive and user-friendly (e.g. (78,104–106)), and MOG provides the researcher with specialized GUIs and methods for exploratory data analysis. MOG's GUI allows direct interactivity with the data through interactive tables, trees and visualizations, so that a researcher can easily explore data from different perspectives.

Most available R-based tools read all data directly into the main memory. Thus, on a laptop/desktop computer, analysis of a big dataset is slow (or crashes) if the available memory is not sufficiently large. For example, a dataset of 100 000 human transcripts over 5000 samples (500 000 000 expression values) requires at least 4GB (8 byte for each value) of free memory to be loaded into memory at once.

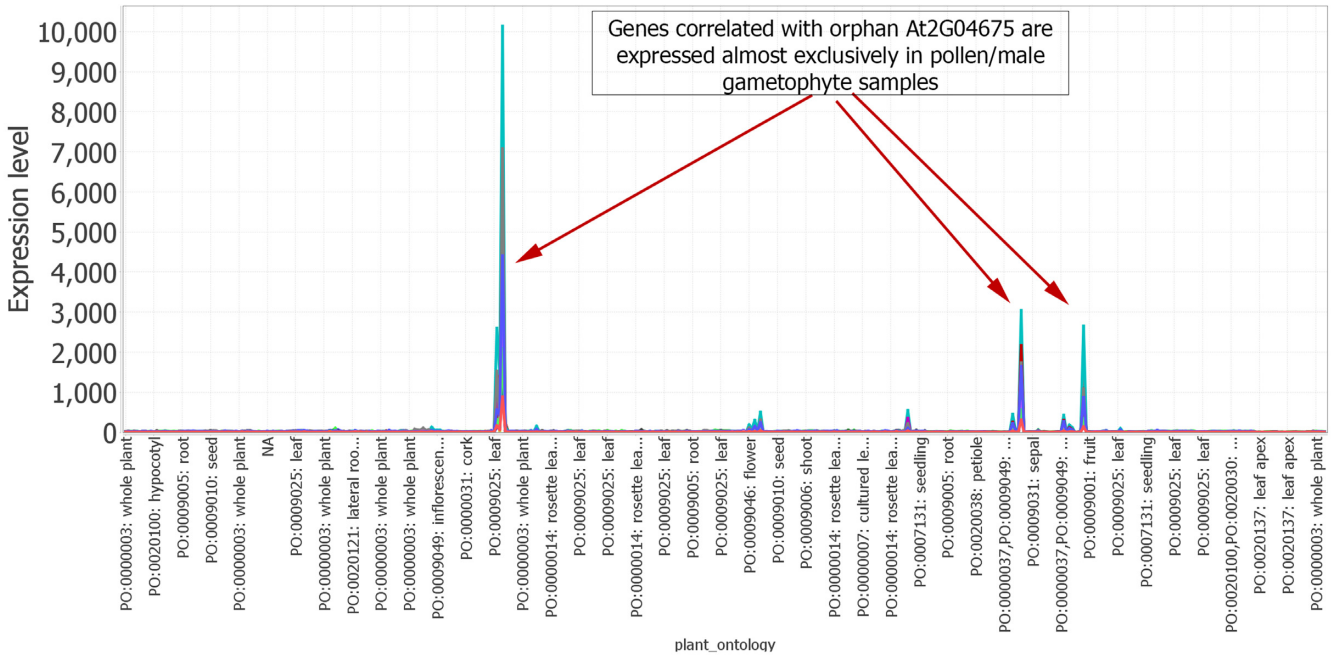

Figure 6. MOG line chart visualization shows the expression of orphan gene At2G04675 over the *AT-microarray-dataset* representing 71 diverse studies and a wide variety of environmental, genetic and developmental conditions (23). X-axis are samples, and Y-axis indicates their expression value. The orphan gene At2G04675 is of no known function, and genes highly correlated with At2G04675 are expressed almost exclusively in pollen/male gametophyte samples. Each line represents a gene. (Lines in this visualization are for clarity and the connections from sample to sample do not imply a relationship).

To circumvent this problem, R developers can use the new DelayedArray (107) framework together with DelayedMatrixStats (108) which can enable efficient handling of big datasets with R. For example, iSEE's (18) code is compatible with using DelayedArray (107) objects.

In contrast, MOG uses an indexing strategy to read data only when it is needed, which drastically reduces the total memory consumption of the system. Table 4 compares five of the most recent tools for exploratory analysis of expression data to MOG. (More details are provided in Supplementary File 8.)

Benchmarking. We benchmarked MOG's performance with the *Hu-cancer-RNASeq-dataset* (18 212 genes over 7142 samples) using a laptop with 64 bit Windows 10, 8 GB RAM and Intel(R) Core(TM) i5-7300HQ CPU; the system's resource utilization was monitored by Windows Performance Monitor tool (WPMT) (109). During benchmarking, only the software being tested was running. MOG's efficiency was compared to that of one of the R-based 'shiny' web-app (13) (choosing PIVOT, because it permits loading normalized data).

PIVOT repeatedly crashed and failed to load the full *Hu-cancer-RNASeq-dataset* (Additional File 6), but was able to load a subset of data consisting only of 410 tumor and non-tumor liver samples. We measured the execution time (time taken to compute and display output) of the Mann–Whitney U test for differentially expressed genes in tumor versus non-tumor samples. The test completed in 21 min with PIVOT, but only seven seconds with MOG (Figure 8). We kept MOG running idle until total runtime reached 30 min and compared memory and processor usage (Supplementary File 8); average memory usage of PIVOT (1869 MB) was about twice that of MOG (995 MB) (Supplementary File 8). Peak % processor time (CPU) was greater for MOG; however, MOG completed its task much more quickly, and over the 30 min, the *average* % processor time was 64% for PIVOT but only 2% for MOG (Figure 8A).

We benchmarked MOG's performance on datasets of different sizes, created by splitting the *Hu-cancer-RNASeq-dataset* by organ type (tumor and non-tumor samples). For each dataset, we performed a Mann–Whitney U test on all the genes for tumor versus non-tumor groups. MOG took only 31 s to compute a Mann–Whitney U test on 18 212 genes over 1054 samples (Figure 8B and Additional File 6). We then measured the execution time for calculating Pearson correlations of one gene with all others. MOG took only a couple of seconds to compute a Pearson correlation over 1000 samples and 16 s to compute over 7142 samples (Figure 8C).

