A Statistical Approach to Uncovering How Tissue Composition Relates to Gene Expression Variability

Chai Ait Jabbour

Under the Supervision of Dr. Theodore Perkins

Table of Contents

Abstract

Background:

Gene expression variability (GEV) plays a crucial role in understanding biological processes and disease pathogenesis, particularly in cancer research. This project investigates the influence of tissue composition, specifically cell type, and individual sample characteristics on GEV, focusing on breast cancer as a model system.

Results:

Decomposing the total variance using the Law of Total Variance highlighted the predominant influence of cell type on gene expression variability. A histogram representation demonstrated a notable concentration of genes exhibiting a substantial contribution of cell type-related variability to the total variance.

Conclusion:

The findings underscore the significance of tissue composition in driving gene expression patterns, with implications for disease diagnosis and treatment, particularly in breast cancer. This analysis provides a foundation for identifying genetic biomarkers and developing personalized therapeutic interventions.

Keywords:

Gene Expression Variability, Law of Total Variance, cell type, tissue composition, breast cancer, biomarkers.

Background

Gene expression variability (GEV) refers to the inherent differences in the activity levels of genes across cells, tissues, or individuals. Understanding the sources and extent of GEV is crucial for elucidating the molecular mechanisms underlying biological processes and disease pathogenesis. In the context of cancer research, GEV analysis can provide valuable insights into the heterogeneity of tumor cells, treatment responses, and disease progression.

This project, titled "A Statistical Approach to Uncovering How Tissue Composition Relates to Gene Expression Variability", aims to investigate the relative contributions of tissue composition and individual sample characteristics to GEV. By leveraging statistical methodologies, the project seeks to discern whether variations in GEV primarily stem from differences in cell type or from inherent individual differences between patients.

Previous studies have demonstrated the significance of GEV in various biological contexts. GEV analysis has been instrumental in identifying biomarkers for disease diagnosisn and treatment stratification. However, the specific factors driving GEV, particularly within the context of tissue composition heterogeneity, remain obscure.

Cell type heterogeneity within tissues poses a significant challenge in dissecting the underlying mechanisms of GEV. Each cell type exhibits distinct gene expression profiles, influenced by both intrinsic cellular programs and extrinsic environmental cues. Understanding how these factors interact to shape GEV is essential for unraveling the complex regulatory networks governing cellular behavior and disease states.

By focusing on breast cancer as a model system, the project aims to shed light on the interplay between tissue composition and GEV in a clinically relevant context. Breast cancer is known for its cellular heterogeneity, which contributes to variations in treatment responses and clinical outcomes among patients. Investigating the role of cell type composition in driving GEV may provide critical insights into the molecular underpinnings of breast cancer progression and therapeutic resistance.

In summary, this project is an effort to bridge the gap between statistical analysis techniques and biological understanding in the study of GEV. By dissecting the contributions of tissue composition to GEV, the project aims to uncover novel insights into the biological processes underlying cancer development and progression, with implications for personalized medicine and targeted therapeutic interventions.

Analysis

The analysis begins with a comprehensive exploration of the dataset, consisting of a multi-dimensional data matrix encapsulating gene expression profiles across various cell types and patient samples. The dataset serves as the foundation for investigating the relationship between tissue composition and gene expression variability (GEV) within the context of breast cancer.

Overview of the Data Matrix

The data matrix consists of rows representing individual genes and columns corresponding to different samples (sample here refers to a pair of patient and cell type), with additional metadata rows providing information such as the number of cells in each sample. Each value in the matrix represents the expression level of a specific gene in a particular sample, quantified through RNA sequencing (RNA-seq).

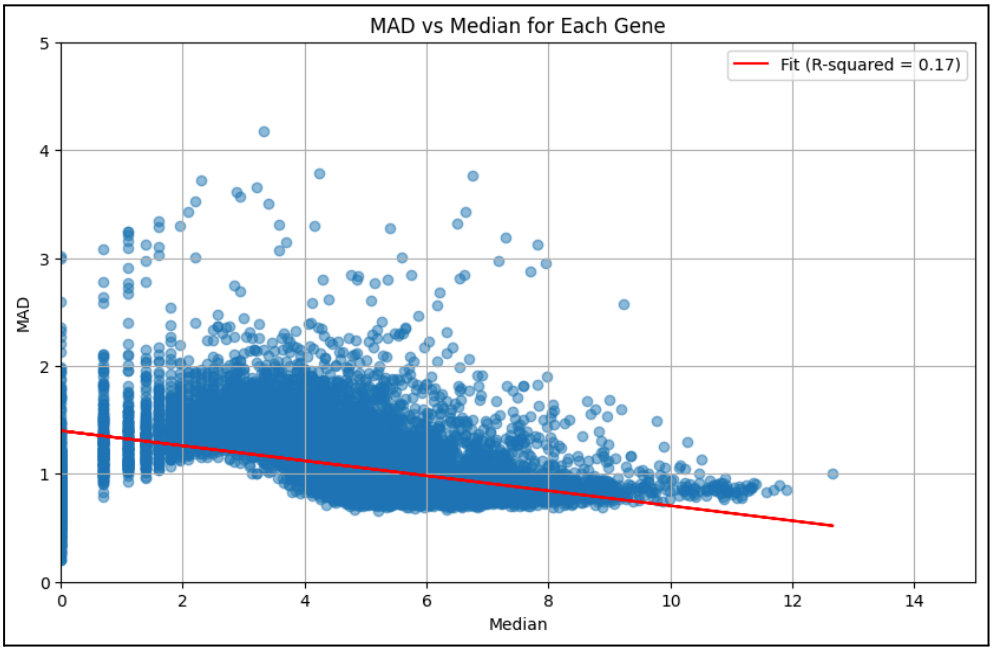
	A	B	C	D	E	F	G	H	I	J
1	sample_id	BIOKEY_1	BIOKEY_1	BIOKEY_1	BIOKEY_1	BIOKEY_1	BIOKEY_1	BIOKEY_1	BIOKEY_1	BIOKEY_1
2	cell_type	B cells	DC	Endothelial	Epithelial	Macrophage	Mast	Mesenchymal	Plasma cells	T/NK
3	n_cells	435	28	14	440	380	0	234	30	2932
4	A1BG	50	15	3	166	216	0	97	28	434
5	A1BG-AS1	13	2	0	15	37	0	18	0	91
6	A2M	4	2	296	31	1363	0	653	0	56
7	A2M-AS1	1	2	0	9	2	0	9	0	25

