الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Frères Mentouri Constantine 1
Faculté des Sciences de la Nature et de la Vie
Département de Biologie Appliquée

جامعة قسنطينة 1 الإخوة منتوري
كلية علوم الطبيعة والحياة
قسم البيولوجيا التطبيقية

Mémoire présenté en vue de l'obtention du diplôme de Master
Domaine : Sciences de la Nature et de la Vie
Filière : Biotechnologies
Spécialité : Bio-informatique

N° d'ordre :
N° de série :

Intitulé :

A Dissimilarity Measure for High-Dimensional Gene Expression Datasets: "Distance-DissimRatio" for Quantifying Transcriptomic Variation.

Présenté par :Soutenu le : 25/06/2025

- AOUISSI BOUCHRA
- HADJ AZZAM SARRA

Jury d'évaluation :

Président(e) : CHEHILI Hamza	M.C.A- UFM Constantine
Encadrant : KENIDRA BILEL	M.C.B- UFM Constantine
Examinateur : MEZIANI Dahbia Yasmina	M.C.B- UFM Constantine

Année universitaire
2024 - 2025

DEDICATION

In the Name of Allah, the Most Merciful, the Most Gracious. All praise is to Allah, the Lord of the worlds, and prayers and peace be upon Mohamed, His servant and messenger.

We have the honor to dedicate this research paper to:

Our dear parents for their unconditional love, support, and prayers which have always led us

through the valley of darkness with an endless light of hope.

Our beloved siblings who have never left our sides, surrounding us with love and care.

Our dear friends who have constantly pushed us forward, chasing insecurities and ensuring that good times keep flowing

BOUCHRA and SARRA

ACKNOWLEDGEMENTS

I extend my sincerest thanks to Mr. Billel KENIDRA, our advisor, whose expert guidance and invaluable advice significantly enhanced the quality of my manuscript. His mentorship has been instrumental in my academic growth, and I am profoundly grateful for the opportunity to have learned from him.

I am also grateful to the committee members who generously dedicated their time to review and evaluate my work.

Furthermore, I would like to acknowledge all the individuals who contributed to this project, whether through intellectual input, moral support, or encouragement. Their assistance has been crucial to the successful completion of this work.

ABSTRACT

The selection of appropriate similarity measures for high-dimensional gene expression datasets represents a critical challenge in computational genomics, particularly as genomic technologies have evolved from microarray platforms to RNA-sequencing and single-cell applications. Traditional metrics such as Euclidean distance, Manhattan distance, Pearson correlation, and Chi-Square distance often prove inadequate when confronted with the complex characteristics of modern genomic datasets, including extreme high-dimensionality, extensive zero-inflation patterns, systematic batch effects, and heterogeneous noise profiles. To address these fundamental limitations, we developed the Distance-DissimRatio (DDR) methodology, a novel similarity measure that integrates mean-centered deviation analysis, direct expression difference assessment, and comprehensive normalization framework. The DDR methodology exhibits unique characteristics including scale independence, bounded influence properties, zero-inflation robustness, and enhanced biological interpretability. Comprehensive experimental validation was conducted across three diverse gene expression datasets: GSE43346 (68 samples × 54,675 genes), GSE10072 (107 samples × 22,283 genes), and GSE13576 (209 samples × 54,675 genes). Performance evaluation using F-measure and Davies-Bouldin Index demonstrated DDR's superior performance compared to traditional metrics. DDR achieved perfect maximum F-measure (1.0) and optimal Davies-Bouldin index (2.2071) for GSE43346, highest mean F-measure (0.9883) for GSE10072, and competitive performance for GSE13576. The methodology demonstrates linear computational complexity O(n), enabling efficient analysis of large-scale genomic datasets. These results establish DDR as a robust, scalable, and biologically interpretable framework with significant implications for precision medicine, biomarker discovery, and gene regulatory network analysis.