DISCUSSION AND CONCLUSION

We demonstrated MOG's functionalities by exploring three different well-validated datasets: a human RNA-Seq dataset from non-tumor and tumor samples (*Hu-cancer-RNASeq-dataset*), an *A. thaliana* microarray dataset (*AT-microarray-dataset*) and an *A. thaliana* metabolomics dataset (*AT-metab-dataset*). In each case, known information was recapitulated in the MOG analysis, and new potential relationships became apparent.

During exploration of the *Hu-cancer-RNASeq-dataset* by MOG, we created a catalog of genes that are differentially expressed in different types of tumors, identifying in this

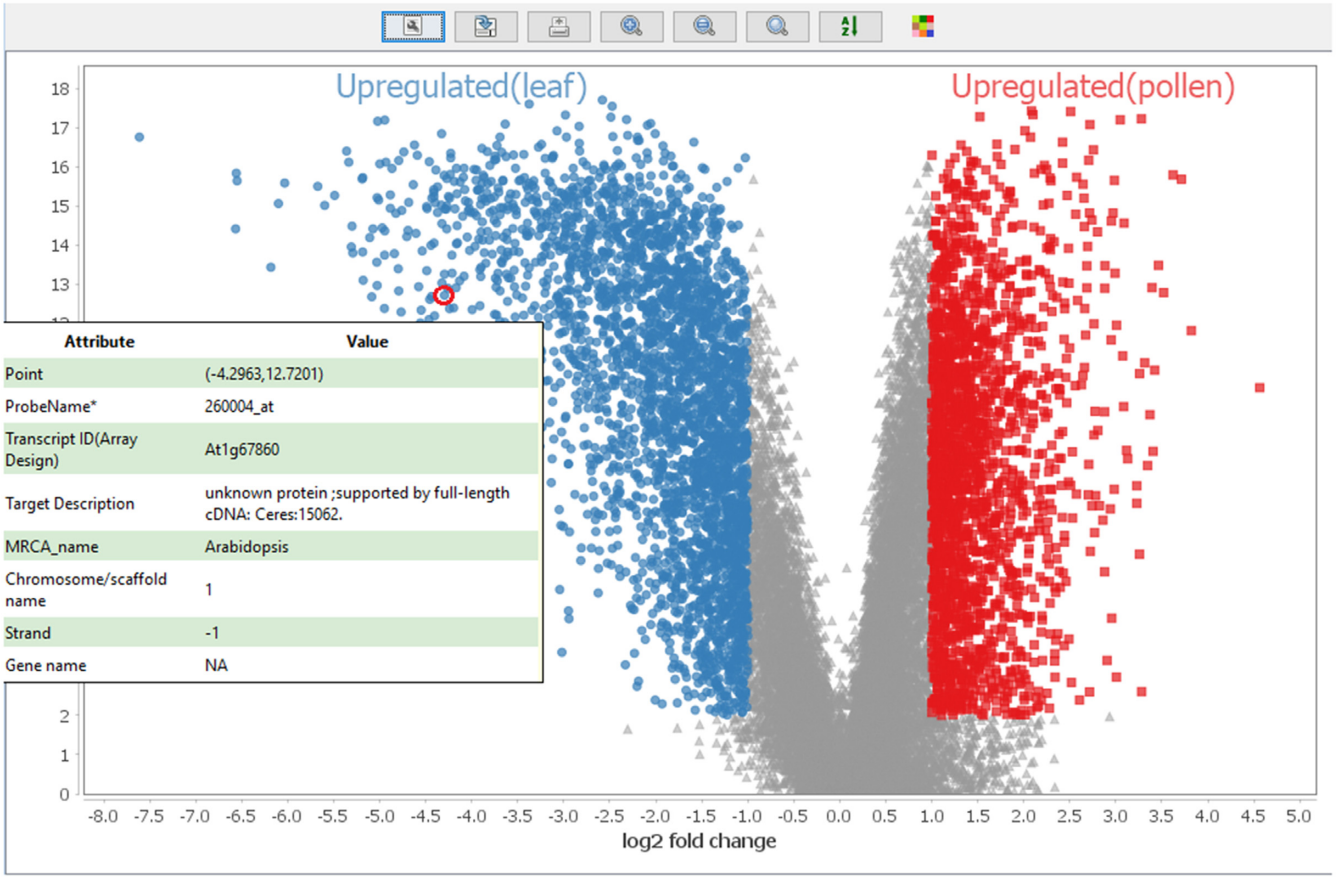

Figure 7. Using MOG for differential expression analysis of leaf and pollen samples, followed by volcano plot visualization (Y-axis: -$\log_{10}$(*P*-value)). At1G67860, an Arabidopsis specific gene with no known function, is 16-fold more highly accumulated in leaves relative to pollen (Mann–Whitney U test; BH corrected *P*-value$<10^{-3}$). The gene metadata is revealed upon hovering the mouse over a data point.

process 35 genes that are consistently upregulated or downregulated in every type of cancer in the dataset. GPC3 (67,71–72) was identified by MOG as a biomarker gene for liver cancer. Gene-level resolution analysis by MOG revealed that the cadre of genes that are co-expressed with the GPC3 gene change drastically among the individual organs, and between tumor samples and corresponding non-tumor samples. By mining the sample and study metadata, we identified genes that showed regulation with cancer progression. Many of the genes we identified have been reported previously in the literature and in THPA to be prognostic biomarkers for different cancers. Many other genes that MOG identified as differentially expressed genes are *not marked as prognostic in THPA*. These genes present potential new biomarkers for disease progression. Because each tumor type has many variations, investigating multiple candidate prognostic markers in individual tumors can provide critical information for personalized medicine (110).