Figure 1 - Data matrix

The first few rows of the data matrix reveal the structure and content of the dataset.
The 'sample_id' row identifies the patient from which this sample was taken.
The first row denotes the cell types: B cells, dendritic cells (DC), endothelial cells, epithelial cells, macrophages, mast cells, mesenchymal cells, plasma cells, and T/NK cells.
Additionally, the 'n_cells' row provides information about the number of cells associated with each sample (a sample denotes a pair of patient and cell type), which serves as a crucial parameter for assessing the statistical significance of gene expression variations.
Further exploration of the data matrix reveals numerical values representing gene expression levels for each gene across different samples. For instance, the expression levels of genes such as 'A1BG', 'A1BG-AS1', 'A2M' and 'A2M-AS1' are recorded across various samples. These expression levels in the matrix indicate the relative abundance or activity levels of each gene within each sample.
These expression measurements serve as the foundation for our analysis.

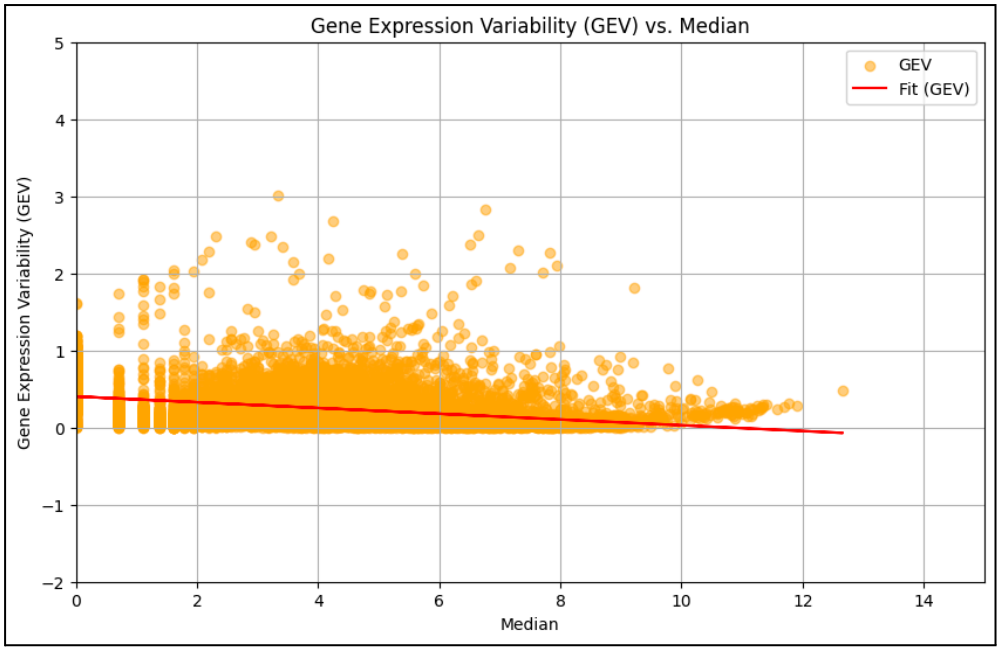
Exploring the Raw Data

In this section, we delve into the initial exploration of our dataset, focusing on two key graphical representations: Median Absolute Deviation (MAD) vs. Median and Gene Expression Variability (GEV) vs. Median. These graphs offer valuable insights into the relationship between gene expression variability and the median expression level of genes.



Graph 1 - Median Absolute Deviation vs. Median of each Gene for Raw Data

The MAD vs. Median graph illustrates the dispersion of gene expression values around their respective median levels. Each data point in this graph represents a gene, with the x-axis denoting the median expression level and the y-axis representing the MAD, a measure of variability.
Upon visual inspection, we observe a slowly descending trend in the linear fit, indicating that genes with higher median expression levels tend to exhibit lower variability in their expression patterns. This trend suggests that genes with higher median expression levels tend to have more stable expression patterns across samples, while genes with lower median expression levels exhibit greater variability.



