Linked Data based Multi-Omics Integration and Visualization for Cancer Decision Networks

Alokkumar Jha[1], Yasar Khan[1], Qaiser Mehmood[1], Dietrich rebholz-schuhmann[1], and Ratnesh Sahay[1]

Insight Centre for Data Analytics, National University of Ireland Galway
firstname.lastname@insight-centre.org

Abstract. Visualization of Gene Expression (GE) is a challenging task since the number of genes and their associations are difficult to predict in various set of biological studies. GE could be used to understand tissue-gene-protein relationships. Currently, Heatmaps is the standard visualization technique to depict GE data. However, Heatmaps only covers the cluster of highly dense regions. It does not provide the Interaction, Functional Annotation and pooled understanding from higher to lower expression. In the present paper, we propose a graph-based technique - based on color encoding from higher to lower expression map, along with the functional annotation. This visualization technique is highly interactive (HeatMaps are mainly static maps). The visualization system here explains the association between overlapping genes with and without tissues types. Traditional visualization techniques (viz-Heatmaps) generally explain each of the association in distinct maps. For example, overlapping genes and their interactions, based on co-expression and expression cut off are three distinct Heatmaps. We demonstrate the usability using ortholog study of GE and visualize GE using GExpressionMap. We further compare and benchmark our approach with the existing visualization techniques. It also reduces the task to cluster the expressed gene networks further to understand the over/under expression. Further, it provides the interaction based on co-expression network which itself creates co-expression clusters. GExpressionMap provides a unique graph-based visualization for GE data with their functional annotation and associated interaction among the DEGs(Differentially Expressed Genes).

Introduction

RNA seq and microarray data generate DEGs with their associated expression value as RPKM counts. GE is primarily responsible for gene silencing and enhancing control by transcription initiation [5]. These genes need to be investigated to dissect the role of GE in cancer, through the networks based on their involvement. Understanding of genes could be achieved by the integration of GE and network data to prioritize disease-associated genes [25]. GE data is crucial to visualize, since the overall data is pattern driven where over/under expression drives the function of a gene. These functions could be understood provided associated functional annotation and GO terms could be displayed along with visualization. Secondly, most of the methods use Heatmaps to visualize GE, as a static representation where each information, such as Gene-Gene association, regulation and co-expression requires distinct visualization. To obtain interference for concluding the overall process of a gene, manual interpretation of the gene using distinct visualization becomes essential. Further, this retrieved knowledge required to be annotated for understanding the mechanism and associated cell cycle processes. We have demonstrated a graph-based method to visualize GE data. Graph-based methods have an added advantage over Heatmaps based visualization regarding GE, such as the basics of Heatmaps visualization is to define the similarity among the group of genes to build a co-expression network [16]. This tool also kept the basic requirement intact by keeping the color annotation based expression visualization as in the case of Heatmaps. Here reduction from darker to lighter color representation explains the higher to lower expression of the genes. Along with this it also generates intermediate interaction graph among transcripts or genes. The key advantage of such a mechanism is to understand the gene association, cluster with a maximum number of disease association and identify the group of critical transcripts associated with the disease or normal condition. One potential advantage could be in knockdown studies where genes group based on expression level could be used for experimental validation to understand the oncogenic properties of the gene. Another advantage could be understood by the use-case presented in this paper where we have demonstrated the relationship between the expression data of human and mouse.

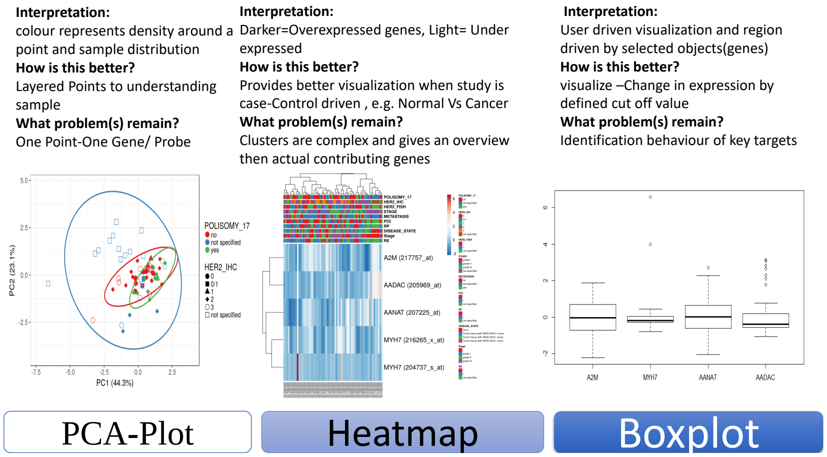
Background

Visualization of GE is key due to its functional relevance in cancer research and other diseases. However, the development of visualization and providing a scientific source such as a mathematical model, functional annotation and associated biological process will make the task of data analytics more structured. The functional annotation will also help to map down other associated biological events like gene fusion, CNV, Methylation to develop scientifically. Since the GExpressionMap approach is mathematical model driven, integration of these concepts to build data-driven discovery will be less cumbersome. The current approaches in GE are Principle Component Analysis (PCA) plot and Box plot in general. As demonstrated in Figure 1 which explains the pros and cons with three existing methods for GE visualization. The first method is the PCA method where principal component analysis has to be performed on the list of genes or transcripts. The outcome is usually being presented using M-A plot [28]. Such plots are primarily useful when working on a limited set of genes, as this approach radically decreases the density of visualization. Further, this visualization is chunky and adding a reference line could be challenging. When working on the RNA seq data where each experiment returns approximately 50000 transcripts and in this case reduction of dimensionality becomes essential. Sometimes due to biased analysis, there are many more “variables” than “observations. Along with this, these types of diagrams are generated either by using ’R’ or ’MATLAB’ which works great with smaller data sets. However, with high throughput data, it creates several issues. Further, the key to any biological outcome its functional annotation and understanding the pattern of the outcome. It is tough to accommodate functional annotation with PCA plot since data points are not so well distinguished. PCA plot is mostly static and supports limited clustering. However, it does not support the functional clustering of genes and is tough to identify sharp data points. Sometimes it is tough to distinguish two distinct clusters if they have a higher amount of overlaps. As demonstrated in Figure 1, another method associated with GE visualization is Heatmaps based visualization. Heatmaps are called as intensity plot or matrix plot, which includes dendrogram and extended Heatmaps as well [1]. As Figure 1 explains, it is a tabular view of a collection of data points, where rows represent genes, columns represent array experiments, and cells represent the measured intensity value or ratio. In GE visualization, Heatmaps provide multi-hue color maps for up- and down-regulation in combination with clustering to place similar profiles next to each other. Other extended versions of these Heatmaps are dendrogram, hierarchical clustering of genes or experiments, often combined with Heatmaps to provide more information about the cluster structures. The critical issue with Heatmaps for GE are, though it provides cluster structure, it is still far from the functional grouping of these clusters due to lack of integrated annotation and GO terms. It also has issues, such as it only supports qualitative interpretation possible due to color coding. It grows vertically with every additional profile and grows horizontally with every additional sample. These problems make the knowledge mining difficult for large-scale data sets, such as RNA-seq GE data. Crucial third method to visualize GE data, as shown in Figure 1, is a one-dimensional box plot approach. Essentially this method is used for a summary of distribution, comparison of several distributions and to see the result of normalization in differentially expressed genes. This visualization is vital to understand the sample-wise or gene-wise distribution. However, due to its 1-D nature, it does not support the multiple data types represented on a single plot. For example, for a single queried gene box can plot the cut-off for the expression. However, it will not provide the entities associated with it, such as overexpressed, underexpressed and not expressed genes. This type of plots are mainly static and as explained in Figure 1 any overlay-ed information, in this case, mutated genes for EGFR. The data points are rich even for the normalized data that it becomes tough to identify the participating entities with each gene.