Using the *AT-microarray-dataset*, we explored expression patterns of genes with unknown functions including orphan genes, identifying 18 mostly plant-restricted genes that are tightly co-expressed with genes central to photosystem assembly. We also identified an Arabidopsis-specific gene, At2G04675, to be highly expressed in pollen development, suggesting a potential involvement of this gene in gametogenesis. With the *AT-metab-dataset*, we identified a potential relationship between arginine and the cell wall defense response. Such exploratory analyses provide clues as to how to approach experimentally testing the function of these genes or metabolites.

Processing multiple heterogeneous RNA-Seq data is a formidable and unsolved challenge. We have intentionally not added capabilities for data processing (e.g. alignment, normalization, and batch-correction to minimize unwanted technical and biological effects) into MOG for two reasons. First, the selection of appropriate statistical and computational methods depends on the data structure and the biological questions to be asked. Different types of data have different characteristics (111–113), and if statistical methods are misapplied during normalization and batch-correction, especially when the data are from multiple heterogeneous studies, the resultant dataset may be misleading. Much as if using R or MATLAB statistical software, a MOG user must consider these technicalities. Second, the data science field is far from unsettled (112–115) with new approaches and variations being developed each year. (GoogleScholar retrieved over 10 000 journal articles from the first half of 2019 for 'RNA-Seq normalization methods'). Potentially a researcher could use MOG as a tool to compare the results of different methods of processing the

Table 4. MOG compared to existing tools for exploratory analysis of expression data

	MOG	PIVOT	iSEE	iGEAK	IRIS-EDA	DEvis
Reference	This paper	(15)	(18)	(16)	(17)	(102)
Year	2019	2018	2018	2019	2019	2019
Platform/GUI	Java/Swing	R/Shiny	R/Shiny	R/Shiny	R/Shiny	R/None
Interactive tables and trees	Yes	No	No	No	No	No
Interactive drag and drop operations	Yes	No	No	No	No	No
Interactive visualizations	Yes	Partial	Partial	Partial	Partial	No
Interactively subset data	Yes	Partial	Partial	Partial	No	No
Save progress	Yes	Yes	Partial (if user saves R code)	No	No	No
Use any R package	Yes	No	No	No	No	No
Supported data types	Omics or other numerical data	RNA-Seq/scRNA-Seq	Omics	RNA-Seq/microarray	RNA-Seq/scRNA-Seq	RNA-Seq
MW U test (sec.)	7	1260	NA	NA	NA	NA

MOG's GUI, designed with Java swing, is fully interactive; in contrast, other available tools are based on R and provide limited or no interactivity. A MOG user can execute any R package/script with interactively selected subsets of data if s/he wishes to perform additional analysis, whereas only a limited number of R-packages are available in the other tools. The last row compares the Mann–Whitney U test's execution time for MOG and PIVOT using the liver tumor and non-tumor datasets (18 212 genes over 410 samples). A more detailed comparison of the tools is available in Supplementary File 8.

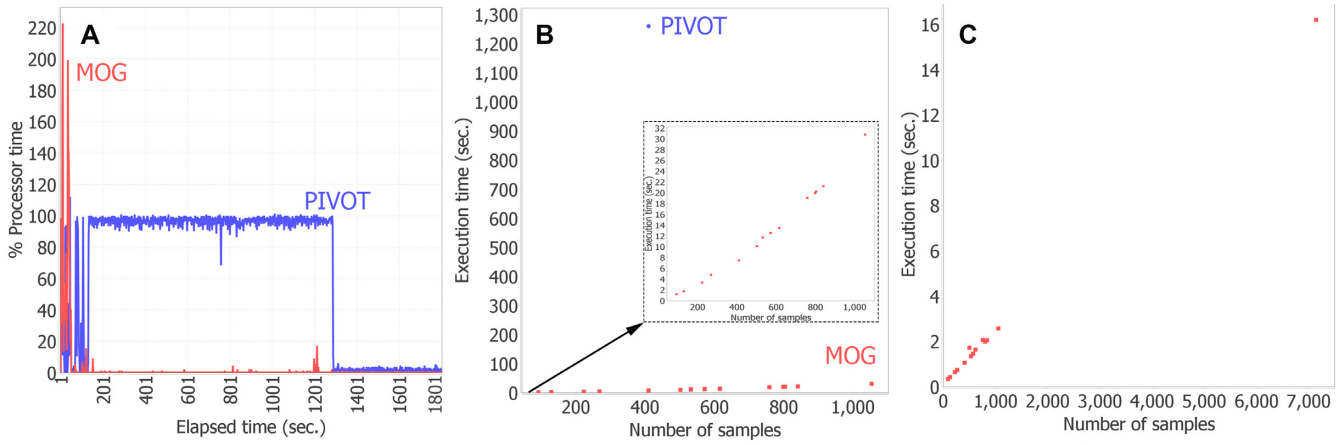

Figure 8. MOG performance benchmarks. MOG was benchmarked using the entire *Hu-cancer-RNASeq-dataset* (18 212 genes over 7142 samples), and using chunks of this dataset.(**A**) Comparison of MOG to R-based (PIVOT). Dataset size was limited to the amount of data that could be loaded in PIVOT (410 samples). % processor time (% CPU utilization) was calculated over 30 min; theoretical maximum value = total processors in computer x 100 (400 in this case). (**B**) Execution times for computing differentially expressed genes using Mann–Whitney U test. Red dots, MOG; blue dot, PIVOT (410 samples). Inset, expanded scale to display MOG execution times. (**C**) MOG execution times for pairwise computations of Pearson correlation of a gene (BIRC5) with all other genes in the datasets. (Other tools cannot perform this computation). Execution times are linear with data size; full dataset analysis took 16 s.

same raw data. Such interactive comparisons would enable biologists to gain insight as to which processing methods best reflect experimentally-established 'ground truths'. This approach would provide a complement to the more typical validation of a dataset by determining GO term enrichment in gene clusters.

Analyses performed while exploring and statistically analyzing datasets on MOG can be saved; by clicking 'save', all the analyses that have been performed are added as objects to the MOG project file. Results obtained with MOG can be shared by sharing the saved MOG project file. If a user wishes to document the information to reproduce the analysis, she/he needs to manually specify the parameters and methods used. In the future, we plan to implement automated report generation for each analysis.