Keywords: Gene expression analysis, similarity measures, high-dimensional data, genomics, clustering, biomarker discovery, precision medicine, computational biology.

ملخص

يمثل اختيار مقاييس التشابه المناسبة لمجموعات بيانات التعبير الجيني عالية الأبعاد تحدياً حاسماً في علم الجينوم الحاسوبي، خاصة مع تطور التقنيات الجينومية من منصات الرقائق الدقيقة إلى تسلسل الحمض النووي الريبوزي والتطبيقات أحادية الخلية. غالباً ما تثبت المقاييس التقليدية مثل المسافة الإقليدية، مسافة مانهاتن، ارتباط بيرسون، ومسافة مربع كاي عدم كفايتها عند مواجهة الخصائص المعقدة لمجموعات البيانات الجينومية الحديثة، بما في ذلك الأبعاد العالية المفرطة، أنماط تضخم الأصفار الواسعة، تأثيرات الدفعات المنهجية، وملفات الضوضاء غير المتجانسة. لمعالجة هذه القيود الأساسية، طورنا منهجية نسبة التباعد-التباين (DDR)، وهي مقياس تشابه مبتكر يدمج تحليل الانحرافات المتمركزة حول المتوسط، تقييم اختلافات التعبير المباشرة، وإطار التطبيع الشامل. تظهر منهجية DDR خصائص فريدة تشمل استقلالية المقياس، خصائص التأثير المحدود، مقاومة تضخم الأصفار، وقابلية التفسير البيولوجي المحسّنة. تم إجراء التحقق التجريبي الشامل عبر ثلاث مجموعات بيانات تعبير جيني متنوعة: GSE43346 (68 عينة × 54,675 جين)، GSE10072 (107 عينة × 22,283 جين)، وGSE13576 (209 عينة × 54,675 جين). أظهر تقييم الأداء باستخدام F-measure ومؤشر Davies-Bouldin أداء DDR المتفوق مقارنة بالمقاييس التقليدية. حققت DDR قياس F أقصى مثالي (1.0) ومؤشر Davies-Bouldin الأمثل (2.2071) لـ GSE43346، أعلى متوسط F-measure (0.9883) لـ GSE10072، وأداء تنافسي لـ GSE13576. تظهر المنهجية تعقيداً حاسوبياً خطياً O(n)، مما يتيح التحليل الفعال لمجموعات البيانات الجينومية واسعة النطاق. تؤسس هذه النتائج DDR كإطار عمل قوي وقابل للتوسع وقابل للتفسير بيولوجياً مع آثار مهمة للطب الدقيق، اكتشاف المؤشرات الحيوية، وتحليل شبكات التنظيم الجيني.

الكلمات المفتاحية:حليل التعبير الجيني، مقاييس التشابه، البيانات عالية الأبعاد، علم الجينوم، التصنيف، اكتشاف المؤشرات الحيوية، الطب الدقيق، علم الأحياء الحاسوبي.

