GenExSt: A Tool to Identify Correlation of Gene Expression after Normalization with Housekeeping Genes

Oliver Bonham-Carter[1] and Yee Mon Thu[2]

{obonhamcarter[1], ythu[2]}@allegheny.edu
Department of Computer Science[1]
Department of Biology[2]
Allegheny College
520 N. Main Street
Meadville, PA 16335
https://allegheny.edu/
Project Link: https://github.com/developmentAC/genExSt

Abstract. Interaction between genes is one driving force that can influence a biological outcome. In a genetic disease such as cancer, understanding genetic interactions may help us elucidate mechanisms sustaining cancer growth. A computational approach is one way to detect genetic interactions in the context of cancer. In this article, we introduce a tool, *GenExSt*, and its underlying method to study gene interactions.

We applied our method to discover gene-pairs whose expressions demonstrate patterns of correlation. For this demonstration, we selected ten breast cancer gene expression data sets from the Genomic Data Commons Data Portal through National Cancer Institute. We focused on genes that suppress genome instability, or instability suppressing genes (GIS), many of which play an important role in cancer.

We applied our method to an inter-comparison across data sets. Here we tested statistical normalization approaches derived from the combined expressions of randomly selected, single, housekeeping (HK) genes, and from the calculated mean of three expressions. In addition, our method derives R^2 values from linear models in which the expressions of all possible pairs of GIS genes are placed in a linear model to produce heatmaps to indicate probable correlations. We show that results from our method are suited to normalized data, extracted from multiple genes simultaneously, rather than using single gene expression values. GenExSt may be used to study gene expression data in other settings provided that the concept of gene interactions is appropriate in the context.

Keywords: Analysis by R-Squared data, normalization, cancer, gene expression, genome instability genes, housekeeping genes, normalization analysis

1 Introduction

1.1 Stability and Pathways for Disease Onset

The human genome sustains an incredible level of stability given a myriad of extrinsic and intrinsic insults that create alterations and potentially compromise the stability. Instability, at a small-scale level, could be beneficial since this is the source of genetic diversity. However, genome instability, at a large-scale level, may inadvertently introduce changes within the genome, which could result in deleterious outcomes. In multi-cellular organisms, genome instability, for example, is a precursor for genetic diseases such as cancer.

Fortunately, multiple cellular mechanisms are in place to ensure that most forms of acquired damage do not perpetuate over multiple generations. These mechanisms may be generally categorized into three rubrics; (1) pathways that promote repair, (2) pathways that prevent the inheritance of unintended alterations, and (3) pathways that ensure cellular destruction, in the case that the acquired damage is irreparable. Genes that are involved in these processes can collectively be designated as *genome stability genes*, as they serve to reduce the disorder, to which biological systems are allergic. Not surprisingly, many genes of genomic stability exhibit changes in expression or function in multiple cancers and it is perhaps the recognition of this unrest that ailments may be discovered.

Discussed in [1], disease onset could be described by inherent biological pathways that may be discovered during a study of gene expression. Similarly, pathways in cancer may allow for its survival and be studied by expression data. For instance, changes in gene expressions provide a survival advantage for cancer cells since they foster genetic diversity within the population [2]. Nevertheless, how individual cancer cells thrive with the burden of a compromised genome is perplexing since multiple complex and redundant pathways exist to hinder propagation of the damaged genome. While healthy cells may undergo cell death (i.e., apoptosis) or cell cycle arrest when genome integrity is compromised, cancer cells may evolve to prosper under the same conditions. These observations suggest that cancer cells, having acquired mechanisms to fend off the challenge of genome instability, have been selected during micro-evolution (i.e., during the tumor's life cycle). High entrophy in biological pathways is not likely to cultivate growth, whether for health, or in the case of cancer, since such growth requires some form of coordinated energy and nutrition to produce mass. Therefore, while cancer is able to find opportunity in unstable biological systems, it still requires some level of stability to maintain its course.

Simultaneously, it is implied that cancer cells may become dependent on these pathways to survive the genome instability. We imply that identifying these pathways may likely reveal vulnerabilities of cancer. This approach has been successfully implemented in the design of targeted cancer therapies. For instance, shown in Figure 1-A, breast cancer cells with a mutation in DNA repair genes, *BRCA1* or *BRCA2*, are more sensitive to the inhibition of another repair pathway, mediated by the protein product of *PARP1*, than healthy cells without the mutation [3]. This sensitivity is as a result of *BRCA1/2* mutant

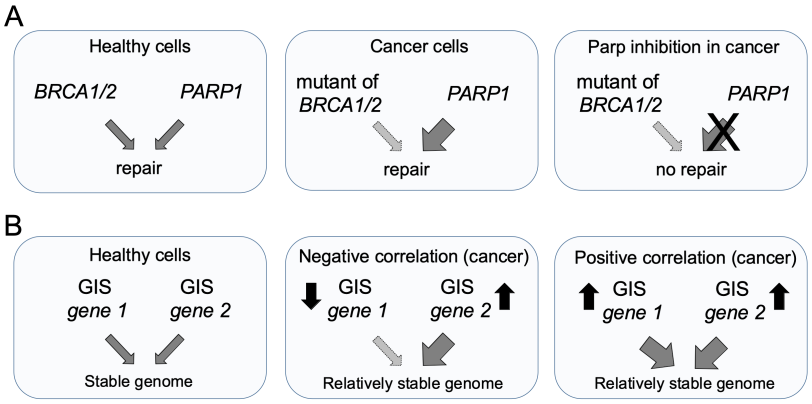

Fig. 1. Genetic interactions between genes responsible for DNA repair or genome stability. (A) Cancer cells with mutated *BRCA1/2* genes become more dependent on the function of the *PARP1* protein product. As a result, they are more vulnerable to inhibition of *PARP1* compared to healthy cells, in which the protein products of *BRCA1/2* genes are functional [3]. (B) A working model to illustrate the genetic dependencies in cancer. In cancer cells, physiological expression levels of GIS genes may be altered (indicated by the up- or down-arrow). A change in the expression of one gene may prompt the dependency of another GIS gene, resulting in a proportional increased or decreased expression of the latter. The nature of this relationship may be manifested in a positive or negative correlation between the expressions of two genes.

cells' dependency on the function of *PARP1*. Other important relationships have become known about these genes [4]. Identify additional vulnerabilities of cancer cells, with mutations that compromise genome stability, will help to encourage novel therapeutic approaches.

1.2 Expression Analysis

To identify cancer cell vulnerabilities, we examined the correlation of gene expressions using a large-scale approach. We reasoned that expression of two genes would positively or negatively correlate if the change in the expression of one gene necessitated the function of another, shown in Figure 1-B.

Due to the immense complexity of the study of expression in the genome, many computational methods, reviewed in [5] are necessary. The use of computational methods, where complex and voluminous amounts of data is involved, is seemingly the only approach for conclusive study, as discussed in [6] (differential analysis), [7] (signature study), [8] (expression profiling), [9] (cancer-specific correlations) and other studies mentioned in [10]. For our own study of diverse pathways, there was no substitution to the application of automated methods. We therefore devised our own method and Tool, called *GenExSt*, to handle our requirements for analysis.