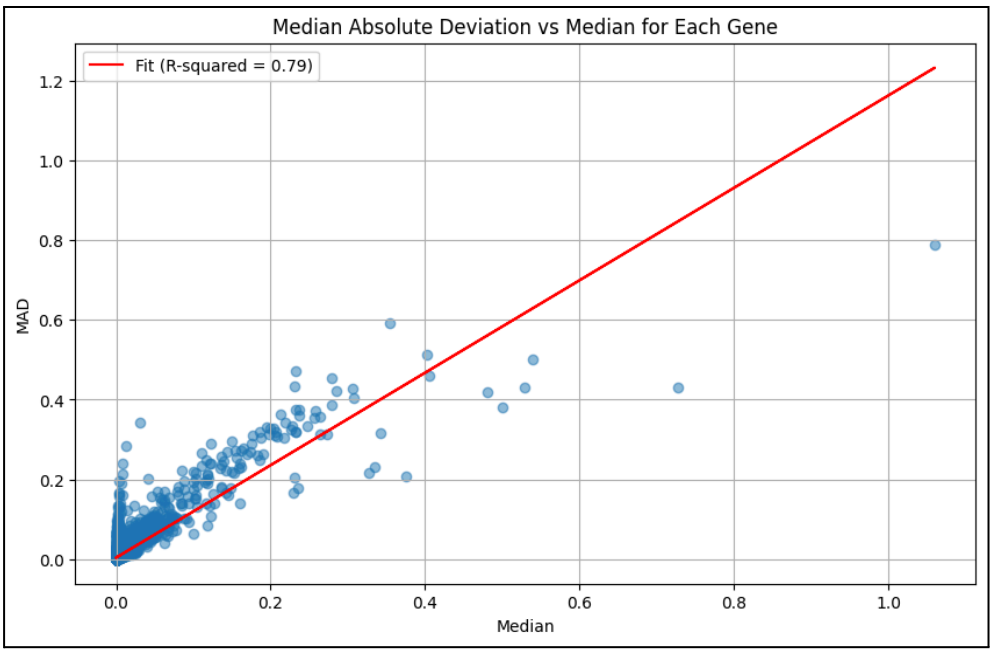
Graph 2 - Gene Expression Variability vs. Median of each Gene for Raw Data

The GEV vs. Median graph depicts the relationship between gene expression variability and the median expression level, where GEV is obtained by taking the absolute value of the difference between the linear fit and the MAD values in Graph 1. Notably, the linear fit in this graph shows a milder downward slope. This suggests that, while there is a slight decrease in gene expression variability with increasing median expression levels, the magnitude of this decrease is much smaller compared to the overall variability observed in the MAD vs. Median graph.

These observations are made on raw data, which means that noise and the sequencing depth difference are unaccounted for. The next section explores normalized data, and will show us to what extent the results in this section are affected by factors that are irrelevant to our study.

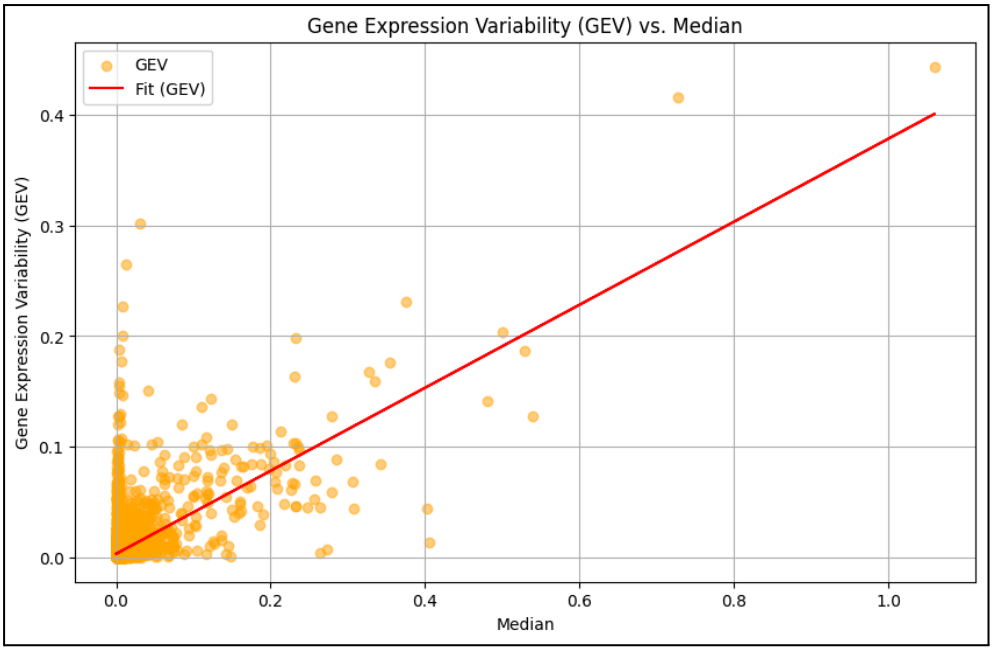
Exploring the Normalized Data

Normalization is a critical step in the preprocessing of gene expression data, aimed at mitigating biases introduced by factors such as sequencing depth and technical variations. Our normalization process involves two key steps: Reads Per Million (RPM) normalization and logarithmic transformation. By applying these normalization techniques, we aim to standardize gene expression measurements across samples, facilitating fair comparisons and accurate analysis.
In this section, we investigate the impact of normalization on the relationship between gene expression variability and median expression levels, as depicted in the MAD vs. Median and GEV vs. Median graphs. These graphical representations allow us to assess how normalization alters the patterns of gene expression variability and explore the underlying biological insights with greater clarity.



Graph 3 - Median Absolute Deviation vs. Median of each Gene for Normalized Data

In this graph, we examine the dispersion of gene expression values around their median levels after normalization. Each data point represents a gene, with the x-axis denoting the median expression level and the y-axis representing the MAD. Unlike the raw data, the linear fit in this graph exhibits a strong positive trend. This suggests that after normalization, genes with higher median expression levels tend to have slightly higher variability in their expression patterns, while genes with lower median expression levels, which is most genes, display lower variability.



Graph 4 - Gene Expression Variability vs. Median of each Gene for Normalized Data

Similarly, the GEV vs. Median graph portrays the relationship between gene expression variability and median expression levels post-normalization. The linear fit in this graph also shows a positive trend, although it's not as positive as the one in Graph 3. This means that a gene with a higher median expression value is more likely to have a higher degree of expression variability.
We can also observe that most genes have a lower GEV and median expression value. This can potentially be explained as follows: most genes have a lower GEV, which means their levels are relatively consistent throughout all samples, because they are needed for the basic functioning of cells in general, while genes with a higher GEV, that are the minority, constitute genes that perform more specialized functions that not all cells need and/or not all individuals have.