RESUME

La sélection de mesures de similarité appropriées pour les jeux de données d'expression génique de haute dimension représente un défi critique en génomique computationnelle, particulièrement avec l'évolution des technologies génomiques des plateformes de puces à ADN vers le séquençage ARN et les applications unicellulaires. Les métriques traditionnelles telles que la distance euclidienne, la distance de Manhattan, la corrélation de Pearson et la distance du Chi-carré s'avèrent souvent inadéquates face aux caractéristiques complexes des jeux de données génomiques modernes, incluant une haute dimensionnalité extrême, des motifs d'inflation de zéros étendus, des effets de lot systématiques et des profils de bruit hétérogènes. Pour adresser ces limitations fondamentales, nous avons développé la méthodologie Distance-DissimRatio (DDR), une mesure de similarité novatrice qui intègre l'analyse des déviations centrées sur la moyenne, l'évaluation des différences d'expression directes, et un cadre de normalisation compréhensif. La méthodologie DDR présente des caractéristiques uniques incluant l'indépendance d'échelle, les propriétés d'influence bornée, la robustesse à l'inflation de zéros, et une interprétabilité biologique améliorée. Une validation expérimentale complète a été menée sur trois jeux de données d'expression génique diversifiés : GSE43346 (68 échantillons × 54 675 gènes), GSE10072 (107 échantillons × 22 283 gènes), et GSE13576 (209 échantillons × 54 675 gènes). L'évaluation de performance utilisant la F-mesure et l'Index de Davies-Bouldin a démontré la performance supérieure de DDR comparée aux métriques traditionnelles. DDR a atteint une F-mesure maximale parfaite (1,0) et un index de Davies-Bouldin optimal (2,2071) pour GSE43346, la F-mesure moyenne la plus élevée (0,9883) pour GSE10072, et une performance compétitive pour GSE13576. La méthodologie démontre une complexité computationnelle linéaire O(n), permettant une analyse efficace des jeux de données génomiques à grande échelle. Ces résultats établissent DDR comme un cadre robuste, évolutif et biologiquement interprétable avec des implications significatives pour la médecine de précision, la découverte de biomarqueurs, et l'analyse des réseaux de régulation génique.

Mots-clés : Analyse de l'expression génique, mesures de similarité, données de haute dimension, génomique, classification, découverte de biomarqueurs, médecine de précision, biologiecomputationnelle.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CHISQ - Chi-Square (distance metric)

DBI - Davies-Bouldin Index

DDR - Distance-DissimRatio (proposed methodology)

DNA - Deoxyribonucleic Acid

EUC - Euclidean (distance metric)

GEO - Gene Expression Omnibus

GSE - Gene Expression Omnibus Series

H3K27me3 - Histone H3 Lysine 27 trimethylation

HG-U133A - Human Genome U133A (Affymetrix platform)

MANH - Manhattan (distance metric)

mRNA - messenger Ribonucleic Acid

NCBI - National Center for Biotechnology Information

NEK2 - NIMA Related Kinase 2 (gene)

PBTs - Pathogenesis-Based Transcript Sets

PRC1 - Protein Regulator of Cytokinesis 1 (gene)

PRC2 - Polycomb Repressive Complex 2

RNA - Ribonucleic Acid

RNA-seq - RNA sequencing

SCLC - Small Cell Lung Cancer

TTK - TTK Protein Kinase (gene)

WGCNA - Weighted Gene Co-expression Network Analysi

GENERAL INTRODUCTION

GENERAL INTRODUCTION

The unprecedented growth of high-throughput genomic technologies, from microarrays to single-cell RNA sequencing (scRNA-seq), has profoundly transformed biological research by enabling large-scale gene expression profiling. These datasets, often high-dimensional, sparse, and heterogeneous, contain complex biological signals that are essential for understanding cellular functions, disease mechanisms, and patient-specific molecular profiles. At the core of most downstream bioinformatics analyses — including clustering, classification, network reconstruction, and biomarker discovery — lies the fundamental task of quantifying similarity between gene expression profiles.

The biological interpretation of these similarity relationships is crucial because they reflect the underlying functional, regulatory, or phenotypic proximity between genes, cells, or biological samples. Accurately capturing these relationships ensures that subsequent analytical outcomes — such as the identification of disease subtypes or the discovery of molecular pathways — are biologically meaningful and clinically actionable.

However, a major challenge in contemporary genomics is that the statistical properties of modern datasets (zero-inflation, high dimensionality, noise, and heterogeneity) violate the assumptions of traditional similarity measures like Pearson correlation and Euclidean distance, originally designed for continuous, normally distributed, low-dimensional data. This mismatch risks obscuring genuine biological signals, producing unreliable results, and limiting the clinical translation of genomic discoveries.