MOG is a novel Java software for interactive exploratory analysis of big 'omics datasets or other datasets. By using an indexing strategy to read data only when it is needed, the total memory consumption of the system is minimized, enabling MOG to perform much more efficiently than the available R-based software. Visualizations produced by MOG are fully interactive, and enable researchers to detect and mine interesting data points and probe the relationships among them. The statistical methods implemented in MOG help to guide exploration of hidden patterns in a user-friendly manner. By integrating metadata, MOG affords an opportunity to extract new insights into the relationships between gene expression and gene structure, gene location, or any of the diverse information entered by scientists about the biology and experimental conditions.

Taken together these features can aid a researcher in developing new, experimentally testable hypotheses.

DATA AVAILABILITY

We subscribe to FAIR data and software practices (116). MOG is free and open source software published under the MIT License. MOG software, user guide and all compiled datasets in this article are freely downloadable from http://metnetweb.gdcb.iastate.edu/MetNet_MetaOmGraph.htm. MOG's source code and user guide is available at https://github.com/urmi-21/MetaOmGraph/. MOG's source code (version 1.8.0) at the time of submission is archived and can be accessed using the DOI:10.5281/zenodo.3520986. Additional files are available at https://github.com/urmi-21/MetaOmGraph/tree/master/MOG_SupportingData.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We especially thank Nick Ransom for his formative role in MOG's early development. We are grateful to our collaborators, Kevin Bassler, Pramesh Singh and Ling Li, for their help and feedback. We much appreciate the efforts of Jing Li, Priyanka Bhandhary, Arun Seetharam and the early-adopters who beta-tested MOG and provided valuable feedback.

FUNDING

National Science Foundation Grant [IOS 1546858, in part]; Orphan Genes, An Untapped Genetic Reservoir of Novel Traits; Center for Metabolic Biology, Iowa State University. Funding for open access charge: National Science Foundation Grant [IOS 1546858].
Conflict of interest statement. None declared.

REFERENCES

1. Brazma,A., Parkinson,H., Sarkans,U., Shojatalab,M., Vilo,J., Abeygunawardena,N., Holloway,E., Kapushesky,M., Kemmeren,P., Lara,G.G. *et al.* (2003) ArrayExpressa public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.*, **31**, 68–71.
2. Kodama,Y., Shumway,M. and Leinonen,R. (2011) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
3. Haug,K., Salek,R.M., Conesa,P., Hastings,J., de Matos,P., Rijnbeek,M., Mahendraker,T., Williams,M., Neumann,S., Rocca-Serra,P. *et al.* (2012) MetaboLightsan open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.*, **41**, D781–D786.
4. Martens,L., Hermjakob,H., Jones,P., Adamski,M., Taylor,C., States,D., Gevaert,K., Vandekerckhove,J. and Apweiler,R. (2005) PRIDE: the proteomics identifications database. *Proteomics*, **5**, 3537–3545.
5. Lazar,C., Meganck,S., Taminau,J., Steenhoff,D., Coletta,A., Molter,C., Weiss-Solís,D.Y., Duque,R., Bersini,H. and Nowé,A. (2012) Batch effect removal methods for microarray gene expression data integration: a survey. *Brief. Bioinform.*, **14**, 469–490.
6. Rhodes,D.R. and Chinnaiyan,A.M. (2005) Integrative analysis of the cancer transcriptome. *Nat. Genet.*, **37**, S31–S37.
7. Li,J., Arendsee,Z., Singh,U. and Wurtele,E.S. (2019) Recycling RNA-Seq Data to Identify Candidate Orphan Genes for Experimental Analysis. bioRxiv doi: https://doi.org/10.1101/671263, 21 June 2019, preprint: not peer reviewed.
8. Ihaka,R. and Gentleman,R. (1996) R: a language for data analysis and graphics. *J. Comput. Graph. Stat.*, **5**, 299–314.
9. Rau,A., Marot,G. and Jaffrézic,F. (2014) Differential meta-analysis of RNA-seq data from multiple studies. *BMC Bioinformatics*, **15**, 91.
10. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.
11. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
12. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
13. Chang,W., Cheng,J., Allaire,J., Xie,Y., McPherson,J. and shiny: Web Application Framework for R (2018) *R package version 1.2.0*.
14. Ma,T., Huo,Z., Kuo,A., Zhu,L., Fang,Z., Zeng,X., Lin,C.-W., Liu,S., Wang,L., Liu,P. *et al.* (2018) MetaOmics: analysis pipeline and browser-based software suite for transcriptomic meta-analysis. *Bioinformatics*, **35**, 1597–1599.
15. Zhu,Q., Fisher,S.A., Dueck,H., Middleton,S., Khaladkar,M. and Kim,J. (2018) PIVOT: platform for interactive analysis and visualization of transcriptomics data. *BMC Bioinformatics*, **19**, 6.
16. Choi,K. and Ratner,N. (2019) iGEAK: an interactive gene expression analysis kit for seamless workflow using the R/shiny platform. *BMC Genomics*, **20**, 177.
17. Monier,B., McDermaid,A., Wang,C., Zhao,J., Miller,A., Fennell,A. and Ma,Q. (2019) IRIS-EDA: An integrated RNA-Seq interpretation system for gene expression data analysis. *PLoS Comput. Biol.*, **15**, e1006792.
18. Rue-Albrecht,K., Marini,F., Soneson,C. and Lun,A.T. (2018) iSEE: interactive summarizedexperiment explorer [version 1; peer review: 3 approved]. *F1000Research*, **7**, 741.
19. Kucukural,A., Yukselen,O., Ozata,D.M., Moore,M.J. and Garber,M. (2019) DEBrowser: interactive differential expression analysis and visualization tool for count data. *BMC Genomics*, **20**, 6.
20. Marini,F. (2018) ideal: Interactive Differential Expression AnaLysis. *Bioconductor*, doi:10.18129/B9.bioc.ideal.
21. Wang,Q., Armenia,J., Zhang,C., Penson,A.V., Reznik,E., Zhang,L., Minet,T., Ochoa,A., Gross,B.E., Iacobuzio-Donahue,C.A. *et al.* (2018) Unifying cancer and normal RNA sequencing data from different sources. *Scientific data*, **5**, 180061.
22. Brazma,A. and Vilo,J. (2000) Gene expression data analysis. *FEBS Lett.*, **480**, 17–24.
23. Mentzen,W.I. and Wurtele,E.S. (2008) Regulon organization of Arabidopsis. *BMC Plant Biol.*, **8**, 99.
24. Almeida-de Macedo,M.M., Ransom,N., Feng,Y., Hurst,J. and Wurtele,E.S. (2013) Comprehensive analysis of correlation coefficients estimated from pooling heterogeneous microarray data. *BMC Bioinformatics*, **14**, 214.
25. Trevino,S., Sun,Y., Cooper,T.F. and Bassler,K.E. (2012) Robust detection of hierarchical communities from *Escherichia coli* gene expression data. *PLoS Comput. Biol.*, **8**, e1002391.
26. Tukey,J.W. (1981) Exploratory Data Analysis. Addison-Wesley Publishing Company Reading, Mass. — Menlo Park, Cal., London, Amsterdam, Don Mills, Ontario, Sydney 1977, XVI, 688 S. *Biometrical J.*, **23**, 413–414.
27. Kelder,T., Conklin,B.R., Evelo,C.T. and Pico,A.R. (2010) Finding the right questions: exploratory pathway analysis to enhance biological discovery in large datasets. *PLoS Biol.*, **8**, e1000472.
28. Shannon,P.T., Grimes,M., Kutlu,B., Bot,J.J. and Galas,D.J. (2013) RCytoscape: tools for exploratory network analysis. *BMC Boinformatics*, **14**, 217.
29. Safran,M., Dalah,I., Alexander,J., Rosen,N., Iny Stein,T., Shmoish,M., Nativ,N., Bahir,I., Doniger,T., Krug,H. *et al.* (2010) GeneCards Version 3: the human gene integrator. *Database*, **2010**, doi:10.1093/database/baq020.

30. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
31. Kersey,P.J., Allen,J.E., Armean,I., Boddu,S., Bolt,B.J., Carvalho-Silva,D., Christensen,M., Davis,P., Falin,L.J., Grabmueller,C. *et al.* (2015) Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Res.*, **44**, D574–D580.
32. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2006) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
33. Lamesch,P., Berardini,T.Z., Li,D., Swarbreck,D., Wilks,C., Sasidharan,R., Muller,R., Dreher,K., Alexander,D.L., Garcia-Hernandez,M. *et al.* (2011) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.*, **40**, D1202–D1210.
34. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 14863–14868.
35. Kumari,S., Nie,J., Chen,H.-S., Ma,H., Stewart,R., Li,X., Lu,M.-Z., Taylor,W.M. and Wei,H. (2012) Evaluation of gene association methods for coexpression network construction and biological knowledge discovery. *PLoS One*, **7**, e50411.
36. Langfelder,P. and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.
37. Faith,J.J., Hayete,B., Thaden,J.T., Mogno,I., Wierzbowski,J., Cottarel,G., Kasif,S., Collins,J.J. and Gardner,T.S. (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol.*, **5**, e8.
38. van Dam,S., Vosa,U., van der Graaf,A., Franke,L. and de Magalhaes,J.P. (2017) Gene co-expression analysis for functional classification and gene–disease predictions. *Brief. Bioinform.*, **19**, 575–592.
39. Vandenbon,A., Dinh,V.H., Mikami,N., Kitagawa,Y., Teraguchi,S., Ohkura,N. and Sakaguchi,S. (2016) Immuno-Navigator, a batch-corrected coexpression database, reveals cell type-specific gene networks in the immune system. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, E2393–E2402.
40. McKenzie,A.T., Katsyv,I., Song,W.-M., Wang,M. and Zhang,B. (2016) DGCA: a comprehensive R package for differential gene correlation analysis. *BMC Syst. Biol.*, **10**, 106.
41. Wang,Y.R. and Huang,H. (2014) Review on statistical methods for gene network reconstruction using expression data. *J. Theor. Biol.*, **362**, 53–61.
42. Daub,C.O., Steuer,R., Selbig,J. and Kloska,S. (2004) Estimating mutual information using B-spline functions–an improved similarity measure for analysing gene expression data. *BMC Bioinformatics*, **5**, 118.
43. Song,L., Langfelder,P. and Horvath,S. (2012) Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics*, **13**, 328.
44. Singh,P., Chen,T., Arendsee,Z., Wurtele,E.S. and Bassler,K.E. (2017) A Regulatory Network Analysis of Orphan Genes in Arabidopsis Thaliana. APS March Meeting 2017, abstract id. V6.005, https://ui.adsabs.harvard.edu/abs/2017APS..MAR.V6005S/abstract.
45. Hedges,L.V. and Vevea,J.L. (1998) Fixed-and random-effects models in meta-analysis. *Psychol. Methods*, **3**, 486.
46. Field,A.P. (2001) Meta-analysis of correlation coefficients: a Monte Carlo comparison of fixed-and random-effects methods. *Psychol. Methods*, **6**, 161.
47. Soneson,C. and Delorenzi,M. (2013) A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, **14**, 91.
48. Fukushima,A. (2013) DiffCorr: an R package to analyze and visualize differential correlations in biological networks. *Gene*, **518**, 209–214.
49. Fisher,R.A. (1915) Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, **10**, 507–521.
50. Edgington,E.S. (1980) Validity of randomization tests for one-subject experiments. *J. Educ. Stat.*, **5**, 235–251.
51. Weisstein,E.W. (2004) Bonferroni correction. From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/BonferroniCorrection.html.
52. Holm,S. (1979) A Simple Sequentially Rejective Multiple TestProcedure. *Scand. J. Stat.*, **6**, 65–70.
53. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stat. Soc.: series B (Methodological)*, **57**, 289–300.
54. Lonsdale,J., Thomas,J., Salvatore,M., Phillips,R., Lo,E., Shad,S., Hasz,R., Walters,G., Garcia,F., Young,N. *et al.* (2013) The genotype-tissue expression (GTEx) project. *Nat.Genet.*, **45**, 580–585.
55. Colaprico,A., Silva,T.C., Olsen,C., Garofano,L., Cava,C., Garolini,D., Sabedot,T.S., Malta,T.M., Pagnotta,S.M., Castiglioni,I. *et al.* (2015) TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.*, **44**, e71.
56. Sondka,Z., Bamford,S., Cole,C.G., Ward,S.A., Dunham,I. and Forbes,S.A. (2018) The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat.Rev. Cancer*, **18**, 696–705.
57. Amberger,J.S., Bocchini,C.A., Schiettecatte,F., Scott,A.F. and Hamosh,A. (2014) OMIM. org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.*, **43**, D789–D798.
58. Yang,Y.H., Dudoit,S., Luu,P., Lin,D.M., Peng,V., Ngai,J. and Speed,T.P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.
59. Arendsee,Z., Li,J., Singh,U., Seetharam,A., Dorman,K. and Wurtele,E.S. (2019) phylostratr: a framework for phylostratigraphy. *Bioinformatics*, **35**, 3617–3627.
60. Fukushima,A., Kusano,M., Mejia,R.F., Iwasa,M., Kobayashi,M., Hayashi,N., Watanabe-Takahashi,A., Narisawa,T., Tohge,T., Hur,M. *et al.* (2014) Metabolomic characterization of knockout mutants in Arabidopsis: development of a metabolite profiling database for knockout mutants in Arabidopsis. *Plant Physiol.*, **165**, 948–961.
61. Hur,M., Campbell,A.A., Almeida-de Macedo,M., Li,L., Ransom,N., Jose,A., Crispin,M., Nikolau,B.J. and Wurtele,E.S. (2013) A global approach to analysis and interpretation of metabolic data for plant natural product discovery. *Natur. Prod. Rep.*, **30**, 565–583.
62. Slattery,M.L., Mullany,L.E., Wolff,R.K., Sakoda,L.C., Samowitz,W.S. and Herrick,J.S. (2018) The p53-signaling pathway and colorectal cancer: Interactions between downstream p53 target genes and miRNAs. *Genomics*,**111**, 762–771.
63. Boyle,E.I., Weng,S., Gollub,J., Jin,H., Botstein,D., Cherry,J.M. and Sherlock,G. (2004) GO:: TermFinderopen source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, **20**, 3710–3715.
64. Supek,F., Bošnjak,M., Škunca,N. and Šmuc,T. (2011) REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One*, **6**, e21800.
65. Kaur,S.P. and Cummings,B.S. (2019) Role of Glypicans in regulation of the tumor microenvironment and cancer progression. *Biochem. Pharmacol.*, **168**, 108–118.
66. Capurro,M.I., Xu,P., Shi,W., Li,F., Jia,A. and Filmus,J. (2008) Glypican-3 inhibits Hedgehog signaling during development by competing with patched for Hedgehog binding. *Dev. Cell*, **14**, 700–711.
67. Gao,W. and Ho,M. (2011) The role of glypican-3 in regulating Wnt in hepatocellular carcinomas. *Cancer Rep.*, **1**, 14–19.
68. Filmus,J. and Capurro,M. (2008) The role of glypican-3 in the regulation of body size and cancer. *Cell Cycle*, **7**, 2787–2790.
69. Blackhall,F.H., Merry,C.L., Davies,E. and Jayson,G.C. (2001) Heparan sulfate proteoglycans and cancer. *Brit. J. cancer*, **85**, 1094–1098.
70. Davoodi,J., Kelly,J., Gendron,N.H. and MacKenzie,A.E. (2007) The Simpson–Golabi–Behmel syndrome causative Glypican-3, binds to and inhibits the dipeptidyl peptidase activity of CD26. *Proteomics*, **7**, 2300–2310.
71. Ho,M. and Kim,H. (2011) Glypican-3: a new target for cancer immunotherapy. *Eur. J. Cancer*, **47**, 333–338.