Overall, exploring the normalized data provides valuable insights into the impact of preprocessing steps on gene expression variability. By accounting for technical biases and variations, we managed to obtain more interesting results compared to when we performed the same analysis on raw data. This allows us to better discern the underlying patterns and relationships within the dataset.

Introducing the Law of Total Variance

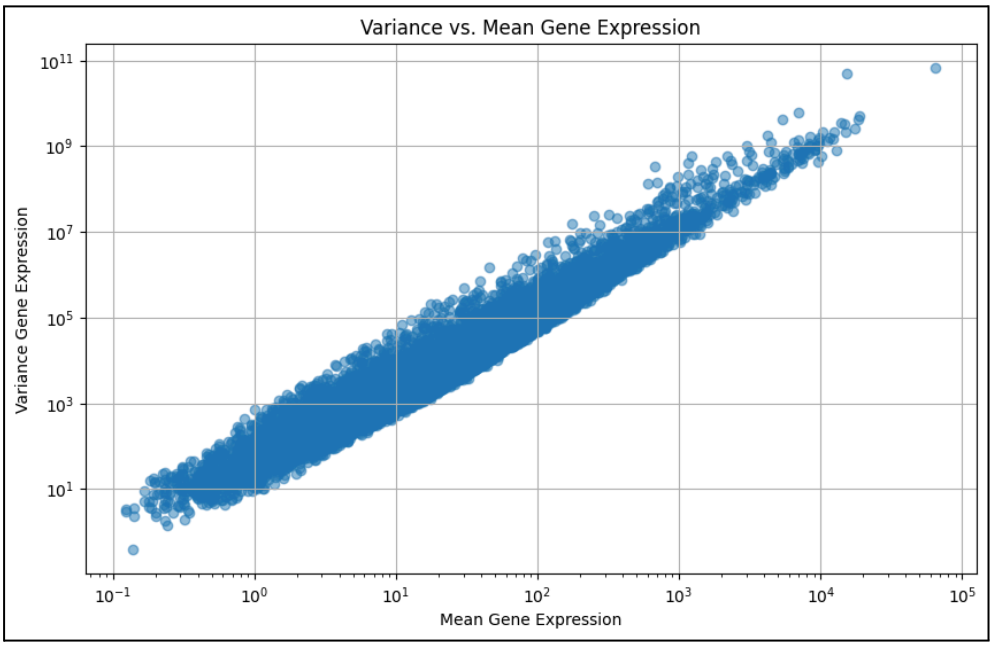
To advance our analysis, it's essential to introduce the Law of Total Variance. This principle provides a framework for understanding the total variability of a random variable. According to this law, the total variance of a variable can be decomposed into two components: the variance of the conditional means of the variable, plus the conditional variance of the variable itself.

$$\mathbf{Var}(Y) = \mathbf{E}[\mathbf{Var}(Y \mid X)] + \mathbf{Var}(\mathbf{E}[Y \mid X]).$$

In simpler terms, the Law of Total Variance suggests that the overall variability of a dataset can be attributed to two main sources: the variability of the average values within groups (conditional means) and the variability of individual data points around those averages.

This concept is particularly relevant in our study of gene expression variability (GEV). By applying the Law of Total Variance to our dataset, we can analyze how much of the overall variability in gene expression can be attributed to differences between cell types (represented by the mean of the conditional variances) and how much is due to individual differences between patients (represented by variance of the conditional means).

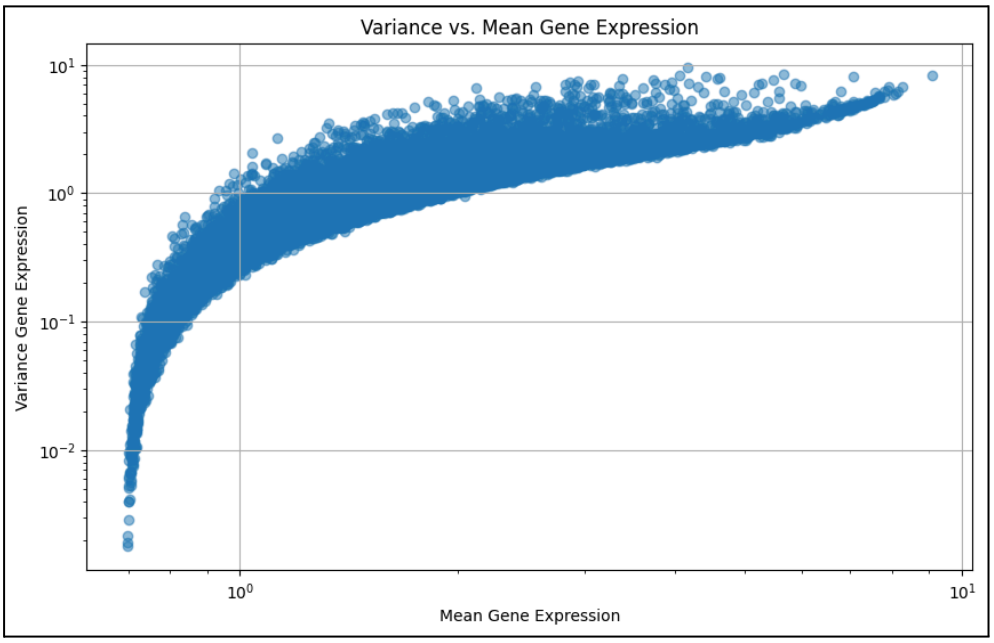
Visualizing the Total Variance of Each Gene for Raw Data



Graph 5 - Total Variance vs. Mean of Each Gene for Raw Data

Upon examining Graph 5, we can see that the scatterplot is arranged in a positive slope line, indicating a positive relationship between the total variance and the mean expression level of each gene. Notably, the total variance exceeds the mean expression level for most genes. We will decompose this total variance into variance caused by individual patient differences and differences in cell type.