Related Work

There are rich set of tools and web applications to visualize GE data and its biological and functional associations. The most related tools for GE visualization are M-A plot, Heatmaps, Scree Plot, Box Plot, Scatter Plot, Wiggle Plot, Profile Plot (also known as Parallel Coordinate Plot), VA Enhanced Profile Plots and Dendrogram. In general, Mayday[6], ClustVis[18], GENE-E [1], MISO[14] are

[1] http://www.broadinstitute.org/cancer/software/GENE-E/index.html



Interpretation:
colour represents density around a point and sample distribution
How is this better?
Layered Points to understanding sample
What problem(s) remain?
One Point-One Gene/ Probe

Interpretation:
Darker=Overexpressed genes, Light= Under expressed
How is this better?
Provides better visualization when study is case-Control driven , e.g. Normal Vs Cancer
What problem(s) remain?
Clusters are complex and gives an overview then actual contributing genes

Interpretation:
User driven visualization and region driven by selected objects(genes)
How is this better?
visualize –Change in expression by defined cut off value
What problem(s) remain?
Identification behaviour of key targets

Fig. 1. Motivational Scenario to develop GExpressionMap for breast cancer data from E-GEOD-29431

some of the most commonly used tools which covers these plots for GE. BiCluster[8] represents GE data by the hybrid approach of Heatmaps and Parallel Coordinate Plots. These plots are interactive, and GE annotations have been formalized with proper color annotations. However, this tool works on Heatmaps, and with massive data points, clusters generated by this tool can only help to infer the functionally enriched region. However, the role of each participating member and their co-expressed expressions cannot be determined. The unavailability of functional annotation and GO terms make it difficult to understand the biological processed involved with each cluster thus the pattern of the expression. INVEX [26] is again a Heatmaps based tool which deals with GE and metabolomics datasets generated from clinical samples and associated metadata, such as phenotype, donor, gender, etc. It is a web-based tool where data size has certain limitations. However, it has built-in support for gene/metabolite annotation along with Heatmaps builder. The Heatmaps builder primarily works on 'R' APIs. Though these tools have great potential due to inbuilt functional annotation, lack of clustering, interactive selection of gene entities and support for large-scale datasets provides further room for improvement. GeneXPress[21] has been developed to improve the functional annotation to reduce the task of post-processing after the obtained list of DEGs. It also contains an integrated clustering algorithm to explore the various binding sites from DEGs. Multi-view representation, which includes graph based interaction map for selected genes and Heatmaps based visualization with functional annotation, makes it a most relevant tool for GE based biological discovery with an integrated motif discovery environment. However, a single source of functional annotation raises the requirement for linked functional annotation. Again the graph visualization is limited to selected genes wherein Heatmaps identification of exact data point is a cumbersome task. GEPAT[24] is also a gene expression visualization tool developed over the Heatmaps and focused on visualization of pathway associated gene expression data. Integrated GO terms enrichment environment makes the tool unique regarding understanding the mechanism of differentially expressed genes. However, the functional annotation is mostly performed manually to have the exact mapping of each transcript involved with a certain loop of biological processed. Integrated cluster maker provides substantial support to the idea of GExpressionMap. ArrayCluster [29] is a tool developed from GE datasets, keeping in mind to resolve analytical and statistical problems associated with data. The ideal co-expression based clustering method and functional annotation of each cluster make it unique and provide a ground for GExpressionMap to include co-expression based clustering of DEGs. However, it has limited support to microarray data and makes it difficult to apply on larger gene sets generated from RNA Seq. Also, it is a Heatmaps based plotting, which makes data point selection difficult. J-Express [2] is again GE data analysis tool which contains almost every type of plot. The inclusion of various plots makes visual analytics from this tool robust. However, most of the plots generated from J-Express are static, thus lacks the key feature to understand the in-depth analysis of each tool. Integrated Gene set enrichment analysis (GSEA), Chromosome (DNA sequence) mapping and analysis, Gaussian kernels and Cross-data class prediction are some of the critical features, which makes this tool unique among others. Tang et al. presented [23] is one of the most earlier interactive visualization tool developed on the concept of ROI (Region of Interest) accommodated using scattered map. Visualization is widely supported with mathematical modeling of GE data for limited data points. This tool provides a strong foundation for GExpressionMap where we have mathematically modeled gene expression for dense and large data sets of transcripts.