72. Anatelli,F., Chuang,S.-T., Yang,X.J. and Wang,H.L. (2008) Value of glypican 3 immunostaining in the diagnosis of hepatocellular carcinoma on needle biopsy. *Am. J. Clin. Pathol.*, **130**, 219–223.
73. Capurro,M., Wanless,I.R., Sherman,M., Deboer,G., Shi,W., Miyoshi,E. and Filmus,J. (2003) Glypican-3: a novel serum and histochemical marker for hepatocellular carcinoma. *Gastroenterology*, **125**, 89–97.
74. Xiang,Y.-Y., Ladeda,V. and Filmus,J. (2001) Glypican-3 expression is silenced in human breast cancer. *Oncogene*, **20**, 7408–7412.
75. Sasisekharan,R., Shriver,Z., Venkataraman,G. and Narayanasami,U. (2002) Roles of heparan-sulphate glycosaminoglycans in cancer. *Nat. Rev. Cancer*, **2**, 521–528.
76. Kim,H., Xu,G.-L., Borczuk,A.C., Busch,S., Filmus,J., Capurro,M., Brody,J.S., Lange,J., D'armiento,J.M., Rothman,P.B. *et al.* (2003) The heparan sulfate proteoglycan GPC3 is a potential lung tumor suppressor. *Am. J. Respir. Cell Mol. Biol.*, **29**, 694–701.
77. Valsechi,M.C., Oliveira,A.B.B., Conceição,A.L.G., Stuqui,B., Candido,N.M., Provazzi,P.J.S., de Araújo,L.F., Silva,W.A., de Freitas Calmon,M. and Rahal,P. (2014) GPC3 reduces cell proliferation in renal carcinoma cell lines. *BMC Cancer*, **14**, 631.
78. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
79. Bader,G.D. and Hogue,C.W. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**, 2.
80. Uhlen,M., Zhang,C., Lee,S., Sjöstedt,E., Fagerberg,L., Bidkhori,G., Benfeitas,R., Arif,M., Liu,Z., Edfors,F. *et al.* (2017) A pathology atlas of the human cancer transcriptome. *Science*, **357**, eaan2507.
81. Sun,X., Zhang,H., Luo,L., Zhong,K., Ma,Y., Fan,L., Fu,D. and Wan,L. (2016) Comparative proteomic profiling identifies potential prognostic factors for human clear cell renal cell carcinoma. *Oncol. Rep.*, **36**, 3131–3138.
82. Swe,M.T., Pongchaidecha,A., Chatsudthipong,V., Chattipakorn,N. and Lungkaphin,A. (2019) Molecular signaling mechanisms of renal gluconeogenesis in nondiabetic and diabetic conditions. *J. Cell. Physiol.*, **234**, 8134–8151.
83. Pan,W., Cheng,Y., Zhang,H., Liu,B., Mo,X., Li,T., Li,L., Cheng,X., Zhang,L., Ji,J. *et al.* (2014) CSBF/C10orf99, a novel potential cytokine, inhibits colon cancer cell growth through inducing G1 arrest. *Sci. Rep.*, **4**, 6812.
84. Tian,Z.-H., Yuan,C., Yang,K. and Gao,X.-L. (2019) Systematic identification of key genes and pathways in clear cell renal cell carcinoma on bioinformatics analysis. *Ann. Transl. Med.*, **7**, 89.
85. You,J., Chen,W., Chen,J., Zheng,Q., Dong,J. and Zhu,Y. (2018) The Oncogenic Role of ARG1 in Progression and Metastasis of Hepatocellular Carcinoma. *Biomed Res. Int.*, **2018**, 1–10.
86. Ren,X., Ji,Y., Jiang,X. and Qi,X. (2018) Downregulation of CYP2A6 and CYP2C8 in tumor tissues is linked to worse overall survival and recurrence-free survival from hepatocellular carcinoma. *Biomed. Res. Int.*, **2018**, 5859415.
87. Yu,T., Wang,X., Zhu,G., Han,C., Su,H., Liao,X., Yang,C., Qin,W., Huang,K. and Peng,T. (2018) The prognostic value of differentially expressed CYP3A subfamily members for hepatocellular carcinoma. *Cancer Manag.Res.*, **10**, 1713–1726.
88. Eun,H.S., Cho,S.Y., Lee,B.S., Kim,S., Song,I.-S., Chun,K., Oh,C.-H., Yeo,M.-K., Kim,S.H. and Kim,K.-H. (2019) Cytochrome P450 4A11 expression in tumor cells: a favorable prognostic factor for hepatocellular carcinoma patients. *J. Gastroenterol. Hepatol.*, **34**, 224–233.
89. de Vries,J.K., Levin,A., Loud,F., Adler,A., Mayer,G. and Pena,M.J. (2018) Implementing personalized medicine in diabetic kidney disease: Stakeholders' perspectives. *Diabetes Obes. Metab.*, **20**, 24–29.
90. Lightbody,G., Haberland,V., Fiona,B., Taggart,L., Zheng,H., Parks,E. and Blayney,J. (2018) Review of applications of high-throughput sequencing in personalised medicine: barriers and facilitators of future progress in research and clinical application. *Brief.Bioinform.*, **20**, 1795–1811.
91. Chen,Y.-E., Su,Y.-Q., Mao,H.-T., Nan,W., Zhu,F., Yuan,M., Zhang,Z.-W., Liu,W.-J. and Yuan,S. (2018) Terrestrial plants evolve highly-assembled photosystem complexes in adaptation to light shifts. *Front. Plant Sci.*, **9**, 1811.