Visualizing the Total Variance of Each Gene for Normalized Data

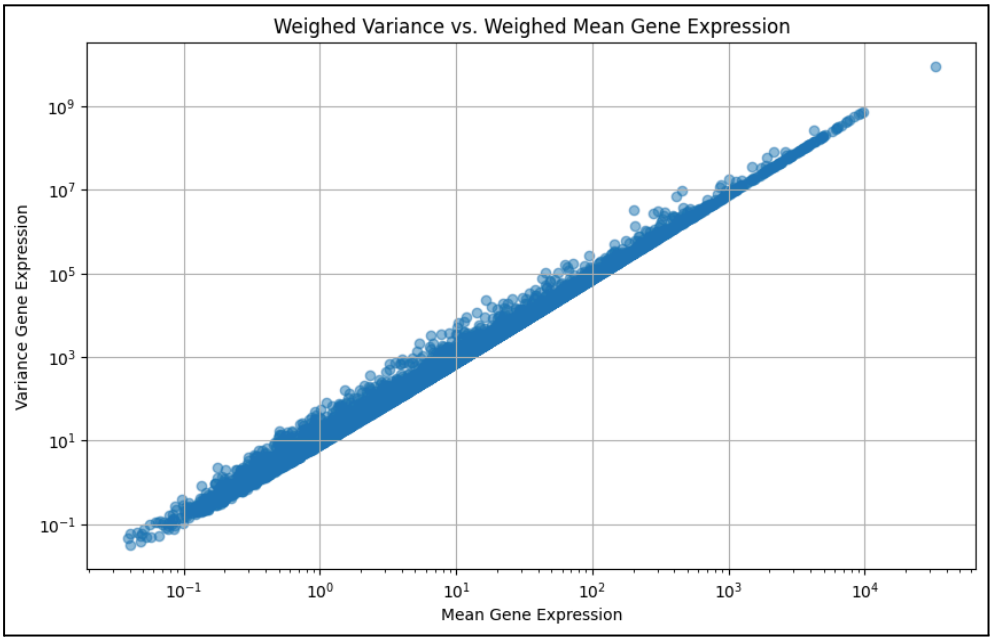


Graph 7 - Total Variance vs. Mean of Each Gene for Normalized Data

Upon inspecting Graph 7, we observe a positive logarithmic relationship between the total variance and the mean expression level of each gene. Unlike the linear relationship observed in the raw data, this relationship takes on a distinctive shape, with the bulk of the data points focused in the middle of the shape as opposed to the extremities. This suggests that, in the context of normalized data, the relationship between total variance and mean expression level is not as straightforward, but rather follows a logarithmic trend.

Visualizing the Total Variance of Each Gene for Weight Scaled Normalized Data

In this section, we attempt to take the effect of the number of cells taken for each sample (pair of patient and cell type) out of the equation. This is done by normalizing the data using the Reads Per Million transform only, then calculating the 'Patient Fraction' for each sample. The 'Patient Fraction' is calculated by taking the number of cells for each sample and dividing it by the total amount of cells for the patient at hand. Then, we proceed to multiply all the gene expression values for this sample by the corresponding 'Patient Fraction'.

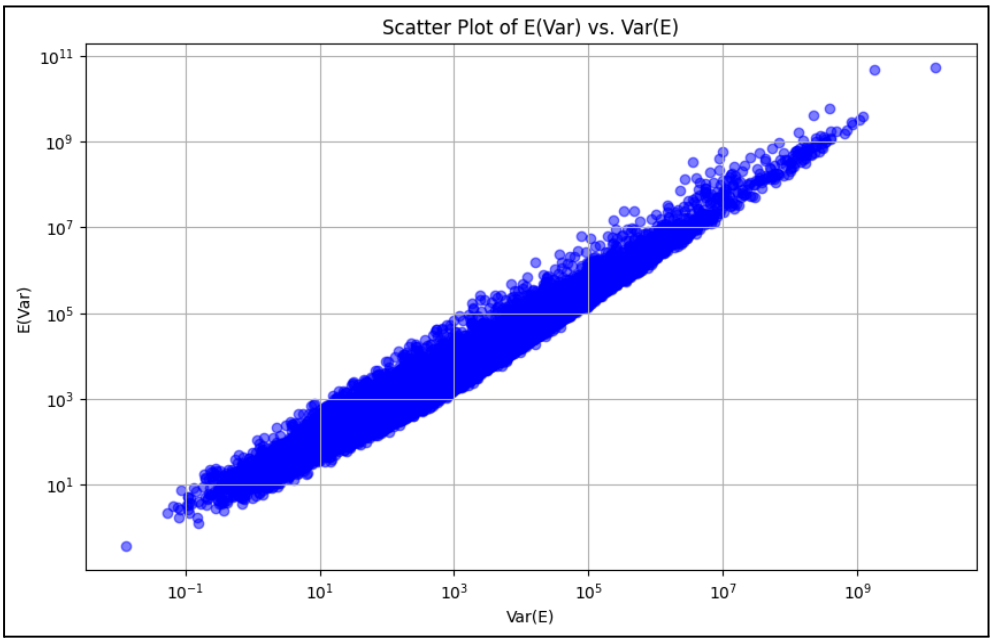


Graph 8 - Total Variance vs. Mean of Each Gene for Normalized Data Scaled by Weight

As we can see in the graph above, there is still a positive relationship between the total variance and the mean for each gene, which goes to show that the difference in the number of cells taken for each sample has little interference with the distribution of the data.