Mathematical Model of GE and Visualization

It is essential to know the spectrum of visualization and behavior of visualizing events. Mathematical modeling of both provides a stable visualization system. Few attempts have been made earlier to model gene expression. Here we have modeled GE based on our requirement where we have identified the up-regulated, down-regulated and not expressed states for the genes and we have used it to identify meaningful data points in the cluster have an accurate co-expression network generated from GE data. GE data are a linear transcription model follows a system of differential equations [3]. The basic understanding of the terms are as follows; **Gene Expression**: Combination of genes code for proteins that are essential for the development and functioning of a cell or organism. **Transcript-based co-expression network** : Set of genes, proteins, small molecules, and their mutual regulatory interactions.

The modeling could be understood by Fig. 2. As per the figure, the system could be realized as

$$\frac{\partial r}{\partial t} = f(p) - Vr, \frac{\partial p}{\partial t} = Lr - Up \tag{1}$$

[2] http://jexpress.bioinfo.no/site/JexpressMain.php

where $V, U = relative\ degradation, L = Translation, r = concentration\ of\ gene, p = concentration\ of\ protein.$

To define the over, under and no expression, and stability of cluster based on interaction network, let's assume that, at given time point t, if the concentration of mRNA is $x1$ and concentration of protein is $p = x2$, then this can be generalized as a continuous function.

$$x_i(t) \in R_{\geq 0} \tag{2}$$

$$\begin{gathered} \dot{x}_i(t) = f_i(x),\ 1 \leq i \leq n \\ Say\ x_1 = mRNA\ concentration, \\ p = x_2 = protein\ concentration \end{gathered} \tag{3}$$

$$\begin{aligned} \dot{x}_1 = \kappa_1 f(x_2) - \gamma_1 x_1,\ \ \dot{x}_2 = \kappa_2 x_1 - \gamma_2 x_2 \\ \kappa_1, \kappa_2 > 0\ production\ rate,\ \ \gamma_1, \gamma_2 > 0\ degradation\ rate \end{aligned} \tag{4}$$

$$\begin{aligned} f(x_2) = f(p) = \frac{\theta^n}{\theta^n + x_2^n} \\ f(p) = \frac{\theta^n}{\theta^n + p^n} \\ if\ \theta > 0\ explains\ genes\ are\ under\ expressed \\ else\ genes\ are\ over\ expressed \end{aligned} \tag{5}$$

$$\begin{aligned} Assume\ \dot{x} = 0 \\ \dot{x_1} = 0 : x_1 = \frac{\kappa_1}{\gamma_1} f(x_2) = \frac{\kappa_1}{\gamma_1} f(p) \\ same\ as \\ \dot{x_2} = 0 : x_1 = \frac{\gamma_2}{\kappa_2} x_2 \\ x_1 = \frac{\gamma_2}{\kappa_2} p \\ for\ x_1\ and\ x_2 > 0,\ genes\ will\ not\ show\ expression \end{aligned} \tag{6}$$

Another key extension of this model will be to understand the interaction model from these under/over/non expressed genes. Lets assume that these interaction networks are continuous function and cluster building follows *rate law*, then Equation 1 can be generalized as;

$$\begin{aligned} x_i = f_i(x), where\ 1 \leq i \leq n \\ where f_i(x) = rate\ law\ for\ each\ interaction \end{aligned} \tag{7}$$

If translation happens with this gene then each cluster will follow a model to take part in post translational modification(PTM)s, that can be understood by,

$$p_i = f_i(p), where\ 1 \leq i \leq n \tag{8}$$

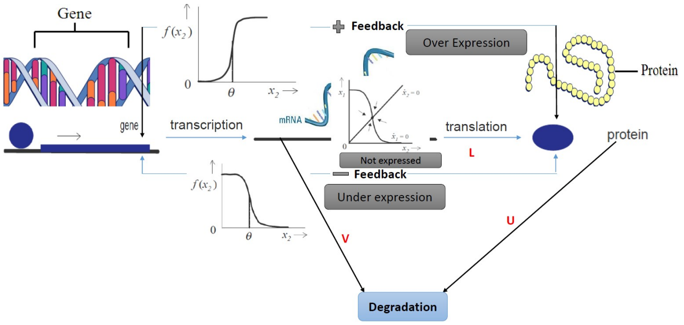


Fig. 2. Mathematical model to understand GE and gene interection based clustering.

$$\frac{\partial r}{\partial t} = f_i(p) - Vr \tag{9}$$
$$Equaltion\ for\ rate\ of$$
$$change\ in\ interection\ for\ mRNA$$

$$\frac{\partial[f_i(p)]}{\partial t} = Lr - Uf_i(p)$$
$$\int Lr.dt = f_i(p) - \int Uf_i(p).dt \tag{10}$$

This equation explains the relevance of GE based clustering and the effect of rate of change in expression. All the data points within this range of equation will have an easy to manageable knowledge mining. This will help to define the boundary and interpretation module from visualization.

Cancer Decision Networks: Integration, Model and Query Processing

Visualization is working as a presentation model with a structured and distributed model underneath for processing, filtering and querying the data. In cancer genomics, if one gene regulated by more than one events, such as gene expression, CNV, and methylation, it is unlikely that retrieved regulation occurred by chance. To realize the aspect of the multi-genomic event-based model, we have constructed a knowledge graph called "**Decision Networks(DN)**." The DN works on two-layer integration, where at first layer, we identify the linking parameters, such as Gene Symbol, CG IDs and Chr: Start-End. The detailed linked scenario is shown in Figure 2 of [12]. However, at this level integration behaved more like an enriched dataset. Instead of building a single integrated graph, we built a virtually integrated Knowledge graph for DNs. We achieved this by federated SPARQL query as mentioned in Listing 7 of our earlier work [12]. The second layer of integration is essential regarding defining the rules to extract the biological insights from multi-omics integrated DNs. Some of the conventional rules of filtering genes without significance to make visualization clinically actionable are as follows.

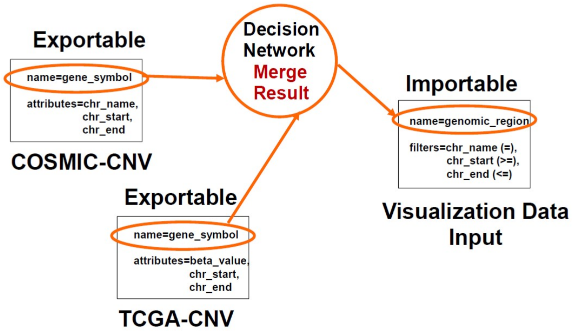


Fig. 3. Data input output model for Visualization

(i) Gene Expression and Methylation are reciprocal to each other. Which means if the gene is hyper-methylated it should be down-regulated.