Moreover, specialized applications like scRNA-seq and multi-omics integration require similarity measures tailored to capture biologically relevant patterns within highly sparse and noisy data while remaining robust to technical variability and outliers. In clinical genomics, where gene expression analyses inform patient stratification and treatment decisions, similarity measures must be both statistically sound and biologically interpretable to ensure reliable, reproducible, and ethically responsible outcomes.

Consequently, the biological aspect of this research lies in developing and rigorously evaluating similarity measures capable of accurately reflecting functional relationships and biological variability in diverse genomic contexts. This work seeks to bridge methodological rigor with biological relevance, ensuring that computational approaches serve to enhance — rather than distort — the interpretation of complex biological systems, ultimately supporting precision medicine, systems biology, and translational genomics.

CHAPTER ONE
BIOLOGICAL FOUNDATIONS OF GENE EXPRESSION AND TRANSCRIPTOMIC VARIATION

1 Introduction

Understanding the biological basis of gene expression and transcriptomic variation is essential to appreciate the significance of computational methods used to analyze gene expression datasets. This chapter introduces key biological concepts, explains the nature of transcriptomic data, and discusses why clustering is a crucial tool in analyzing such data.

2 Gene Expression: Definition and Biological Significance

Gene expression is the fundamental biological process through which the information encoded in DNA is converted into functional products, mainly proteins or functional RNA molecules. This process primarily involves transcription, where a gene's DNA sequence is copied into messenger RNA (mRNA), which then guides protein synthesis or performs regulatory roles (*NHGRI, 2025*).

Gene expression can be thought of as both an “on/off switch” determining whether a gene is active, and a “volume control” regulating the amount of gene product produced. It is tightly regulated and varies between cell types, developmental stages, and environmental conditions, reflecting the dynamic nature of cellular function.

3 Transcriptomics: Exploring the Transcriptome

Transcriptomics is the study of the transcriptome — the complete set of RNA transcripts produced by the genome under specific conditions or in particular cell types (*Microbe Notes, 2024*). Unlike the static genome, the transcriptome is dynamic and reflects the genes actively expressed at any given time.

Transcriptomics includes the analysis of various RNA types: mRNA, ribosomal RNA (rRNA), transfer RNA (tRNA), and non-coding RNAs that regulate gene expression. Modern high-throughput technologies such as DNA microarrays and RNA sequencing (RNA-Seq) enable the simultaneous measurement of thousands of transcripts, providing a snapshot of cellular activity (*Longdom, 2024*).

4 Transcriptomic Data and Its Biological Applications

Transcriptomic data reveal how genes respond to internal and external stimuli, enabling insights into cellular processes, disease mechanisms, and developmental biology. For example, changes in gene expression patterns can indicate disease states or responses to treatments.

One major application of transcriptomic data is clustering, which groups genes or samples based on expression similarity. This helps identify gene modules with related functions or classify biological samples (*Yousefi et al., 2024*).

5 DNA Microarrays and Transcriptomic Data Generation

DNA microarrays are one of the earliest and widely used technologies for transcriptomic profiling. They consist of thousands of DNA probes fixed on a chip, which hybridize with labeled RNA from samples. The resulting fluorescence intensities form a matrix representing gene expression levels across samples.

The output matrix from microarrays is a high-dimensional dataset where rows correspond to genes and columns to samples. This matrix is the starting point for computational analyses such as clustering, which seeks to uncover patterns in gene expression (*Bolshakova et al., 2005*).