Decomposing the Total Variance of Raw Data using the Law of Total Variance with X as the Cell Type

In this section, we employ the Law of Total Variance to delve deeper into the variability of gene expression within our dataset, with X representing the cell type. This is done by compiling all columns with the same cell type in different sub-datasets and using these datasets for our analysis. The Law of Total Variance offers a powerful framework for dissecting the sources of variability in our data, shedding light on the contributions of different factors to overall gene expression variability (GEV).

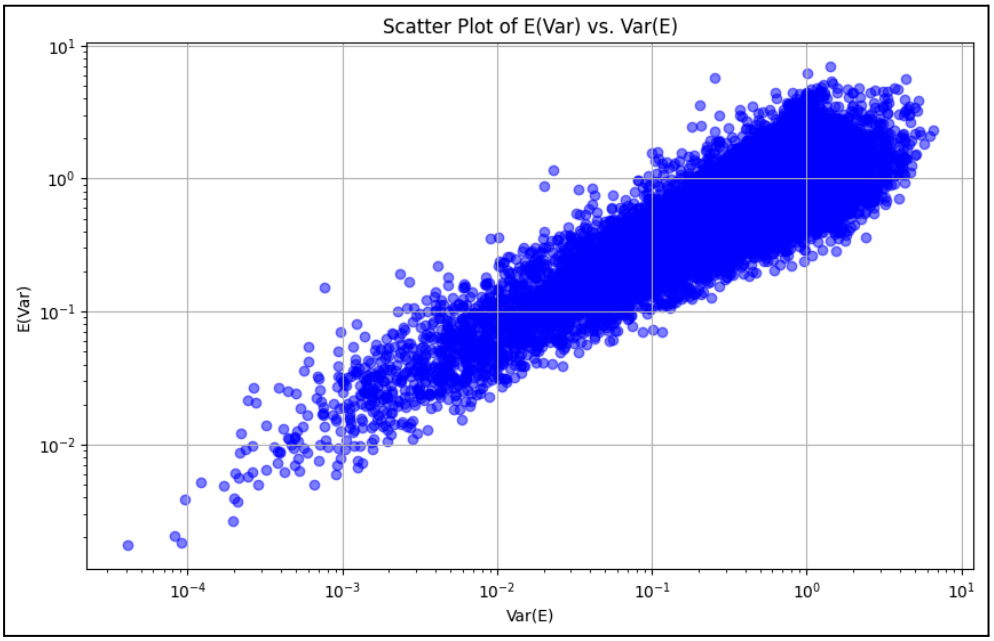


Graph 9 - Mean of Variances vs. Variance of Means of Each Gene across Cell Types for Raw Data

In Graph 9, we analyze the decomposition of total variance into two components: the mean of variances, which represents the variance caused by differences in cell type, and the variance of means, which represents the variance caused by patient differences, across different cell types. We observe a positive slope line, indicating a positive relationship between the mean of variances and the variance of means. However, the mean of variances slightly dominates over the variance of means, suggesting that the variability attributed to differences between cell types (mean of variances) plays a more significant role in driving overall gene expression variability than individual differences between patients (variance of means).

Decomposing the Total Variance of Normalized Data using the Law of Total Variance with X as the Cell Type

In this section, we perform the same analysis as the last section on normalized data.

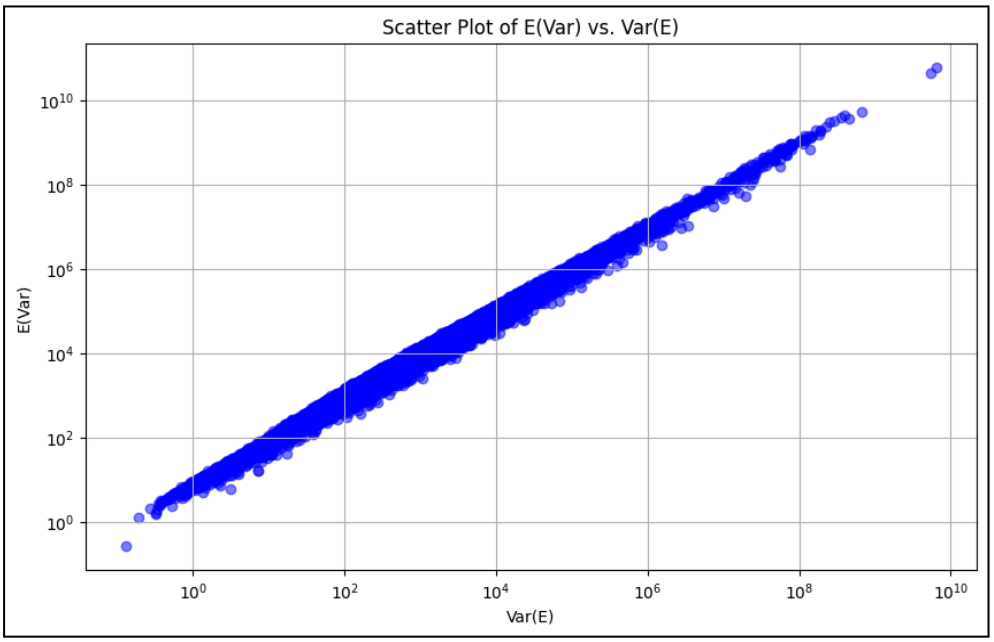


Graph 10 - Mean of Variances vs. Variance of Means of Each Gene across Cell Types for Normalized Data

Graph 10 presents the decomposition of total variance into the mean of variances and the variance of means across different cell types for normalized data. Interestingly, the graphical representation vaguely resembles the shape of a bat, with a distinctive curve and spread. Consistently with Graph 9, this graph shows that differences in cell type account for slightly more of the total variance than individual patient differences.