1.3 An Overview of *GenExSt*

Development in software to assist in normalized gene expressions analysis is readily drawing much attention in Bioinformatics; [11] (bayNorm) and [12] (PEATmoss) are such current projects.

In our own work, we were forced to create original software from code. Written in Python using the Streamlit library (https://www.streamlit.io/) our tool, *GenExSt* (read, "Gene-next"), allows users to determine correlations between GIS gene pairs after normalizing the expression of these genes to the average expression of selective housekeeping (HK) genes.

Shown in Figure 2, *GenExSt* supports interaction with a browser tab for convenient tests of data. Since a browser-centric approach to using *GenExSt* may not lend itself to processing several data sets in tandem, we offer a command-line and parameter-driven version of the tool, shown in Figure 3, to further enable automated processes. In both versions, the analysis is the same.

The method and tool supports automatic normalizations of HK gene expression data for *GenExSt* to generate linear models from gene pairs. Normalization in this work is detailed in the Methods Section and is used to mathematically transform expression values for comparison purposes. The linear models (discussed in Section 2) provide R^2 values and are shown in heatmaps to suggest potential correlations.

We have made *GenExSt* available as an open source project on GitHub (https://github.com/developmentAC/genExSt) for investigators to use the code, or to modify for customized applications.

1.4 Housekeeping Genes

Central to this work are the HK genes which exhibit expression stability, even under diverse conditions and may be conveniently compared with analogous values obtained from diverse samples. Conversely, non-HK genes have varying expression values and are not necessarily compatible between diverse samples. The gene, *GAPDH* is a HK gene and its expression value is considered to be stable even when exposed to cellular stress factors [13]. Other well-known HK genes exhibiting minor variation are *TUBB* and *TUBA1A*. The predictability of their expression values make HK genes particularly well suited as controls in comparative analyses for which we designed *GenExSt*.

Due to the multitude of HK genes that have been uncovered in biological systems, *GenExSt* allows these genes to be conveniently inputted into analysis projects as single values or in groups of genes for normalization. In addition to inputting HK genes, GIS genes may also be inputted into *GenExSt* as arbitrarily-sized groups to facilitate large-scale projects.

1.5 Heatmaps

We note that the results of *GenExSt* are suggested correlations of GIS gene pairs. We used heatmaps to exhibit results such as the one shown in Figure

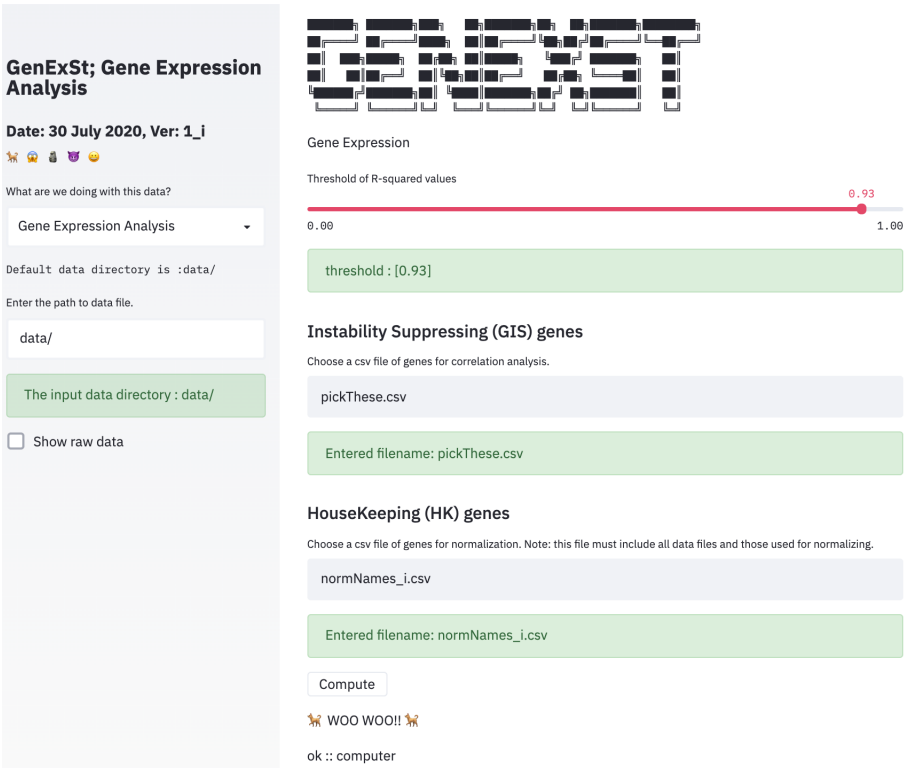

Fig. 2. A screenshot of our method and tool, *GenExSt*, for the study of normalized HK gene expressions and correlation of GIS genes using an analysis of R^2 values from linear models of gene pairs. Heatmap visualizations are created to visually suggest potential correlations between GIS genes. The threshold slider allows for the selection of an upper-bound of R^2 values to include in the outputted heatmaps.

4, in which each cell is color-coded according to the R^2 value obtained from the linear model of the gene pairs at the cross section. In the supplied figure, genes names are provided by their Ensemble IDs, however, during the below experiment we discuss to showcase functionality, the gene names are represented human-readable form. In Figure 4, we note that there are two cells of lighter colors to indicate that the genes at their cross sections may potentially correlate, according to the raised R^2 values from their linear models. The cells of darker colors indicate that the genes at their cross sections do not have such elevated R^2 values and are not suggested to have a correlation by this method. The heatmap graphics were prepared by *GenExSt* using the Plotly library in Python [14].

```
--------------------------------------------------------------------------------
 24 July 2020 | version: 2_ii | Oliver Bonham-Carter | obonhamcarter@allegheny.edu
--------------------------------------------------------------------------------

      The GeneExPy program to perform linear regression over GDP datasets.

      Library installation notes:
      plotly:
      pip3 install plotly, or try running python3 -m pip install scikit-learn
      scikit-learn:
      python3 -m pip install scikit-learn, maybe necessary: pip3 install scipy

      + 😀  USAGE: programName <any key to launch>
      + INPUT directory: (your data files are here)     :  data/
      + OUTPUT directory: (your output is placed here)  :  0out/

      + Note:
         Use parameter <<heatmap>> or <rsqu>>> to ensure that
         Plotly and the statistical libraries have been correctly installed.

       Note: the data directory cannot handle subdirectories holding data. Please
      place the text files into this data directory without using subdirectories.
```

Fig. 3. A screenshot of the command-line version of *GenExSt* to facilitate fully automated processes.