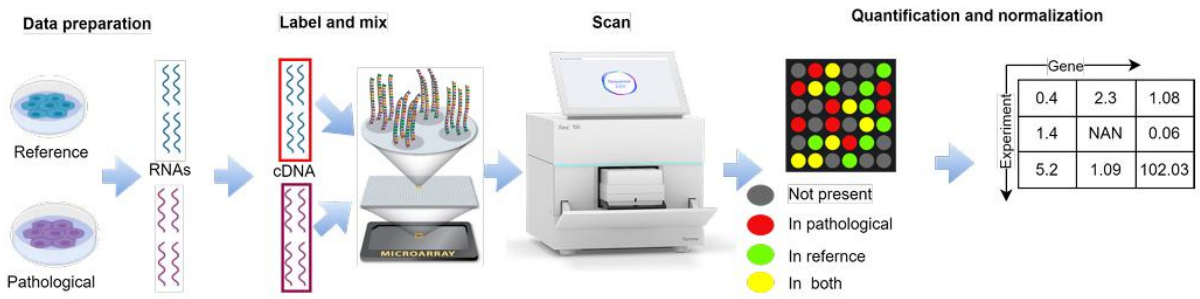

Figure 2 General workflow of DNA microarrays.

6 Why Clustering Is Used in Biological Data Analysis

Clustering is a powerful unsupervised learning technique that organizes data points into groups (clusters) based on similarity. In transcriptomics, clustering is used to:

- Identify groups of co-expressed genes that may share regulatory mechanisms or biological functions.
- Classify samples into biologically meaningful subtypes.
- Reduce data complexity to facilitate interpretation.

Because transcriptomic datasets are typically high-dimensional and noisy, clustering helps reveal underlying biological structure and generate hypotheses for further validation (*Yousefi et al., 2024*).

7 Bioinformatics Approaches in Clustering Gene Expression Data

Bioinformatics integrates computational tools and biological knowledge to analyze transcriptomic data. Various clustering algorithms (hierarchical, k-means, consensus clustering) are applied to gene expression matrices to identify meaningful patterns.

Recent advances include methods tailored for long-read RNA sequencing data, improving clustering accuracy and computational efficiency (*Ma et al., 2024*). Consensus clustering techniques increase robustness by aggregating results from multiple clustering runs, addressing variability in biological data (*Yousefi et al., 2024*)

CHAPTER TWO
SIMILARITY METHODS USED IN HIGH-DIMENSIONAL GENOMIC DATA (RELATED WORKS)

1 Euclidean Distance in Gene Expression Dataset Analysis

1.1 Preamble

Euclidean distance represents one of the most fundamental and widely applied distance metrics in computational biology, particularly in the analysis of high-dimensional gene expression datasets (*Hastie et al., 2009*). As genomic technologies have advanced from microarrays to RNA sequencing, researchers have increasingly relied on distance-based methods to quantify similarities and differences between gene expression profiles across samples, conditions, or time points (*Quackenbush, 2001*). The concept of Euclidean distance, rooted in classical geometry and extended to n-dimensional space, provides an intuitive and mathematically tractable approach for measuring the dissimilarity between gene expression vectors.

In the context of gene expression analysis, Euclidean distance serves as the foundation for numerous analytical techniques, including hierarchical clustering, k-means clustering, principal component analysis, and nearest neighbor classification methods (*Eisen et al., 1998; Tamayo et al., 1999*). The metric's geometric interpretation allows researchers to conceptualize gene expression data as points in high-dimensional space, where each dimension represents the expression level of a specific gene, and the distance between points reflects the overall similarity of expression patterns (*D'haeseleer, 2005*).

The application of Euclidean distance to gene expression datasets presents both opportunities and challenges that merit careful consideration. While its computational simplicity and intuitive interpretation make it an attractive choice for exploratory data analysis, the high-dimensional nature of genomic data and the presence of noise, outliers, and correlated variables can impact its effectiveness (*Jiang et al., 2004*). Understanding these strengths and limitations is crucial for researchers seeking to apply distance-based methods appropriately in their genomic analyses.