(ii) A gene cannot be up- and down-regulated at the same time.

(iii) Functional annotation follows the central dogma of disease evolution where expression is captured first and then mutation, CNV, and Methylation, respectively.

(iv) Cancer is a heterogeneous disease, and any change in one genomic event is not sufficient to understand the mechanism.

(v) Beta-value in Methylation data where negative value represents Hypo- and a positive value represents Hyper- Methylation, respectively.

(vi) The CNV, the germline DNA for a given gene, can only be risk associated it falls outside the range of USCS defined gene length.

(vii) CNV for each cancer type changes based on two parameters, namely cancer are rare **frequency** and potentially confer high penetrance called as **odds ratios**.

(viii) Any pathways represented by the change in CNV, GE and Methylation will always be given a priority in studies and thus in visualization.

After filtering the data based on rules (i - viii), as mentioned above, the systems pre-process the data as shown in Figure 3. Figure 3 shows the key instances of input data, such as *Gene_Symbol, Chr, start, end.* The Decision Network layer we perform the integration and then visualize the filtered data. The result queried, and the filtered result can also be exported for further analysis. The use case was taken from E-GEOD-29431 - Identifying breast cancer biomarkers [3]. We have used the same genes for visualization and in Figure 1. Figure 1 shows the data types used in the study on visualization with various techniques. Whereas Figure 8 shows the solution on same gene as Figure 1.

[3] https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-29431/samples/

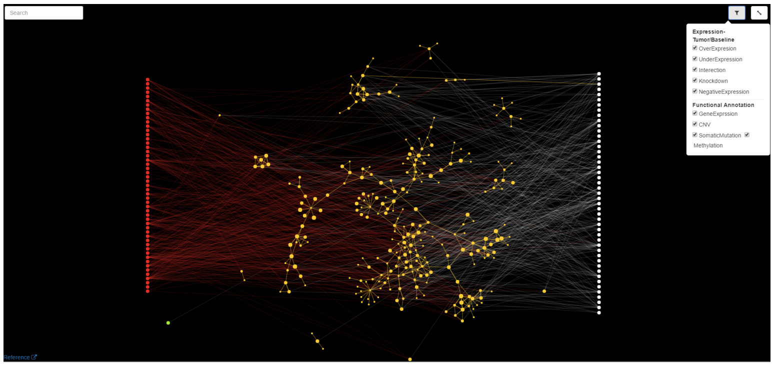
Functional Annotation

Integrated functional annotation is one of the key advantages associated with visual mining of GE data sets. We have used a semantic web approach to link distinct data sets from COSMIC, TCGA and ICGC. In comparison with existing data linking methods, our approach has linked data sets based on the semantics within the data. For example, we have extracted CNV, GE, Mutation and DNA Methylation data from TCGA, COSMIC, and ICGC and linked them to have enriched semantics, which in turn leads to having an improved coverage of the genome for each genomics profile. Each of these genomic signatures has its dedicated SPARQL endpoints. These SPARQL endpoints will be iteratively enriched with other associated similar data types to have maximum coverage of genome for each genomic profile. In the present paper, all differentially expressed genes from use case annotated using GE data from our Linked functional annotation platform [10].

Table 1. Genomics Data Statistics

No.	Data	Triples	Subjects	Predicates	Objects	Size (MB)
1	COSMIC GE	1184971624	148121454	18	148240680	10000
2	COSMIC GM	83275111	3620658	23	9004153	1400
3	COSMIC CNV	8633104	863332	10	921690	122
4	COSMIC Methylation	170300300	8292057	22	603135	2800
5	TCGA-OV	81188714	10974200	15	4774584	3774
6	TCGA-CESC	3763470	627652	43	481227	49557
7	TCGA-UCEC	553271744	19233824	91	68370614	84687
8	TCGA-UCS	1120873	183602	36	188970	10018
9	KEGG	50197150	6533307	141	6792319	4302
10	REACTOME	12471494	2465218	237	4218300	957
11	GOA	28058541	5950074	36	6575678	5858
12	ICGC	577 M	-	-	-	39000
13	CNVD	1,552,025	194,590	9	512,307	71

Table 1 shows the overall statistics of RDFization of COSMIC, TCGA and CNVD data and external (RDF) datasets used: rows 1-4 represent the number of triples and its size for COSMIC gene expression, gene mutation, CNV and Methylation data sets, respectively. Rows 5-8 represent the number of triples, and it is size for TCGA-OV, TCGA-CESC, TCGA-UCEC and TCGA-UCS data, respectively. The RDFization statistics for CNVD data are shown in row 13. Rows 9-12 represents the statistics of external datasets (available in RDF format), namely KEGG, REACTOME, GOA, and ICGC. To query data, we have used an adapted version of SAFE [15], a federation engine to query data from multiple endpoints in a policy-driven approach which may be a key element from the user while the user is selecting his/her hypothesis from visualization and unique functional annotation module based on the distributed concept in genomics.



Expression-Tumor/Baseline

☑ OverExpresion
☑ UnderExpression
☑ Interection
☑ Knockdown
☑ NegativeExpression

Functional Annotation

☑ GeneExprssion
☑ CNV
☑ SomaticMutation ☑ Methylation

Fig. 4. The GExpressionMap main view where the left side represents the lower and right side represents the higher gene expression

GExpressionMap

GExpressionMap has been built over a robust mathematical model of gene expression which defines that GE is linear and having a graph-based visualization for linear model provides the better visual representation of the events. In addition to visualization, we have built linked data based decision networks where we have contributed TCGA-OV, TCGA-UCS, TCGA-UCSC and TCGA-CESC (Methylation, CNV, Gene expression, and Complete Mutation) data along with COSMIC (GE, CNV, GM, and Methylation) and CNVD extending our earlier work [11–13]. These datasets will provide a platform for link identification and federation and addition to Linked Open Data.