Decomposing the Total Variance of Raw Data using the Law of Total Variance with X as the Patient

In this section, we extend our analysis to raw data with X representing the patient this time instead of cell type, which means that we are taking all columns with the same patient id and grouping them in smaller datasets. The Law of Total Variance is applied to unravel the components of gene expression variability. By examining the resulting graph, we aim to elucidate the intricate relationships between gene expression variability, individual differences between patients, and differences between cell types.

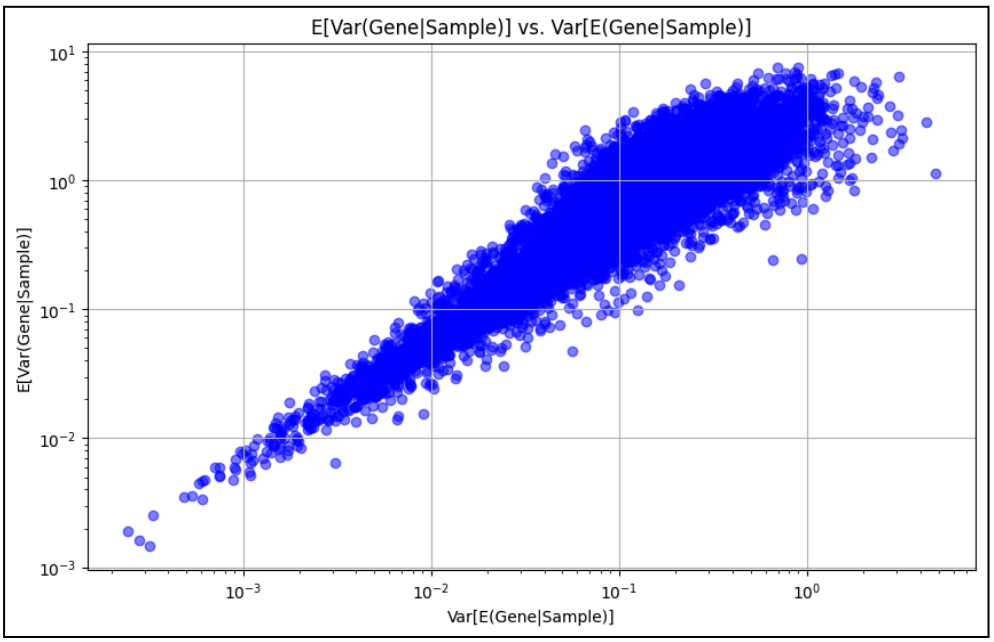


Graph 11 - Mean of Variances vs. Variance of Means of Each Gene across Patients for Raw Data

Graph 11 illustrates the decomposition of total variance into the mean of variances and the variance of means across different patients for raw data. Notably, the graphical representation reveals a very precise linear relationship between the mean of variances and the variance of means. This precise linear trend suggests that, in the context of patient-specific differences, the variability attributed to differences between cell types (represented by the mean of variances) dominates over the variability attributed to differences in gene expression patterns across patients (represented by the variance of means).

Decomposing the Total Variance of Normalized Data using the Law of Total Variance with X as the Patient

In this section, we employ the Law of Total Variance to decompose the total variance of normalized data with X representing the patient. Through graphical analysis, we aim to unravel the complex interplay between gene expression variability and individual patient characteristics.

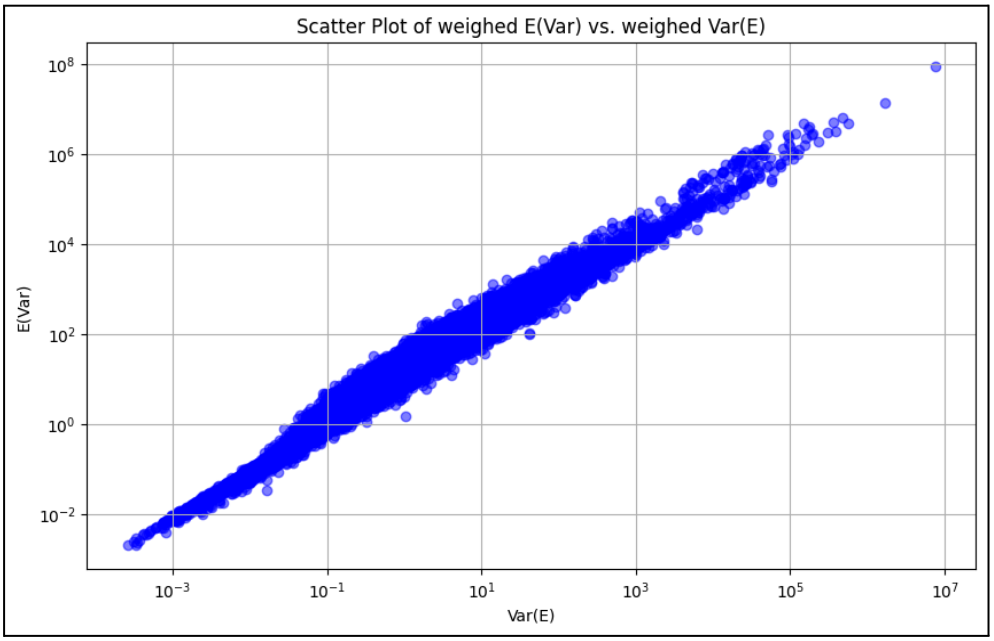


Graph 12 - Mean of Variances vs. Variance of Means of Each Gene across Patients for Normalized Data

Graph 12 showcases the breakdown of total variance into the mean of variances and the variance of means across different patients for normalized data. Notably, the graphical representation reveals a distinctive shape reminiscent of a bat, characterized by a narrow base and a prominent, rounded end. This unique shape suggests that, within the context of normalized data, the variability attributed to differences between cell types (represented by the mean of variances) predominates over the variability stemming from differences across patients (represented by the variance of means).