1.2 Mathematical Formula

The Euclidean distance between two gene expression profiles, represented as vectors **x** and **y** in n-dimensional space, is defined mathematically as (*Duda et al., 2001*):

$$\mathbf{d(x, y) = \sqrt{[\Sigma_{i=1}^{n} (x_i - y_i)^2]}}$$

Where:

- $\mathbf{x} = (x_1, x_2, ..., x_n)$ represents the expression levels of n genes in sample x

- $\mathbf{y} = (y_1, y_2, ..., y_n)$ represents the expression levels of n genes in sample y
- x_i and y_i are the expression values for gene i in samples x and y, respectively
- n is the total number of genes being compared

For computational implementation, the squared Euclidean distance is often used to avoid the computational cost of the square root operation:

$$\mathbf{d^2(x, y) = \Sigma_{i=1}^{n} (x_i - y_i)^2}$$

This squared version preserves the relative ordering of distances while reducing computational complexity, making it particularly suitable for large-scale genomic datasets (Hastie et al., 2009).

1.3 Strengths

Computational Efficiency and Scalability

Euclidean distance offers significant computational advantages in gene expression analysis, with a time complexity of O(n) for comparing two samples with n genes (*Tan et al., 2005*). This efficiency becomes particularly important when analyzing large-scale datasets containing thousands of genes and hundreds of samples, where pairwise distance calculations can quickly become computationally intensive. The straightforward arithmetic operations required for Euclidean distance calculation make it amenable to parallel processing and hardware acceleration, enabling analysis of increasingly large genomic datasets (*Bar-Joseph et al., 2001*).

Intuitive Geometric Interpretation

The geometric foundation of Euclidean distance provides researchers with an intuitive framework for understanding relationships between gene expression profiles (*Brazma & Vilo, 2000*). Samples with similar expression patterns cluster together in the high-dimensional gene expression space, while dissimilar samples are positioned farther apart. This spatial metaphor facilitates the interpretation of clustering results and helps researchers identify groups of samples with similar biological characteristics or experimental conditions.

Mathematical Properties and Analytical Tractability

Euclidean distance satisfies the mathematical properties of a true metric, including symmetry, non-negativity, and the triangle inequality (*Deza & Deza, 2009*). These properties ensure consistent and predictable behavior in downstream analyses and enable the application of

established mathematical frameworks from metric space theory. Additionally, the relationship between Euclidean distance and correlation-based measures provides theoretical connections to other commonly used similarity metrics in gene expression analysis (*Jaskowiak et al., 2014*).

Sensitivity to Magnitude Differences

In gene expression analysis, Euclidean distance is sensitive to both the magnitude and direction of expression changes, making it particularly useful for identifying samples with dramatically different expression levels (*Jiang et al., 2004*). This sensitivity can be advantageous when analyzing datasets where the absolute magnitude of expression changes is biologically meaningful, such as dose-response studies or time-course experiments where progressive changes in gene expression are expected.

Foundation for Advanced Methods

Euclidean distance serves as the foundation for numerous sophisticated analytical techniques commonly applied to gene expression data, including k-means clustering, hierarchical clustering with average linkage, and various machine learning algorithms (*Jain et al., 1999*). The widespread adoption of these methods in the genomics community means that results based on Euclidean distance can be easily compared across studies and integrated with existing analytical workflows.

1.4 Limitations

Curse of Dimensionality

Gene expression datasets typically contain thousands to tens of thousands of genes, creating a high-dimensional space where Euclidean distance can become less discriminative (*Beyer et al., 1999*). In high-dimensional spaces, the phenomenon known as the "curse of dimensionality" causes all pairwise distances to become increasingly similar, reducing the method's ability to distinguish between truly similar and dissimilar samples (*Aggarwal et al., 2001*). This limitation is particularly problematic in genomic datasets where the number of genes often far exceeds the number of samples.

Sensitivity to Noise and Outliers

The squared terms in the Euclidean distance formula make it highly sensitive to outliers and measurement noise, which are common in gene expression datasets due to technical variability in measurement platforms (*Quackenbush, 2002*). A single gene with an aberrant expression value can disproportionately influence the overall distance calculation, potentially

leading to misclassification of samples or incorrect clustering results. This sensitivity is particularly concerning given the inherent noise in both microarray and RNA-sequencing technologies.