GExpressionMap has been divided into four modules to identify critical challenges associated with GE data sets in biology. The first mode called as **Expression mode** talk about the conditional expression and track the changes in the property of transcripts or genes based on the changes in the expression level and identify the role on non-expressed genes in various cell cycle processes. Another mode called **knockdown mode** identifies the changes in various clusters representing a group or a biological process. This mode will also help to understand the effect from a knockdown to knockout. Knockdown studies are essential to solve various biological problems, such as a natural mechanism for silencing gene expression, specific inhibition of the function of any chosen target gene to understand the role in cancer and other diseases. Tracking these changes in the graph-based on motif building or destructing and cluster changes provides a visual impact to this biological discovery [19]. Another critical challenge while dealing with the group of genes or transcripts is to understand the pattern or bias of the network/data to understand the mechanism of the experiment. GExpressionMap provides an integrated annotated genes with their functional annotation and further cluster them based on their RPKM values means their expression pattern. By this way, the experimenter will conclude that how reliable is a cleave from a cluster what functional processes they are involved in and what can be cumulative effect reported from validated and patient data sources based on the linked functional annotation and GO annotations. This dimension of work is called **Annotation, Clustering and GO processes** mode. It is always crucial to find the strongest and weakest cluster based on matrices, such as the number of overexpressed genes connected with a cluster, number of underexpressed genes connected with a cluster or participation of individual gene in a cluster. On the other hand, if critical genes, such as TP53, EGFR, BRCA, and other biomarker is associated with large no of network or clusters can drive the progression in the disease like cancer. However, the number of over and under-expressed genes with this network will explain the functioning. This is how **Interaction and co-expression** mode have revealed the crux of the network. The aerial view of the expression map is depicted in Figure 4. Details of each mode explained in following subsections.

Expression Mode

Expression mode overlays the GE data either from microarray or RNA seq based on RPKM count from lower to the higher expression. As explained in Figure 4 red color bar demonstrated the gene with lower expression value whereas the white expression bar explains the value with higher expression value. The list of bubbles is the genes are either highly or lowest expressed based on their expression value. The expression scale in the bottom is the log scale which explains the range of expression considered maximum to minimum as RPKM/FPKM values. As it can be observed from Figure 6 that bottom expression line annotated as **D** is being used in such a way to have two-way side slider pointer. The major use of this approach is to identify the most significant genes since the expression value from RNA seq has a broader range mostly. The Value as mentioned in **A** explains about two types of values annotated as *OE*-Overexpressed and UE-Under-expressed. The value is constantly displayed as per the change from slider annotated on Figure 6 as **B**. The example of this has been shown as **C** where the for value 49848 expressed of *PSMD9* having been displayed. The overall impact of this mode would be to retain the ease of expression scale as in the case of Heatmaps, however covering the broad spectrum of the gene with added functionalities.

Knockdown Mode

Knock-down studies play a key role in biological experiments to understand the overall impact of a *gene* or *mRNA*. For example in cancer networks where if we consider one GE network contains the expression interaction from normal and adjacent normal called as normal sample expression network. Another network could be the expression network obtained from cancer tissues. Now to understand the behavior is important to understand the knockdown effect of most affected genes. As GExpressionMap also provides *bottleneck genes* based on cluster binding and a number of the associated cluster with that gene, the strength of the cluster. If a single gene has different expression level in both normal and cancer

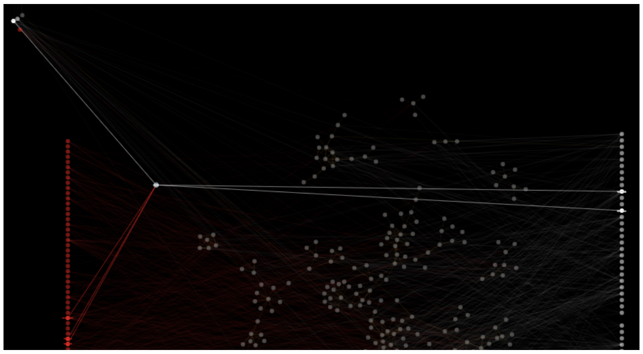


Fig. 5. A bottleneck view to understanding the effect of expression change and associations

network, it would be key to understand the impact of losing that gene and then understand the overall pattern of the network. Especially cancer network can get distorted after losing these bottleneck gene or highly expressed genes. A key observation such as the presence of certain genes with higher cluster binding in normal network however absence in cancer network can lead to key outcomes in cancer studies. Figure 5 provides a snippet of one such case. As explained in the figure knockdown of *PSMD9* will affect two genes from higher expression pole and two genes from a lower expression pole. Further, the cluster associated with it and having lower expression will have loss of connectivity and will cause insatiability in the network. This is a typical example of cancer progression and loss of connectivity in the cancer networks.

Interaction and co-expression mode

Dynamic changing property from normal to cancer networks reveals common system-level properties and molecular properties of prognostic genes across cancer types [27]. However current methods to generate co-expression network are basically for microarray data since they have been defined based on probe ids. This types of the network will not be able to cope-up to identify the changes in the broader level in co-expression networks [9]. This paper builds the co-expression network based on raw RPKM/FPKM values, or it can also accommodate expression value as log2 fold change values [17]. One of the key impacts of building a co-expression network using these expression counts is to bring similar associations or cell functions together after clustering. Usually in cancer networks, one of the major issues is to identify missing links and predict the fill-ins for the missing links. Since the RPKM values are experiment specific becomes essential to track the change and loss of expression for same tissue across different experiments. Building a co-expression network by this approach will automatically define the causality of the network if changes are abrupt. If a certain transcript is not at all expressed or lost the connectivity due to some treatment in any of the control would be easy to track. Apart from this differentially

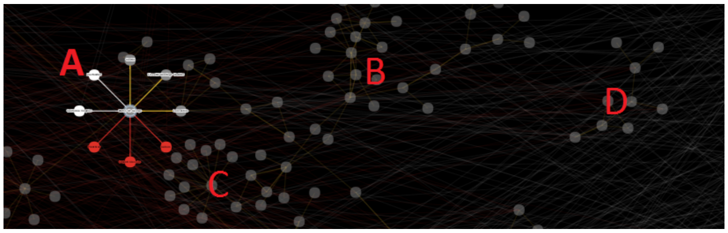


Fig. 6. GExpressionMap leveling and interacting partner association to visually mine functional annotations.