Decomposing the Total Variance of Weight Scaled Normalized Data using the Law of Total Variance with X as the Patient

Given the dataset where each gene expression value was scaled by the 'Patient Fraction', which is the ratio of the number of cells in the sample to the number of cells in all the samples corresponding to the same patient, we decompose the total variance of each gene according to the Law of Total Variance.

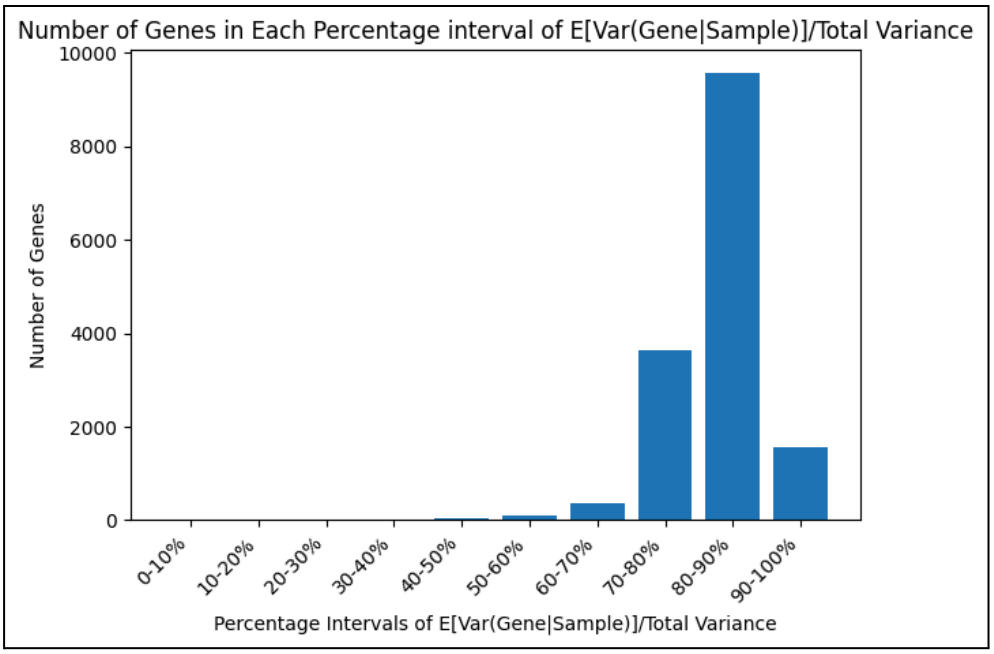


Graph 13 - Mean of Variances vs. Variance of Means of Each Gene across Patients for Raw Data

The graph above shows that the mean of variances is always higher than the variance of means for each gene. This corroborates the findings of the graphs in the sections above, and goes to show that the amount of cells in each sample has little effect on the analysis, and that tissue composition is the primary reason for the variability of gene expression values.

Results

In our investigation into the decomposition of total variance in normalized data using the Law of Total Variance, focusing on the patient as the variable of interest, we made significant discoveries. One particularly striking observation arises from the histogram representation of the ratio of the mean of variances to total variance across all genes. Within this distribution, a notable concentration of genes falls within the 80-90% interval of this ratio. This concentration suggests that the majority of genes exhibit a substantial contribution of the mean of variances to the total variance.
Such findings highlight the predominant influence of cell type on gene expression variability within our dataset. This histogram not only offers valuable insights into the underlying factors shaping gene expression patterns but also provides a foundation for further exploration and interpretation of our results.



Histogram 14 - Histogram Showing Number of Genes in Each Percentage interval of the ratio of the Mean of Variances of Each Gene over the Total Variance

Conclusion

Our investigation into the decomposition of total variance in normalized data using the Law of Total Variance has yielded compelling insights into the factors influencing gene expression variability.
Our findings underscore the significant role of cell type in driving gene expression patterns, with a notable concentration of genes exhibiting a substantial contribution of cell type-related variability to the total variance. This revelation highlights the intricate interplay between tissue composition and gene expression dynamics within our dataset.

These discoveries hold profound implications for our understanding of gene regulation mechanisms and their relevance to disease states, particularly in the context of breast cancer. By elucidating the predominant influence of cell type on gene expression variability, this analysis opens possibilities for the identification of genetic biomarkers that could serve as valuable tools for breast cancer diagnosis and treatment.
Leveraging this knowledge, future research endeavors can further explore the intricate relationship between cell type-specific gene expression patterns and disease progression, ultimately facilitating more effective personalized therapeutic interventions.

In conclusion, this analysis lets us know that there are benefits in studying the expression variability of certain genes that could be associated with breast cancer, and holds promise for the development of novel approaches to combat breast cancer and other complex diseases.

References

Bashkeel, N., Perkins, T.J., Kærn, M. *et al.* Human gene expression variability and its dependence on methylation and aging. *BMC Genomics* 20, 941 (2019). https://doi.org/10.1186/s12864-019-6308-7