Scale Dependency and Normalization Requirements

Euclidean distance is sensitive to the scale and dynamic range of gene expression values, which can vary dramatically across different genes (*Bolstad et al., 2003*). Genes with naturally higher expression levels or greater variability can dominate the distance calculation, potentially masking important patterns in genes with lower but biologically significant expression changes. This limitation necessitates careful data preprocessing and normalization, which can introduce additional complexity and potential sources of bias into the analysis pipeline.

Assumption of Linear Relationships

The geometric foundation of Euclidean distance assumes that relationships between variables are fundamentally linear and that equal weights should be given to differences in all dimensions (*Jain & Dubes, 1988*). However, gene expression data often exhibit complex, non-linear relationships due to regulatory networks, feedback loops, and other biological processes. This assumption may lead to suboptimal performance when analyzing datasets where gene interactions follow non-linear patterns.

Inability to Capture Correlation Patterns

Euclidean distance focuses on the magnitude of differences between corresponding genes but does not directly capture correlation patterns or co-expression relationships that are often biologically meaningful (*D'haeseleer et al., 2000*). Two samples might have very different absolute expression levels but exhibit similar patterns of relative gene expression changes, which would be considered dissimilar by Euclidean distance despite potentially representing similar biological states or responses.

Limited Biological Interpretation

While mathematically well-defined, Euclidean distance may not always correspond to meaningful biological differences between samples (*Jaskowiak et al., 2014*). The metric treats all genes as equally important and independent, which may not reflect the hierarchical organization of biological systems or the varying functional importance of different genes in specific biological processes.

2 Manhattan Distance in Gene Expression Dataset Analysis

2.1 Preamble

Manhattan distance represents a fundamental alternative to Euclidean distance in the analysis of high-dimensional gene expression datasets (*Aggarwal et al., 2001*). Named after the grid-like street pattern of Manhattan, this distance metric measures the sum of absolute differences between corresponding elements of two vectors, providing a unique geometric perspective on similarity relationships in genomic data (*Duda et al., 2001*). As computational biology has evolved to handle increasingly complex and noisy datasets, Manhattan distance has emerged as a robust alternative that offers distinct advantages in specific analytical contexts.

The adoption of Manhattan distance in gene expression analysis stems from its mathematical properties that make it particularly well-suited for handling the challenges inherent in genomic datasets (*Jaskowiak et al., 2014*). Unlike Euclidean distance, which emphasizes large differences through squared terms, Manhattan distance treats all deviations equally regardless of magnitude, potentially providing more balanced representations of expression pattern differences (*Tan et al., 2005*). This characteristic has proven valuable in applications ranging from gene regulatory network inference to sample classification, where the preservation of subtle but consistent expression changes across multiple genes may be more biologically meaningful than dramatic changes in individual genes.

The growing recognition of Manhattan distance in genomics research reflects broader trends toward robust statistical methods that can handle the noise, outliers, and high dimensionality characteristic of modern gene expression datasets (*Bar-Joseph et al., 2001*). RNA-sequencing technologies, in particular, have introduced new sources of variability and sparse expression patterns that can benefit from distance metrics less sensitive to extreme values (*Robinson et al., 2010*). Furthermore, the discrete nature of RNA-seq count data aligns well with Manhattan distance's focus on absolute differences rather than squared deviations.

Contemporary applications of Manhattan distance in gene expression analysis extend beyond traditional clustering and classification tasks to include novel approaches in single-cell genomics, time-series analysis, and comparative genomics (*Kiselev et al., 2017*). The metric's computational efficiency and interpretability have made it particularly attractive for exploratory data analysis and visualization techniques, where researchers seek to identify meaningful patterns in complex, multi-dimensional expression landscapes while maintaining analytical tractability.