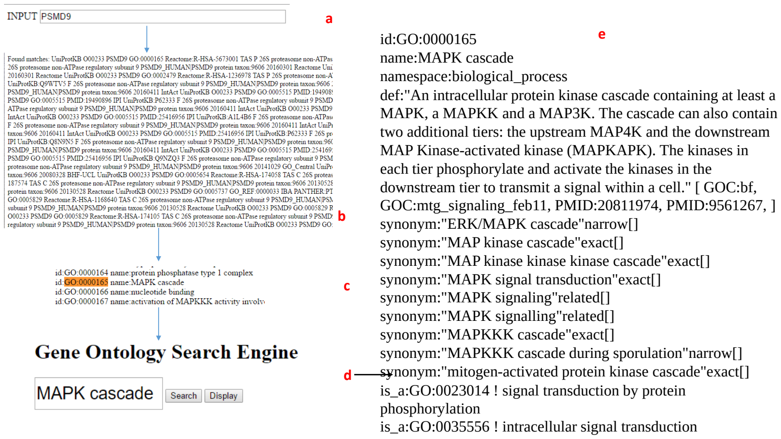
expressed genes could be easily extended to differentially expressed pathways based on co-expression network. This could be one of the potential outcomes. Figure 6 provides a glimpse of a co-expression network. One of the key points in this visualization is that it highlights the high expression network and keeps the less expressed network in light color annotations. Figure 6 clearly indicates that cluster A is highly expressed than B,C,D among these co-expressed networks.

Annotation, Clustering and GO processes mode
This mode of GExpressionMap involves the key features such as retrieval of *GO:ID* for a bottleneck gene identified based on clustering. To reduce the complexity in the visualization GExpressionMap have placed annotation based on user request. As depicted in Figure 6 where **C** indicates the bottleneck since holding three expression cluster. Now if the user is interested in functional annotation of this gene, they need to retrieve GO biological process and as mentioned in 4 as G clicking on this would provide a to an interface to obtain annotations as displayed in Figure 7. As we click on **G** of Figure 6 it takes to the **a** of Figure 7 and user need to enter the bottleneck gene obtained. Then interface queried a flat file for annotations [4]. This will display Go Ids and other Ids associated with the input query gene represented as **b** in Figure 7. Once we have obtained the GO Ids we have used gene ontology search engine obtained from [4] and embedded with our system. Then we query for obtained GO Id and outcome of some can be represented as **d** and **e** in Figure 7. This way we have contributed a web application with visualization to annotated the gene with associated expression visualization and identification of bottleneck gene or protein. Another key is to identification and understanding of clusters. One of such cluster based on our use case having been shown in Figure 8. The details of these clusters and associated methods will be discussed in the Result section.

Case Study, Results and Discussion

To demonstrate the feasibility of the proposed approach in biology, we have demonstrated a use-case from Monaco, Gianni, et al. [20]. This paper represents the comparative GE data between human and mouse. We have used

[4] https://github.com/zweiein/pyGOsite



INPUT PSMD9

```
Found matches: UniProtKB O00233 PSMD9 GO:0000165 Reactome:R-HSA-5673001 TAS P 26S proteasome non-ATPas
26S proteasome non-ATPase regulatory subunit 9 PSMD9_HUMAN|PSMD9 protein taxon:9606 20160301 Reactome Uni
20160301 Reactome UniProtKB O00233 PSMD9 GO:0002479 Reactome:R-HSA-1236978 TAS P 26S proteasome non-A
UniProtKB:Q9WTV5 F 26S proteasome non-ATPase regulatory subunit 9 PSMD9_HUMAN|PSMD9 protein taxon:9606
PSMD9_HUMAN|PSMD9 protein taxon:9606 20160411 IntAct UniProtKB O00233 PSMD9 GO:0005515 PMID:194908
PSMD9 GO:0005515 PMID:19490896 IPI UniProtKB:P62333 F 26S proteasome non-ATPase regulatory subunit 9 PSMD
ATPase regulatory subunit 9 PSMD9_HUMAN|PSMD9 protein taxon:9606 20160411 IntAct UniProtKB O00233 PSMD9
IntAct UniProtKB O00233 PSMD9 GO:0005515 PMID:25416956 IPI UniProtKB:A1L4B6 F 26S proteasome non-ATPas
F 26S proteasome non-ATPase regulatory subunit 9 PSMD9_HUMAN|PSMD9 protein taxon:9606 20160411 IntAct UniPr
taxon:9606 20160411 IntAct UniProtKB O00233 PSMD9 GO:0005515 PMID:25416956 IPI UniProtKB:P62333 F 26S pr
IPI UniProtKB:Q8N9N5 F 26S proteasome non-ATPase regulatory subunit 9 PSMD9_HUMAN|PSMD9 protein taxon:96
PSMD9_HUMAN|PSMD9 protein taxon:9606 20160411 IntAct UniProtKB O00233 PSMD9 GO:0005515 PMID:254169
PSMD9 GO:0005515 PMID:25416956 IPI UniProtKB:Q9NZQ3 F 26S proteasome non-ATPase regulatory subunit 9 PSM
proteasome non-ATPase regulatory subunit 9 PSMD9_HUMAN|PSMD9 protein taxon:9606 20141029 GO_Central UniPr
taxon:9606 20080328 BHF-UCL UniProtKB O00233 PSMD9 GO:0005654 Reactome:R-HSA-174058 TAS C 26S proteas
187574 TAS C 26S proteasome non-ATPase regulatory subunit 9 PSMD9_HUMAN|PSMD9 protein taxon:9606 2013052
protein taxon:9606 20130528 Reactome UniProtKB O00233 PSMD9 GO:0005737 GO_REF:0000033 IBA PANTHER:PT
GO:0005829 Reactome:R-HSA-1168640 TAS C 26S proteasome non-ATPase regulatory subunit 9 PSMD9_HUMAN|PSM
subunit 9 PSMD9_HUMAN|PSMD9 protein taxon:9606 20130528 Reactome UniProtKB O00233 PSMD9 GO:0005829 R
O00233 PSMD9 GO:0005829 Reactome:R-HSA-174105 TAS C 26S proteasome non-ATPase regulatory subunit 9 PSMD
regulatory subunit 9 PSMD9_HUMAN|PSMD9 protein taxon:9606 20130528 Reactome UniProtKB O00233 PSMD9 GO:
```

```
id:GO:0000164 name:protein phosphatase type 1 complex
id:GO:0000165 name:MAPK cascade
id:GO:0000166 name:nucleotide binding
id:GO:0000167 name:activation of MAPKKK activity involv
```