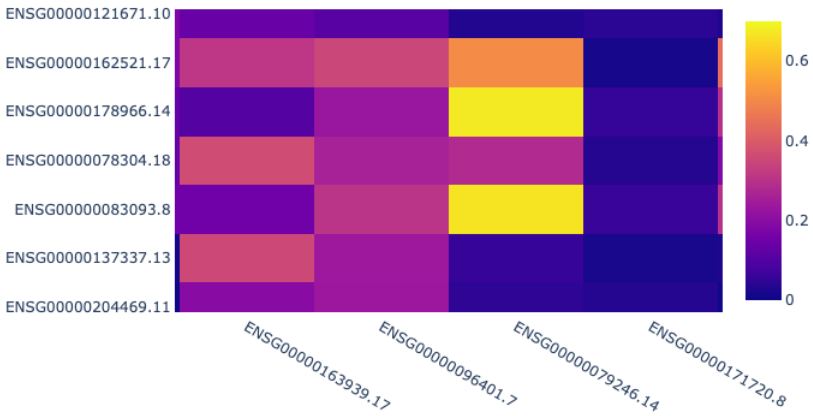

Fig. 4. A heatmap from *GenExSt* suggests potential correlations based on R^2 values. Each cell is the cross-section of the gene pairs, placed in a linear model. Here we label the heatmap with the Ensemble IDs of genes.

2 Methods

The *Transcriptome Profiling* data for this project was downloaded from the Genomic Data Commons (GDC) Data Portal, (https://portal.gdc.cancer.gov/) available from the National Cancer Institute (NCI). From these data sets, we

selected a subset of genes that suppress genome instability (GIS genes), as categorized in the study by Putnam *et al.* [15]. These genes were regarded in the author's article as GIS genes since perturbation in the function of these genes lead to small-scale or large-scale changes within the genome.

We used FPKM data sets to show how to determine the existence of a positive or negative correlation between the expression of two given GIS genes in cancer, as shown in Figure 1-B. These correlations can reveal if two GIS genes coordinate, or if an alteration in the expression of one GIS gene may increase the dependency of cancer cells on another GIS gene, as seen in Figure 1-B. Co-expression between pairs of genes has been studied in cancer cell lines [16]. However, to our knowledge, a similar type of analysis has not been performed using gene expression data from cancer tissues.

To exhibit the functionality of *GenExSt*, we report a preliminary data on co-expression of pairs of GIS genes using breast cancer gene expression data. We observed some consistent correlations between certain pairs although different groups of HK genes were used for normalization. In the future, our goal is to expand the number of data sets in order to identify pairs of genes that exhibit consistent co-expression. Furthermore, it would be beneficial to determine the reproducibility of these correlations. Below we discuss the mechanism of the tool and then proceed to the results.

The data that we selected for our work was arbitrarily chosen for its association with Breast Invasive Carcinoma and its affiliation with the GDC's "TCGA-BRCA" listing, as specified by the GDC website. In Table 1, we display the ID's and file names of the individual sets used for the current work. Each of our chosen sets were from GDC's *Transcriptome Profiling* category, of type *Gene Expression Quantification* and RNA-Seq experimental strategy.

We defined our *superset* of gene expressions to imply the complete listing of Ensemble IDs originating from each file in our data set of Table 1. Each file contained exactly the same listing of Ensemble IDs and their associated gene expression values, which vary from data set to data set.

From the human expression data described in Table 1, we defined a subset of 273 GIS genes (i.e., genes that are responsible for maintaining integrity of the genome), which were the focus for our study. Note, these genes were entered into *GenExSt* by the CSV file *pickThese.csv*, as shown in Figure 2. The GIS genes were obtained from the study by Putnam *et al.* [15]. This subset was important for two general reasons, (1) expression of these genes directly addressed our research question mentioned in Introduction, and (2) we determined that there was too much gene diversity to cover in the total data (60484 genes total) to permit a concentrated study.

Due to the inherent noise and other sources of error associated with the quantification and collection of expression data, normalization is necessary to allow for studies of comparison. The overarching goal of our study was to determine normalizing factors that make it possible for us to discover pairs of GIS genes which show consistent correlation. To this end, we studied single and multiple gene expressions to be used to normalize data, as discussed below.

Table 1. The gene expression data downloaded from the GDC Data Portal available from NCI. The ID and the data set make up the directory and filename of the data set (i.e., `ID/data set.ext`.) In this work, each set is labeled by the index d, for $1 \leq d \leq 10$.

Set	**ID**/*Filename*
1	**10e813bc-63fe-4f32-a000-88527d5444b2/** *e1f960fc-6efa-4492-9f44-8e645ec8bfc2.FPKM.txt.gz*
2	**1d6732d2-c627-4883-b2a3-c822a5023a06/** *a6f657a3-439e-4afd-be04-413f2ed02828.FPKM-UQ.txt.gz*
3	**1da554d3-d50d-4210-886a-b750b33cf4de/** *93904035-e387-47e3-90f0-ea0d5f726dac.FPKM-UQ.txt.gz*
4	**2237e742-56da-4688-904e-8caaba7b1831/** *7d9d7119-4a9e-46f6-b758-153bfcc869a5.FPKM.txt.gz*
5	**27c15e58-e219-432c-a286-9835a9f9bdff/** *00511204-3512-4a5e-b664-60271e968903.FPKM.txt.gz*
6	**5c480d48-8117-48f1-b6d4-c8b420769a43/** *a154c740-f4b4-4b62-a4a8-2273f8f834b1.htseq.counts.gz*
7	**89e04995-ee7b-4bba-b964-2c53a389e5e1/** *bfe796aa-7f29-49d2-a72f-79a9acf12365.htseq.counts.gz*
8	**9caffd9b-0aa8-4aac-8a0d-598267a9f293/** *d578e27f-537c-4aaa-8903-6ffe68346276.FPKM.txt.gz*
9	**e011d682-e543-4f6f-95d2-e86cce259644/** *e29ce54a-49a1-47a8-82fd-7687cec0d1bb.FPKM.txt.gz*
10	**e4fba51b-cab5-4714-a0b3-9aa1e3786a27/** *cbf61237-f5ba-4fa0-97be-e79e8e75a887.FPKM.txt.gz*

2.1 Single Expression Normalization

Our first step was to study the possibilities of normalization using the expression of single genes as inputs in our calculations. For this, we arbitrarily selected the following three HK genes; *TUBB* (ENSG00000196230.11), *TUBA1A* (ENSG00000167552.12) and *GAPDH* (ENSG00000111640.13).

Table 2. The legend of variable names of the normalizing factors that were used to normalize the gene expressions. For each set, these variables were created to be used to normalize all expressions of the set.

Variable	**Expr. in Norm. Fact.**
HKG_{Tubb}	Single expression of *TUBB*
HKG_{Tuba1a}	Single expression of *TUBA1A*
HKG_{Gapdh}	Single expression of *GAPDH*
$HKG_{AvgTubbTuba1aGapdh}$	Avg of three genes
HKG_{AvgG1}	Avg of ten genes in group 1
HKG_{AvgG2}	Avg of ten genes in group 2
HKG_{AvgG3}	Avg of ten genes in group 3

Table 3. The variance and mean of values described in R^2 heatmaps across normalizing factors. The heatmaps were designed to illustrate trends in the numbers of gene-gene correlations according to their elevated R^2 values. Too many high mean values (as noted in the single gene) are biologically suspect and we noted that when ten genes were used to create the normalizing factors, there were fewer gene-gene correlations (lower means). These results are in-keeping with the natural high-complexity to make such gene-gene interactions possible in biology [17].