92. Ruban,A.V. and Johnson,M.P. (2015) Visualizing the dynamic structure of the plant photosynthetic membrane. *Nat. Plants*, **1**, 15161.
93. Nosek,L., Semchonok,D., Boekema,E.J., Ilík,P. and Kouřil,R. (2017) Structural variability of plant photosystem II megacomplexes in thylakoid membranes. *Plant J.*, **89**, 104–111.
94. Bhuiyan,N.H., Friso,G., Poliakov,A., Ponnala,L. and van Wijk,K.J. (2015) MET1 is a thylakoid-associated TPR protein involved in photosystem II supercomplex formation and repair in Arabidopsis. *Plant Cell*, **27**, 262–285.
95. Fan,X., Zhang,J., Li,W. and Peng,L. (2015) The NdhV subunit is required to stabilize the chloroplast NADH dehydrogenase-like complex in Arabidopsis. *Plant J.*, **82**, 221–231.
96. Arendsee,Z.W., Li,L. and Wurtele,E.S. (2014) Coming of age: orphan genes in plants. *Trends Plant Sci.*, **19**, 698–708.
97. Gollery,M., Harper,J., Cushman,J., Mittler,T., Girke,T., Zhu,J.-K., Bailey-Serres,J. and Mittler,R. (2006) What makes species unique? The contribution of proteins with obscure features. *Genome Biol.*, **7**, R57.
98. Marchler-Bauer,A., Derbyshire,M.K., Gonzales,N.R., Lu,S., Chitsaz,F., Geer,L.Y., Geer,R.C., He,J., Gwadz,M., Hurwitz,D.I. *et al.* (2014) CDD: NCBI's conserved domain database. *Nucleic Acids Res.*, **43**, D222–D226.
99. Sumner,L.W., Lei,Z., Nikolau,B.J. and Saito,K. (2015) Modern plant metabolomics: advanced natural product gene discoveries, improved technologies, and future prospects. *Nat. Prod. Rep.*, **32**, 212–229.
100. Quanbeck,S.M.M., Brachova,L., Campbell,A.A., Guan,X., Perera,A., He,K., Rhee,S.Y., Bais,P., Dickerson,J., Dixon,P. *et al.* (2012) Metabolomics as a hypothesis-generating functional genomics tool for the annotation of Arabidopsis thaliana genes of unknown function. *Front. Plant Sci.*, **3**, 15.
101. del Rıo,L.A., Corpas,F.J. and Barroso,J.B. (2004) Nitric oxide and nitric oxide synthase activity in plants. *Phytochemistry*, **65**, 783–792.
102. Price,A., Caciula,A., Guo,C., Lee,B., Morrison,J., Rasmussen,A., Lipkin,W.I. and Jain,K. (2019) DEvis: an R package for aggregation and visualization of differential expression data. *BMC Bioinformatics*, **20**, 110.
103. Furtună,T.F. and Vinte,C. (2016) Integrating R and Java for Enhancing Interactivity of Algorithmic Data Analysis Software Solutions. *Rom. Stat. Rev*, **64**, 29–41.
104. López-Fernández,H., Reboiro-Jato,M., Glez-Peña,D., Laza,R., Pavón,R. and Fdez-Riverola,F. (2018) GC4S: a bioinformatics-oriented Java software library of reusable graphical user interface components. *PLoS One*, **13**, e0204474.
105. Ignatchenko,V., Ignatchenko,A., Sinha,A., Boutros,P.C. and Kislinger,T. (2015) VennDIS: A JavaFX-based Venn and Euler diagram software to generate publication quality figures. *Proteomics*, **15**, 1239–1244.
106. Kirov,I., Khrustaleva,L., Van Laere,K., Soloviev,A., Meeus,S., Romanov,D. and Fesenko,I. (2017) DRAWID: user-friendly java software for chromosome measurements and idiogram drawing. *Comp.Cytogenet.*, **11**, 747–757.
107. Pags,H. and with contributions from Peter Hickeywith contributions from Peter Hickey and Lun,A. (2019) DelayedArray: A unified framework for working transparently with on-disk and in-memory array-like datasets. *R package version 0.10.0.*
108. Hickey,P. (2019) DelayedMatrixStats: functions that apply to rows and columns of 'DelayedMatrix' objects. *R package version 1.6.0*.
109. Microsoft (2014) Overview of Windows Performance Monitor. *Microsoft Docs*, https://techcommunity.microsoft.com/t5/ask-the-performance-team/windows-performance-monitor-overview/ba-p/375481.
110. Cieślik,M. and Chinnaiyan,A.M. (2018) Cancer transcriptome profiling at the juncture of clinical translation. *Nat.Rev. Genet.*, **19**, 93–109.
111. Chawade,A., Alexandersson,E. and Levander,F. (2014) Normalyzer: a tool for rapid evaluation of normalization methods for omics data sets. *J. Proteome Res.*, **13**, 3114–3120.
112. Hicks,S.C., Okrah,K., Paulson,J.N., Quackenbush,J., Irizarry,R.A. and Bravo,H.C. (2017) Smooth quantile normalization. *Biostatistics*, **19**, 185–198.