Gene Ontology Search Engine

MAPK cascade

```
id:GO:0000165
name:MAPK cascade
namespace:biological_process
def:"An intracellular protein kinase cascade containing at least a
MAPK, a MAPKK and a MAP3K. The cascade can also contain
two additional tiers: the upstream MAP4K and the downstream
MAP Kinase-activated kinase (MAPKAPK). The kinases in
each tier phosphorylate and activate the kinases in the
downstream tier to transmit a signal within a cell." [ GOC:bf,
GOC:mtg_signaling_feb11, PMID:20811974, PMID:9561267, ]
synonym:"ERK/MAPK cascade"narrow[]
synonym:"MAP kinase cascade"exact[]
synonym:"MAP kinase kinase kinase cascade"exact[]
synonym:"MAPK signal transduction"exact[]
synonym:"MAPK signaling"related[]
synonym:"MAPK signalling"related[]
synonym:"MAPKKK cascade"exact[]
synonym:"MAPKKK cascade during sporulation"narrow[]
synonym:"mitogen-activated protein kinase cascade"exact[]
is_a:GO:0023014 ! signal transduction by protein
phosphorylation
is_a:GO:0035556 ! intracellular signal transduction
```

Fig. 7. The Go ontology and functional annotation for the human-mouse model use case.

GExpressedMap to visualize this data and draw some of the key conclusions using visual representation. Based on the steps mentioned earlier, we have developed an expression map where Figure 8 represents one of the key clusters from this expression map. As we can observe from the diagram, human genes *A2M* have a close expression concerning mouse genes such as as*Aanat, Aadac, Amap, Abat, Abca1, and Aars*. Here, the key observation is that this cluster also holds other clusters and becomes bottleneck genes in human-mouse expression network. On the other hand, the only A2M human gene is underexpressed, and has a strong correlation with underexpressed genes in mouse (such as *Aanat, Aadac, Amap, Abat, Abca1*) as well as an overexpressed gene in mouse (such as *Aars*). One of the key outcomes of this cluster could be to identify detectable expression differences between species or individuals. The expression could logically divided into selectively neutral (or nearly neutral) differences and those underlying observable phenotypic [7]. To dig in further to identify the fact we have extracted the GO ids for each of the genes involved in the cluster. Where A2M highly associated with GO terms such as GO:0003824, GO:0004867, GO:0010951 and GO:0070062. Where, GO:0003824 is responsible for *catalytic activity* and has close correlation with GO:0003674, GO:0004867 associated with *serine-type endopeptidase inhibitor activity* and has close association with GO:0004866:*endopeptidase inhibitor activity*, whereas GO:0010951 and GO:0070062 are associated with negative regulation of endopeptidase activity and extracellular exosome respectively. To establish an association between human-mouse cluster, we have used the MGD [2] database, as the current version of GExpressionMap only supports *homospaiens*. The annotations for *Aanat, Aadac, Amap, Abat, Abca1* are a protein-coding gene which has the relation of *A2M*. These genes Aanat(cellular response to cAMP circadian rhythm, melatonin biosynthetic process, N-terminal protein amino acid acetylation), Aadac

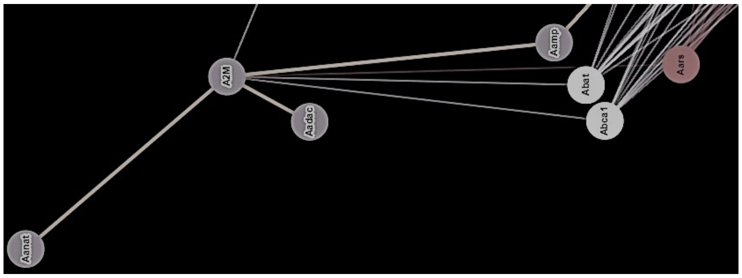


Fig. 8. Cluster representing diseasome for human-mouse

(carboxylic ester hydrolase activity, deacetylase activity, endoplasmic reticulum, endoplasmic reticulum membrane), Amap, Abat (aging, behavioral response to cocaine, catalytic activity, copulation), Abca1 (anion transmembrane transporter activity, apolipoprotein A-I binding, apolipoprotein A-I-mediated signaling pathway, apolipoprotein A-I receptor activity). Where the only highly expressed gene in mouse *Aars(alanine-tRNA ligase activity, cellular response to unfolded protein, skin development, tRNA modification)* having relation with *A2M.* Based on the biological process, this cluster represents *Membranoproliferative Glomerulonephritis, X-Linked Tangier Disease; TGD* and A2M are also involved with X-Linked Tangier Disease. In Summary, the visual identification of cluster, mapping of GE for each associated gene with the cluster, identification of expression level and functional annotation provides a key solution to how orthologs data with GExpressionMap have helped to mine the gene association to predict possible disease based on expression data. The proposed case study and results have just provided initial insight into a hidden treasure that can dig down visually using GExpressionMap. The expression extended for time series co-expression data where expression change happens on a certain time interval. For instance effect of ZIKA virus [22] where expression of top genes visualized for 12, 48 and 96 hours.

Conclusions

GExpressionMap is a key mechanism developed for visualization of gene expression data which is highly user-friendly, interactive, modular and visually informative. Integrated functional annotation, clustering, and co-expression network based on scientifically selected color annotations make it highly informative, usable and associative towards biological discovery based on genes expression.

Acknowledgment

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289, co-funded by the European Regional Development Fund.