Normalizing Factor	**Mean**	**Variance**
HKG_{TUBB}	0.28617	0.06526
HKG_{Tuba1a}	0.71638	0.08067
HKG_{GAPDH}	0.55054	0.12963
$HKG_{AvgTubbTuba1aGapdh}$	0.26050	0.05792
HKG_{AvgG1}	0.22230	0.05274
HKG_{AvgG2}	0.19935	0.04645
HKG_{AvgG3}	0.20610	0.04732

In each of our data sets, the gene expression values for selected genes were obtained and applied to the normalization of the set (i.e., those featured in Table 1) from which the particular expression values were obtained. The names of our normalizing variables follow the convention set-out in Table 2. In this table, one will note that the expression of a HK gene (i.e., *TUBB*, *TUBA1A* or *GAPDH*) in each data set is used as a normalizing factor for that data set. In other words, for single expression normalization, each data set has three normalizing factors, that are unique to that set.

For example, the normalizing factor that was created for each data set using the *TUBB* gene expression, can be written, HKG_{Tubb} and similarly for the other genes, *TUBA1A* (HKG_{Tuba1a}) and *GAPDH* (HKG_{Gapdh}). Each data set featured in Table 1 has its own specific HKG_{Tubb}, HKG_{Tuba1a} and HKG_{Gapdh} values.

Using these values, we calculated a normalizing expression for each GIS gene (total N = 272) for each data set in Table 1. The natural log of the normalized values, denoted by, y_n, of all GIS genes ($gene_n$ for $1 \leq n \leq N$) were calculated from expression values (denoted by x_n) using the normalizing factors featured in Table 2. For example, the equation for normalized expressions in a set using *TUBB* as a normalizing factor (HKG_{Tubb}) was the following.

$$y_n \leftarrow log\,(x_n / HKG_{Tubb})$$

With log-transformation, we were able to achieve normal distribution of values, that may otherwise be skewed [18].

2.2 Multiple Expression Normalization

Three genes: In an effort to diversify the normalizing factors for each data set featured in Table 1, we used the average of the expressions of the three

genes described above (i.e., *TUBB*, *TUBA1A* or *GAPDH*) instead of expression of individual genes. For example, described in Figure 5, the natural log of the normalized value (denoted by y_n) for each GIS gene ($gene_n$ where $1 \leq n \leq N$) was obtained in each data set by the following equation. We note that *GenExSt* handled this calculation automatically.

$$y_n \leftarrow log\,(x_n / HKG_{AvgTubbTuba1aGapdh})$$

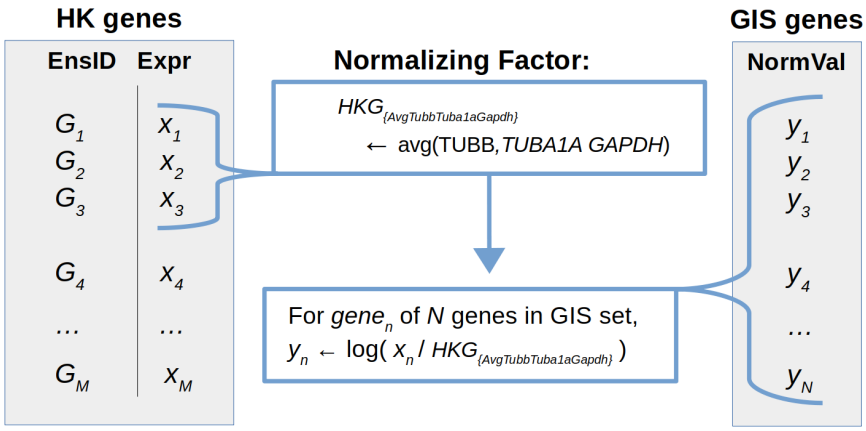

Fig. 5. Each expression value was normalized by *GenExSt* according to an average derived from selected gene expression values in the same data set. This average was calculated in each data set using the selected Ensemble IDs.

Ten genes: For this approach, we used the average of ten gene expressions to obtain normalizing factors for each set shown in Table 1. We speculated, that the average of ten genes, chosen for their functional attributes, would create a normalizing factor that would better represent the diversity of HK genes.

As part of this approach, we first identified a list of approximately 3800 HK genes, whose expressions are relatively uniform across multiple tissues [19]. This is an important criteria since we intend to use this normalization method for comparison of gene expression derived from cancers of different tissue of origins. HK genes from this list were subject to gene ontology analysis (http://geneontology.org/) so that we could systematically categorize them based on cellular components to which they belonged.

After sorting according to fold enrichment values, the top ten non-redundant cellular components were chosen and one representative gene from each functional category was randomly selected to build a list of HK genes. The same process was repeated to generate two additional lists of HK genes. We identify the selected HK genes and their ontologies in Tables 4, 5, and 6. We speculated that we would get consistent results if ten genes that we selected correctly represented the diversity of HK genes. The average expression of ten genes in each list

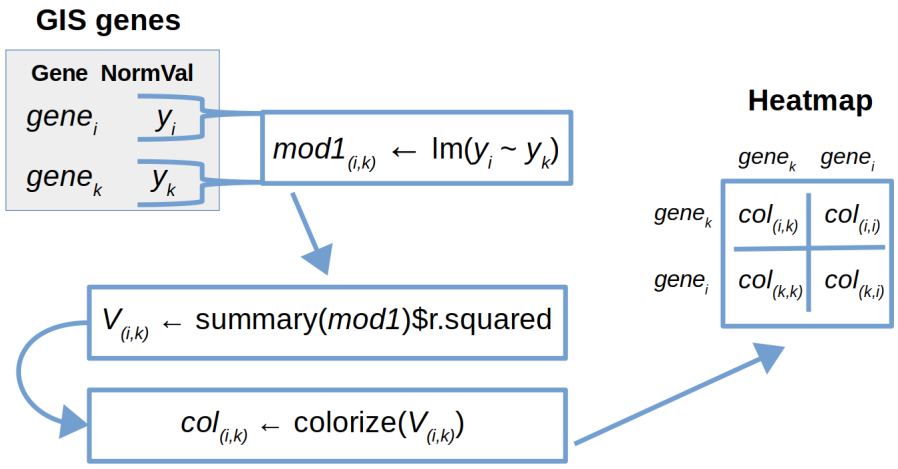

Fig. 6. Collected, are the R^2 values from linear models where the normalized expression quantity of one gene is regressed over another. We note that $gene_i$ and $gene_k$ represent all normalized values obtained across the data sets of Table 1. Heatmaps are generated from R^2 values gained from linear models. The values have been colorized according to their placement in the natural range of R^2 (i.e., $[0 \leq R^2 \leq 1]$.)

was determined, denoted by, HKG_{AvgG1}, HKG_{AvgG2} and HKG_{AvgG3}. These three averages were determined to obtain three normalizing factors unique for each set in Table 1.

Using HKG_{AvgG1} unique for each data set, the natural log of normalized values (denoted by y_n) for each GIS $gene_n$ where $1 \leq n \leq N$) was obtained by the following equation.

$$y_n \leftarrow log\,(x_n / HKG_{AvgG1})$$

The same procedure was performed using corresponding HKG_{AvgG2} and HKG_{AvgG3}.