113. Evans,C., Hardin,J. and Stoebel,D.M. (2017) Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief. Bioinform.*, **19**, 776–792.
114. Paulson,J.N., Chen,C.-Y., Lopes-Ramos,C.M., Kuijjer,M.L., Platig,J., Sonawane,A.R., Fagny,M., Glass,K. and Quackenbush,J. (2017) Tissue-aware RNA-Seq processing and normalization for heterogeneous and sparse data. *BMC Bioinformatics*, **18**, 437.
115. Schmidt,F., List,M., Cukuroglu,E., Köhler,S., Göke,J. and Schulz,M.H. (2018) An ontology-based method for assessing batch effect adjustment approaches in heterogeneous datasets. *Bioinformatics*, **34**, i908–i916.
116. Wilkinson,M.D., Dumontier,M., Aalbersberg,I.J., Appleton,G., Axton,M., Baak,A., Blomberg,N., Boiten,J.-W., da Silva Santos,L.B., Bourne,P.E. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci.Data*, **3**, 160018.
117. Vander Ark,A., Cao,J. and Li,X. (2018) TGF-β receptors: In and beyond TGF-β signaling. *Cell. Signal.*, **52**, 112–120.
118. Nandi,P., Lim,H., Torres-Garcia,E.J. and Lala,P.K. (2018) Human trophoblast stem cell self-renewal and differentiation: role of decorin. *Sci. Rep.*, **8**, 8977.
119. Luo,D., Chen,H., Lu,P., Li,X., Long,M., Peng,X., Huang,M., Huang,K., Lin,S., Tan,L. *et al.* (2017) CHI3L1 overexpression is associated with metastasis and is an indicator of poor prognosis in papillary thyroid carcinoma. *Cancer Biomark.*, **18**, 273–284.
120. Huang,Y., Prasad,M., Lemon,W.J., Hampel,H., Wright,F.A., Kornacker,K., LiVolsi,V., Frankel,W., Kloos,R.T., Eng,C. *et al.* (2001) Gene expression in papillary thyroid carcinoma reveals highly consistent profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 15044–15049.
121. Qiu,J., Zhang,W., Zang,C., Liu,X., Liu,F., Ge,R., Sun,Y. and Xia,Q. (2018) Identification of key genes and miRNAs markers of papillary thyroid cancer. *Biol. Res.*, **51**, 45.
122. Ysuhiro,I., Hiroshi,Y., Kennichi,K., Yasushi,N., Kanji,K. and Akira,M. (2006) Inverse relationships between the expression of MMP-7 and MMP-11 and predictors of poor prognosis of papillary thyroid carcinoma. *Pathology*, **38**, 421–425.
123. Chen,S.-T., Liu,D.-W., Lin,J.-D., Chen,F.-W., Huang,Y.-Y. and Hsu,B. R.-S. (2012) Down-regulation of matrix metalloproteinase-7 inhibits metastasis of human anaplastic thyroid cancer cell line. *Clin. Exp. Metastasis*, **29**, 71–82.
124. Zhang,H., Cai,Y., Zheng,L., Zhang,Z., Lin,X. and Jiang,N. (2018) Long noncoding RNA NEAT1 regulate papillary thyroid cancer progression by modulating miR-129-5p/KLK7 expression. *J. Cell. Physiol.*, **233**, 6638–6648.
125. Zhang,Y., Hu,J., Zhou,W. and Gao,H. (2018) LncRNA FOXD2-AS1 accelerates the papillary thyroid cancer progression through regulating the miR-485-5p/KLK7 axis. *J. Cell. Biochem.*, **120**, 79527961.
126. Zhan,S., Li,J., Wang,T. and Ge,W. (2018) Quantitative proteomics analysis of sporadic medullary thyroid cancer reveals FN1 as a potential novel candidate prognostic biomarker. *Oncologist*, **23**, 1415–1425.