References

1. Battke, F., Symons, S., Nieselt, K.: Mayday-integrative analytics for expression data. BMC bioinformatics 11(1), 121 (2010)
2. Blake, J.A., Richardson, J.E., Bult, C.J., Kadin, J.A., Eppig, J.T., Group, M.G.D., et al.: Mgd: the mouse genome database. Nucleic acids research 31(1), 193–195 (2003)
3. Chen, T., He, H.L., Church, G.M., et al.: Modeling gene expression with differential equations. In: Pacific symposium on biocomputing. vol. 4, p. 4 (1999)
4. Consortium, G.O., et al.: Gene ontology consortium: going forward. Nucleic acids research 43(D1), D1049–D1056 (2015)
5. Delgado, M.D., León, J.: Gene expression regulation and cancer. Clinical and Translational Oncology 8(11), 780–787 (2006)
6. Dietzsch, J., Gehlenborg, N., Nieselt, K.: Mayday-a microarray data analysis workbench. Bioinformatics 22(8), 1010–1012 (2006)
7. Dowell, R.D.: The similarity of gene expression between human and mouse tissues. Genome Biol 12(1), 101 (2011)
8. Heinrich, J., Seifert, R., Burch, M., Weiskopf, D.: Bicluster viewer: a visualization tool for analyzing gene expression data. In: Advances in Visual Computing, pp. 641–652. Springer (2011)
9. Hong, S., Chen, X., Jin, L., Xiong, M.: Canonical correlation analysis for rna-seq co-expression networks. Nucleic acids research 41(8), e95–e95 (2013)
10. Jha, A., Khan, Y., Iqbal, A., Zappa, A., Mehdi, M., Sahay, R., Rebholz-Schuhmann, D.: Linked functional annotation for differentially expressed gene (DEG) demonstrated using illumina body map 2.0. In: Malone, J., Stevens, R., Forsberg, K., Splendiani, A. (eds.) Proceedings of the 8th Semantic Web Applications and Tools for Life Sciences International Conference, Cambridge UK, December 7-10, 2015. CEUR Workshop Proceedings, vol. 1546, pp. 23–32. CEUR-WS.org (2015)
11. Jha, A., Khan, Y., Iqbal, A., Zappa, A., Mehdi, M., Sahay, R., Rebholz-Schuhmann, D.: Linked functional annotation for differentially expressed gene (deg) demonstrated using illumina body map 2.0. In: SWAT4LS. pp. 23–32 (2015)
12. Jha, A., Khan, Y., Mehdi, M., Karim, M.R., Mehmood, Q., Zappa, A., Rebholz-Schuhmann, D., Sahay, R.: Towards precision medicine: discovering novel gynecological cancer biomarkers and pathways using linked data. Journal of biomedical semantics 8(1), 40 (2017)
13. Jha, A., Mehdi, M., Khan, Y., Mehmood, Q., Rebholz-Schuhmann, D., Sahay, R.: Drug dosage balancing using large scale multi-omics datasets. In: VLDB Workshop on Data Management and Analytics for Medicine and Healthcare. pp. 81–100. Springer (2016)
14. Katz, Y., Wang, E.T., Airoldi, E.M., Burge, C.B.: Analysis and design of rna sequencing experiments for identifying isoform regulation. Nature methods 7(12), 1009–1015 (2010)
15. Khan, Y., Saleem, M., Iqbal, A., Mehdi, M., Hogan, A., Ngomo, A.C.N., Decker, S., Sahay, R.: Safe: Policy aware sparql query federation over rdf data cubes. In: SWAT4LS (2014)
16. Khomtchouk, B.B., Van Booven, D.J., Wahlestedt, C.: Heatmapgenerator: high performance rnaseq and microarray visualization software suite to examine differential gene expression levels using an r and c++ hybrid computational pipeline. Source code for biology and medicine 9(1), 1 (2014)

17. Kommadath, A., Bao, H., Arantes, A.S., Plastow, G.S., Tuggle, C.K., Bearson, S.M., Guan, L., Stothard, P.: Gene co-expression network analysis identifies porcine genes associated with variation in salmonella shedding. BMC genomics 15(1), 1 (2014)
18. Metsalu, T., Vilo, J.: Clustvis: a web tool for visualizing clustering of multivariate data using principal component analysis and heatmap. Nucleic acids research 43(W1), W566–W570 (2015)
19. Mocellin, S., Provenzano, M.: Rna interference: learning gene knock-down from cell physiology. Journal of translational medicine 2(1), 39 (2004)
20. Monaco, G., van Dam, S., Ribeiro, J.L.C.N., Larbi, A., de Magalhães, J.P.: A comparison of human and mouse gene co-expression networks reveals conservation and divergence at the tissue, pathway and disease levels. BMC evolutionary biology 15(1), 259 (2015)
21. Segal, E., Yelensky, R., Kaushal, A., Pham, T., Regev, A., Koller, D., Friedman, N.: Genexpress: a visualization and statistical analysis tool for gene expression and sequence data. In: Proceedings of the 11th International Conference on Intelligent Systems for Molecular Biology (ISMB). vol. 18 (2004)
22. Singh, P.K., Khatri, I., Jha, A., Pretto, C.D., Spindler, K.R., Arumugaswami, V., Giri, S., Kumar, A., Bhasin, M.K.: Determination of system level alterations in host transcriptome due to zika virus (zikv) infection in retinal pigment epithelium. Scientific reports 8(1), 11209 (2018)
23. Tang, C., Zhang, L., Zhang, A.: Interactive visualization and analysis for gene expression data. In: System Sciences, 2002. HICSS. Proceedings of the 35th Annual Hawaii International Conference on. pp. 9–pp. IEEE (2002)
24. Weniger, M., Engelmann, J.C., Schultz, J.: Genome expression pathway analysis tool–analysis and visualization of microarray gene expression data under genomic, proteomic and metabolic context. BMC bioinformatics 8(1), 179 (2007)
25. Wu, C., Zhu, J., Zhang, X.: Integrating gene expression and protein-protein interaction network to prioritize cancer-associated genes. BMC bioinformatics 13(1), 182 (2012)
26. Xia, J., Lyle, N.H., Mayer, M.L., Pena, O.M., Hancock, R.E.: Invex—a web-based tool for integrative visualization of expression data. Bioinformatics 29(24), 3232–3234 (2013)
27. Yang, Y., Han, L., Yuan, Y., Li, J., Hei, N., Liang, H.: Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. Nature communications 5 (2014)
28. Yeung, K.Y., Ruzzo, W.L.: Principal component analysis for clustering gene expression data. Bioinformatics 17(9), 763–774 (2001)
29. Yoshida, R., Higuchi, T., Imoto, S., Miyano, S.: Arraycluster: an analytic tool for clustering, data visualization and module finder on gene expression profiles. Bioinformatics 22(12), 1538–1539 (2006)