2.3 Linear models

When the gene expressions of a data set are normalized by a single or multiple expression normalizing factor, we determined correlation of expression between all possible pairs of genes within the GIS list. Correlation is represented by R^2 values, generated from a test where the normalized expression values of GIS genes underwent an *all-against-all* linear model test. Our method and tool, *GenExSt*, automatically creates the linear models of gene pairs. To visualize the whole data set, we used heatmaps in which R^2 values are represented as a gradient of colors. The R^2 statistical metric is a measurement of proximity of data points to the fitted regression line. Also known as a "coefficient of determination", this statistic describes the percentage of the response variable variation, as explained by a linear model and is on a scale that ranges from 0 to 100 percent. The values

Table 4. The first set of HK genes for multiple expression normalization. In the group, the expression values of all ten were averaged in each set d to determine its normalizing factor. Here, we provide the ontology group, Uniprot ID, *human* gene ID and Ensemble IDs.

Ontology	**UnitProtKB**	**Gene ID**	**EnsNum**
Proteasome	Q9Y5K5	*UCHL5*	ENSG00000116750
Ribosome	Q96EL2	*MRPS24*	ENSG00000062582
ER	Q9P0I2	*EMC3*	ENSG00000125037
ER targeting	O76094	*SRP72*	ENSG00000174780
splisosome	O15234	*CASC3*	ENSG00000108349
Prp19 complex (splicing)	Q9BZJ0	*CRNKL1*	ENSG00000101343
splisosomal snRNP complex	P62310	*LSM3*	ENSG00000170860
Proteasome	P62195	*PSMC5*	ENSG00000087191
Methylsome	Q9BQA1	*WDR77*	ENSG00000116455
Translation	Q9UBQ5	*EIF3K*	ENSG00000178982

Table 5. The second set of HK genes for multiple expression normalization.

Ontology	**UnitProtKB**	**Gene ID**	**EnsNum**
Proteasome	Q9NRR5	*UBQLN4*	ENSG00000160803
Ribosome	Q9H2W6	*MRPL46*	ENSG00000259494
ER	Q8N766	*EMC1*	ENSG00000127463
ER targeting	P37108	*SRP14*	ENSG00000140319
splisosome	P38919	*EIF4A3*	ENSG00000141543
Prp19 complex (splicing)	Q99459	*CDC5L*	ENSG00000096401
splisosomal snRNP complex	Q53GS9	*USP39*	ENSG00000168883
Proteasome	Q06323	*PSME1*	ENSG00000092010
Methylsome	O14744	*PRMT5*	ENSG00000100462
Translation	Q7Z478	*DHX29*	ENSG00000067248

Table 6. The third set of HK genes for multiple expression normalization.

Ontology	**UnitProtKB**	**Gene ID**	**EnsNum**
Proteasome	P14735	*IDE*	ENSG00000119912.14
Ribosome	Q96BP2	*CHCHD1*	ENSG00000172586.7
ER	Q9BV81	*EMC6*	ENSG00000127774.6
ER targeting	P49458	*SRP9*	ENSG00000143742.11
splisosome	Q9HCG8	*CWC22*	ENSG00000163510.12
Prp19 complex (splicing)	P11142	*HSPA8*	ENSG00000109971.12
splisosomal snRNP complex	O43447	*PPIH*	ENSG00000171960.9
Proteasome	Q15008	*PSMD6*	ENSG00000163636.9
Methylsome	Q99873	*PRMT1*	ENSG00000126457.19
Translation	Q7L2H7	*EIF3M*	ENSG00000149100.11

in our heatmaps range between 0 and 1, in keeping with the lower- and upper-bounds, respectively, of their R^2 values. The mean and variance values were collected for each normalizing factor and are shown in Table 3.

In Figure 6, we describe the steps to create heatmaps from the R^2 values of linear models between pairs of genes. The generation of heatmaps from normalizations using normalizing factor values (i.e., single or multiple gene expressions) was similar across all prepared figures. Our heatmap graphics were prepared by *GenExSt* using the Plotly library in Python [14].

3 Results and Discussion

In this study, we developed a Tool and method called *GenExSt* that was applied to a study where normalization allowed us to identify consistent correlation between expression of two GIS genes. Our tool outputted heatmaps which allowed us to visually study our genes for correlations by locating lighter colors for gene pairs in resulting heatmaps.

We describe in this section how we used *GenExSt* to conduct an experiment and analysis. Ten gene expression data sets generated by RNA-sequencing of breast cancer tissues were randomly chosen for our data analysis. We employed different new normalization approaches to compare gene expression across multiple samples, as detailed in the Methods section.

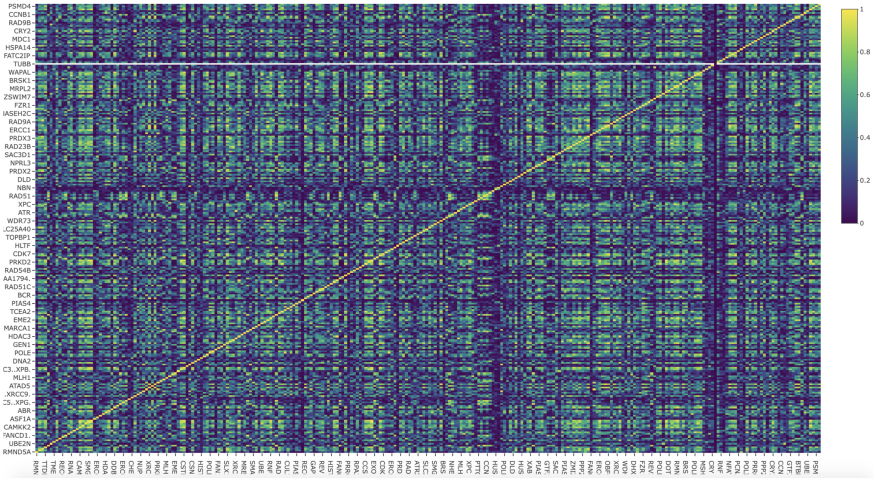

Fig. 7. Heatmaps of R^2 values, derived from the normalizing factor, HKG_{Tubb}.

When the *TUBB* expression was used for normalization, almost half of the gene pairs showed correlations, as shown in Figure 7. However, this level of codependency between genes is biologically improbable, suggesting that most of

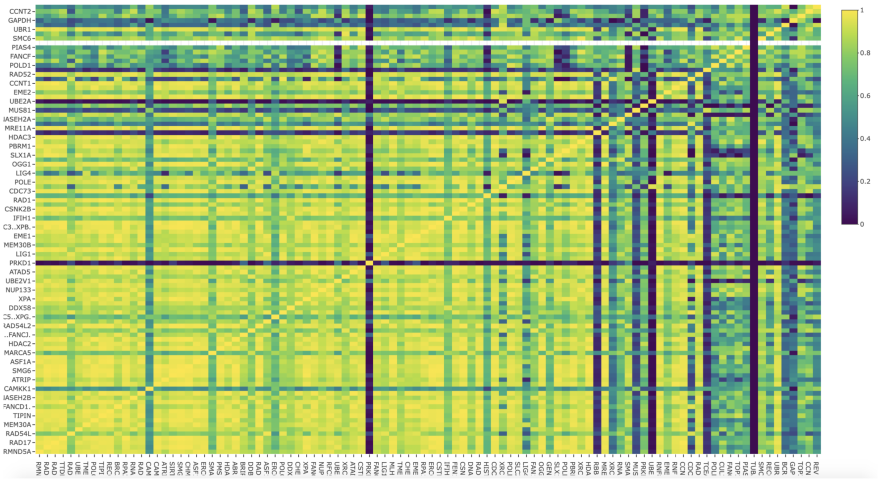

Fig. 8. Heatmaps of R^2 values, derived from the normalizing factor, HKG_{Tuba1a}.

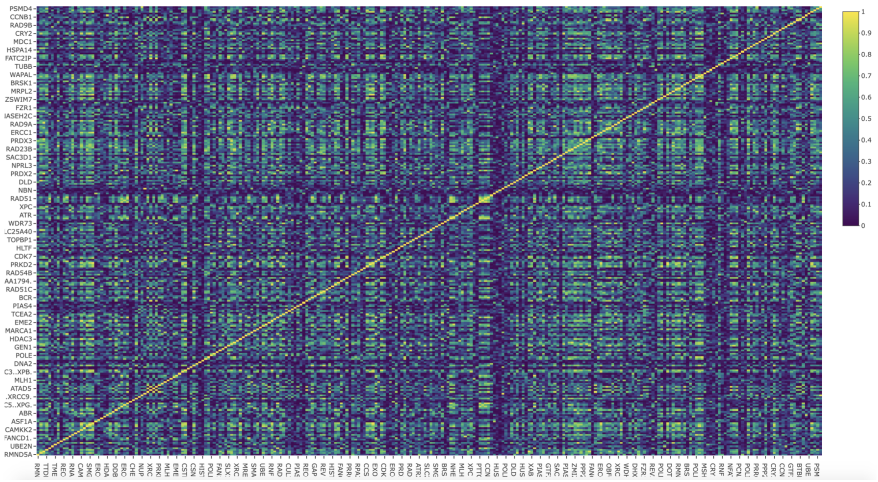

Fig. 9. Heatmaps of R^2 values, derived from the normalizing factor, $HKG_{AvgTubbTuba1aGapdh}$.

the data are false positives. When the expression values were normalized using another single gene expression value (*TUBA1A*), we noted that there were also many false positives in terms of R^2 values showing high correlations, as shown in Figure 8. Furthermore, this result indicated a lack of consistency among single expression normalizations.

Normalizing with $HKG_{AvgTubbTuba1aGapdh}$ slightly reduced the number of pairs with high linear correlation values, as shown in Figure 9. This observation suggests that using only three housingkeeping genes is not sufficient to reduce

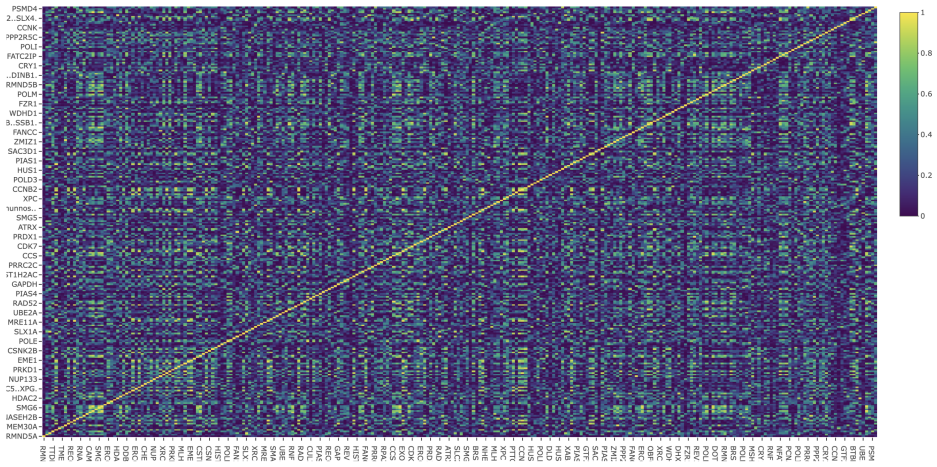

Fig. 10. Heatmaps of R^2 values, derived from the normalizing factor, HKG_{AvgG1}.

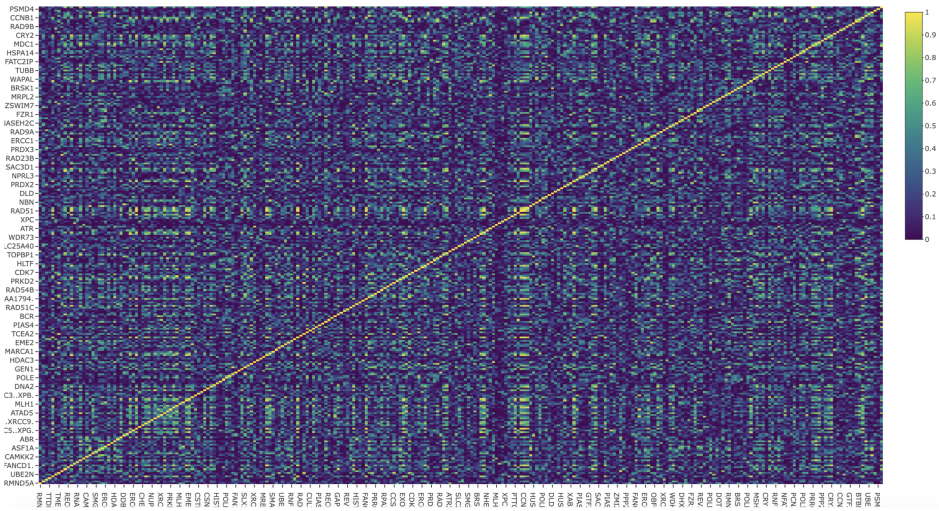

Fig. 11. Heatmaps of R^2 values, derived from the normalizing factor, HKG_{AvgG2}.

false positives. In the gene normalization survey by Vandesompele *et al.* [20], ten house keeping genes were sufficient for normalization purposes, where-as, single genes selected for normalization often led to misleading results.

We therefore applied the average expression of ten HK genes for normalization, as described in the Methods of Section 2. Interestingly, the heatmap generated using HKG_{AvgG1}, HKG_{AvgG2} or HKG_{AvgG3}, of Figures 10, 11 and 12, respectively, shows a pattern similar to the counterpart generated by the $HKG_{AvgTubbTuba1aGapdh}$ normalization, but was significantly different from the

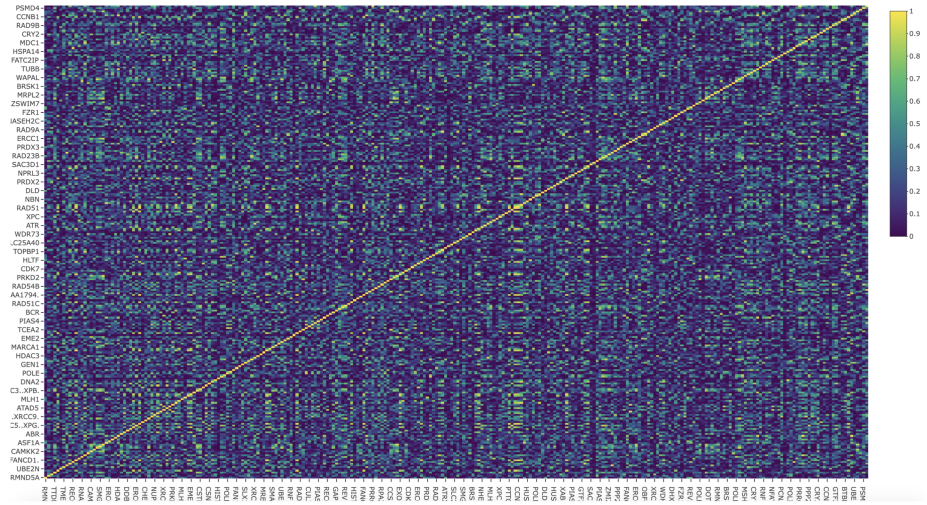

Fig. 12. Heatmaps of R^2 values, derived from the normalizing factor, HKG_{AvgG3}.

one generated by the *TUBB* normalization. Pairs with R^2-values close to 1 were comparatively rare in the heatmaps of Figures 10, 11 and 12 for normalizing factors, HKG_{AvgG1}, HKG_{AvgG2}, and HKG_{AvgG3}, respectively.

Shown in Table 3, the variance and mean values were collected from each normalizing factor experiment. We noted that when the normalizing factors were larger, the gene-gene correlations appeared to be more reliable in spite of the natural high complexity of gene-gene interactions [17].

If these correlations are true positives, the pattern should be reproducible regardless of what normalizing factors were used. In support of this idea, the heatmaps generated by average of ten expression values appear similar to each other. In other words, correlation in expression is preserved for certain pairs of genes regardless of what normalizing factors are used, as shown in all heatmaps in Figures 10, 11 and 12. In future works, we intend to construct experiments with more statistical analysis across larger data sets to be able to determine more of the wide-spread consistencies concerning gene-gene correlations.

Linear correlation between two genes imply that alteration in expression of one gene necessitates a proportional response of the other, shown in Figure 1-B. The idea of genetic dependency stems from "genetic interaction," a concept that has been long rooted in classical genetics. For example, large-sale genetic interactions have been studied in Drosophila [21], Mendelian diseases [22] and determined with relative ease in simple eukaryotic model organisms, such as *Saccharomyces cerevisiae* [23].

In general, genetic interaction between two genes exist if the absence of both genes produces a phenotype, that is different from phenotypes resulting from deficiency of individual genes [24]. For instance, combined deficiency of two genes

result in cell death, whereas deficiency of each gene permits cellular survival with reduced fitness. This scenario exemplifies a type of widely known genetic interaction and is succinctly referred to as "synthetic lethality." This idea was first described by Bridges in 1922, and has been proposed to use for cancer therapy since 1997 by Hartwell and Friend [24].

Since then, the search for synthetic lethal interactions between genetic conditions specific to cancer has been ongoing. The genetic screen performed in cancer cells, using either CRISPR or RNAi, to understand which gene expression changes are necessary for cancer with a specific genetic background reveal multiple synthetic lethal interactions [25], [26]. In addition, multi-pronged approach such as DAISY has been developed to further solidify identification of synthetic lethal gene pairs [16]. Our study complements these efforts due to two main reasons. First, co-expression analysis described in DAISY was performed using data derived from cell lines, whereas our study uses gene expression data from cancer tissues. Second, our preliminary study begins to address the reproducibility of co-expression that may be universal or tissue specific. With a more comprehensive analysis, we anticipate that our study will offer insights into genetic dependencies within patient tumor tissues.

4 Conclusions

In this work, we showcase how our method and tool, *GenExSt*, could be used to facilitate an automated, comparative, *all-against-all*, gene study. Using *GenExSt*, we discovered GIS genes whose expression values positively or negatively correlated with the expressions of other genes. Our results were gathered utilizing computational methods for FPKM gene expression normalization, allowing us to regress each gene over all the others in the study to collect adjusted R^2 values, from which we were able to detect potential correlations. Heatmaps were used to visualize the R^2 values to enable us to spot correlations according to elevated values.

To accomplish this goal, we determined normalizing factors (i.e., derived values from house keeping genes to be used to normalize the gene expression values of GIS genes) that served for comparison purposes across the unique data sets chosen for our study. An important part of this comparison involved linear models where the normalized expression data of genes was regressed over that of all other genes. In our study, consistency represented a plausible correlation between GIS gene expressions that were normalized by normalizing factors. Consistent correlations, we reasoned, would show patterns which would be also found from the successful normalizing factors that would be derived from the expressions of other HK genes.

Results By Single Expression Normalization: We found that the normalizing factors derived from single HK genes did not provide generally consistent or meaningful results. For instance, when using single genes to create the normalizing factors the results of each normalizing factor were dissimilar. In addition, our results described too many biologically improbable correlations.

We rejected these results as false positives. In total, since the single expression values were generally unable to achieve consistent results with each other, we found that this approach was not effective.

The normalizing factor (HKG_{Tubb}), derived from the single gene expression of $TUBB$, provided results to suggest that about half of the genes had high correlations with each other. While some of these correlations were were likely to be biologically relevant, there were too many to be biologically probable. This observation suggested that there were a large number of false positives created by this normalizing factor. The normalizing factor (HKG_{Tuba1a}), derived from the single gene expression of $TUBA1A$, provided results suggesting many high correlations between the GIS genes of our study. Again, we noted that there were too many high correlations, which suggested false negatives. The heatmaps generated from from HKG_{Tubb} and HKG_{Tuba1a} (Figures 7 and 8, respectively) did not maintain consistency as due to their many lacks of agreement in the correlations of genes.

Results By Multiple Expression Normalization: We found that using the averaged expression values of multiple HK genes was an effective approach to finding consistency across our data sets. When we used the average of three genes to normalize expression data, we found fewer cases that we estimated were biologically improbable. However, on close inspection of the results, we still noted that there were a large number of likely false positives. We noticed a pattern emerge; the more genes we used to create normalizing factors, then fewer pairs of genes exhibited correlations. Our preliminary method enabled us to identify the co-expression of gene pairs in breast cancer tissues. This technique allows for reproducibility across data sets and to compare approaches involving diverse normalization factors when detecting correlation patterns.

One of the main goals for the development of our method and tool, *GenExSt* was to streamline the search for genetic interactions in a biological system of interest using gene expression data. Although we chose to showcase the functionality of *GenExSt* using widely available cancer data, this tool may be used to identify pairs of genes whose expressions correlate in other diseased states if the concept of genetic interactions is applicable to those conditions.

4.1 Future Work

In future studies, we will aim to validate the correlations that we found in this preliminary study. For example, we will determine if *GenExSt* can predict pairs of genes that have been shown to have genetic interactions in literature. In addition,we will further develop our method and tool to be able to manage multiple projects where diverse normalizing factors are derived and compared. Such a development would integrate machine learning to be able to assist in recognizing correlations which are noteworthy.

4.2 Acknowledgment

We would like to thank Janyl Jumadinova for her help in proofing this manuscript.

References

1. T. Gaudelet, N. Malod-Dognin, J. Sánchez-Valle, V. Pancaldi, A. Valencia, and N. Pržulj, "Unveiling new disease, pathway, and gene associations via multi-scale neural network," *PloS one*, vol. 15, no. 4, p. e0231059, 2020.
2. R. A. Burrell, N. McGranahan, J. Bartek, and C. Swanton, "The causes and consequences of genetic heterogeneity in cancer evolution," *Nature*, vol. 501, no. 7467, p. 338, 2013.
3. H. Farmer, N. McCabe, C. J. Lord, A. N. Tutt, D. A. Johnson, T. B. Richardson, M. Santarosa, K. J. Dillon, I. Hickson, C. Knights *et al.*, "Targeting the dna repair defect in brca mutant cells as a therapeutic strategy," *Nature*, vol. 434, no. 7035, p. 917, 2005.
4. L. B. Conrad, K. Y. Lin, T. Nandu, B. A. Gibson, J. S. Lea, and W. L. Kraus, "Adp-ribosylation levels and patterns correlate with gene expression and clinical outcomes in ovarian cancers," *Molecular cancer therapeutics*, vol. 19, no. 1, pp. 282–291, 2020.
5. C. Soneson and M. D. Robinson, "Bias, robustness and scalability in single-cell differential expression analysis," *Nature methods*, vol. 15, no. 4, p. 255, 2018.
6. E. Luidy-Imada, T. Matam, L. Collado-Torres, W. Dinalankara, A. Stupnikov, C. Wilks, A. E. Jaffe, B. Langmead, J. T. Leek, A. Favorov *et al.*, "Differential analysis of gene expression across the human genome using recount2 and fantom-cat," 2018.
7. B. P. de Almeida, A. F. Vieira, J. Paredes, M. Bettencourt-Dias, and N. L. Barbosa-Morais, "Pan-cancer association of a centrosome amplification gene expression signature with genomic alterations and clinical outcome," *PLoS computational biology*, vol. 15, no. 3, p. e1006832, 2019.
8. L. Liu, C. Dalal, B. Heineike, and A. R. Abate, "High throughput gene expression profiling of yeast colonies with microgel-culture drop-seq," *Lab on a Chip*, 2019.
9. J. C. Spainhour, H. S. Lim, S. V. Yi, and P. Qiu, "Correlation patterns between dna methylation and gene expression in the cancer genome atlas," *Cancer informatics*, vol. 18, p. 1176935119828776, 2019.
10. H. Chen, C. Li, X. Peng, Z. Zhou, J. N. Weinstein, S. J. Caesar-Johnson, J. A. Demchok, I. Felau, M. Kasapi, M. L. Ferguson *et al.*, "A pan-cancer analysis of enhancer expression in nearly 9000 patient samples," *Cell*, vol. 173, no. 2, pp. 386–399, 2018.
11. W. Tang, F. Bertaux, P. Thomas, C. Stefanelli, M. Saint, S. Marguerat, and V. Shahrezaei, "baynorm: Bayesian gene expression recovery, imputation and normalization for single-cell rna-sequencing data," *Bioinformatics*, vol. 36, no. 4, pp. 1174–1181, 2020.
12. N. Fernandez-Pozo, F. B. Haas, R. Meyberg, K. K. Ullrich, M. Hiss, P.-F. Perroud, S. Hanke, V. Kratz, A. F. Powell, E. F. Vesty *et al.*, "Peatmoss (physcomitrella expression atlas tool): a unified gene expression atlas for the model plant physcomitrella patens," *The Plant Journal*, vol. 102, no. 1, pp. 165–177, 2020.
13. Z. W. Atwan, "Gapdh spike rna as an alternative for housekeeping genes in relative gene expression assay using real-time pcr," *Bulletin of the National Research Centre*, vol. 44, no. 1, pp. 1–8, 2020.
14. P. T. Inc. (2015) Collaborative data science. Montreal, QC. [Online]. Available: https://plot.ly
15. C. D. Putnam, A. Srivatsan, R. V. Nene, S. L. Martinez, S. P. Clotfelter, S. N. Bell, S. B. Somach, J. E. De Souza, A. F. Fonseca, S. J. De Souza *et al.*, "A genetic

network that suppresses genome rearrangements in saccharomyces cerevisiae and contains defects in cancers," *Nature communications*, vol. 7, p. 11256, 2016.
16. L. Jerby-Arnon, N. Pfetzer, Y. Y. Waldman, L. McGarry, D. James, E. Shanks, B. Seashore-Ludlow, A. Weinstock, T. Geiger, P. A. Clemons *et al.*, "Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality," *Cell*, vol. 158, no. 5, pp. 1199–1209, 2014.
17. H. J. Cordell, "Detecting gene–gene interactions that underlie human diseases," *Nature Reviews Genetics*, vol. 10, no. 6, pp. 392–404, 2009.
18. F. Changyong, W. Hongyue, L. Naiji, C. Tian, H. Hua, L. Ying *et al.*, "Log-transformation and its implications for data analysis," *Shanghai archives of psychiatry*, vol. 26, no. 2, p. 105, 2014.
19. E. Eisenberg and E. Y. Levanon, "Human housekeeping genes, revisited," *TRENDS in Genetics*, vol. 29, no. 10, pp. 569–574, 2013.
20. J. Vandesompele, K. De Preter, F. Pattyn, B. Poppe, N. Van Roy, A. De Paepe, and F. Speleman, "Accurate normalization of real-time quantitative rt-pcr data by geometric averaging of multiple internal control genes," *Genome biology*, vol. 3, no. 7, pp. research0034–1, 2002.
21. L. Dehnen, M. Janz, J. K. Verma, O. E. Psathaki, L. Langemeyer, F. Fröhlich, J. J. Heinisch, H. Meyer, C. Ungermann, and A. Paululat, "A trimeric metazoan rab7 gef complex is crucial for endocytosis and scavenger function," *Journal of Cell Science*, 2020.
22. K. Rahit and M. Tarailo-Graovac, "Genetic modifiers and rare mendelian disease," *Genes*, vol. 11, no. 3, p. 239, 2020.
23. J. Van Leeuwen, C. Pons, J. C. Mellor, T. N. Yamaguchi, H. Friesen, J. Koschwanez, M. M. Ušaj, M. Pechlaner, M. Takar, M. Ušaj *et al.*, "Exploring genetic suppression interactions on a global scale," *Science*, vol. 354, no. 6312, p. aag0839, 2016.
24. S. M. Nijman, "Synthetic lethality: general principles, utility and detection using genetic screens in human cells," *FEBS letters*, vol. 585, no. 1, pp. 1–6, 2011.
25. A. Tsherniak, F. Vazquez, P. G. Montgomery, B. A. Weir, G. Kryukov, G. S. Cowley, S. Gill, W. F. Harrington, S. Pantel, J. M. Krill-Burger *et al.*, "Defining a cancer dependency map," *Cell*, vol. 170, no. 3, pp. 564–576, 2017.
26. X. Wang, A. Q. Fu, M. E. McNerney, and K. P. White, "Widespread genetic epistasis among cancer genes," *Nature communications*, vol. 5, p. 4828, 2014.