2.2 Mathematical Formula

The Manhattan distance between two gene expression profiles, represented as vectors **x** and **y** in n-dimensional space, is mathematically defined as (*Hastie et al., 2009*):

$$\mathbf{d(x, y) = \Sigma_{i=1}^{n} |x_i - y_i|}$$

Where:

- $\mathbf{x} = (x_1, x_2, ..., x_n)$ represents the expression levels of n genes in sample x
- $\mathbf{y} = (y_1, y_2, ..., y_n)$ represents the expression levels of n genes in sample y
- x_i and y_i are the expression values for gene i in samples x and y, respectively
- $|x_i - y_i|$ denotes the absolute value of the difference between expression levels
- n is the total number of genes being compared

2.3 Strengths

Robustness to Outliers and Extreme Values

Manhattan distance demonstrates superior robustness compared to Euclidean distance when analyzing gene expression datasets containing outliers or extreme expression values (*Huber & Ronchetti, 2009*). The use of absolute differences rather than squared differences prevents individual genes with aberrant expression levels from disproportionately influencing the overall distance calculation. This property is particularly valuable in RNA-sequencing datasets, where technical artifacts, low-abundance transcripts, or highly variable genes can introduce extreme values that might mislead Euclidean-based analyses (*Bullard et al., 2010*).

Computational Efficiency and Simplicity

The computational requirements for Manhattan distance are minimal, involving only addition and absolute value operations without the need for expensive square root calculations required by Euclidean distance (*Tan et al., 2005*). This efficiency becomes particularly important when analyzing large-scale genomic datasets with thousands of genes and samples, where the reduced computational complexity can significantly impact analysis runtime. The simplicity of the calculation also makes Manhattan distance amenable to hardware acceleration and parallel processing architectures commonly used in bioinformatics applications.

Performance in High-Dimensional Spaces

Manhattan distance exhibits better discriminative power than Euclidean distance in high-dimensional gene expression datasets, where the curse of dimensionality can cause all

pairwise distances to become similar (*Aggarwal et al., 2001*). The linear rather than quadratic relationship between individual gene differences and overall distance helps maintain meaningful distinctions between samples even when analyzing datasets with tens of thousands of genes. This property is particularly important in modern genomics applications where the dimensionality often far exceeds the number of available samples.

Suitability for Sparse Data

The mathematical properties of Manhattan distance make it particularly well-suited for analyzing sparse gene expression datasets, such as single-cell RNA-sequencing data where many genes show zero or near-zero expression in individual cells (*Kiselev et al., 2017*). Unlike Euclidean distance, which can be dominated by a few highly expressed genes, Manhattan distance provides a more balanced representation of expression patterns across all detected genes, including those with low but potentially significant expression levels.

2.4 Limitations

Loss of Geometric Intuition

While Manhattan distance offers computational advantages, it sacrifices some of the geometric intuition provided by Euclidean distance (*Jain & Dubes, 1988*). The restriction to axis-aligned movement creates diamond-shaped rather than circular distance contours, which can be less intuitive for researchers accustomed to thinking about similarity in terms of traditional geometric proximity. This altered geometric interpretation can complicate the visualization and understanding of clustering results, particularly when communicating findings to audiences familiar with Euclidean-based analyses.

Insensitivity to Magnitude Differences

The equal weighting of all absolute differences in Manhattan distance can be problematic when analyzing gene expression datasets where the magnitude of expression changes carries biological significance (*Quackenbush, 2001*). Large expression differences that might indicate important biological responses receive the same per-unit contribution as small differences that could represent technical noise. This limitation can be particularly problematic in dose-response studies or time-course experiments where the magnitude of expression changes is expected to correlate with biological effect strength.

Reduced Sensitivity to Correlated Expression Patterns

Manhattan distance focuses on absolute differences in individual gene expression levels but does not explicitly account for correlation patterns or co-expression relationships that are

often biologically meaningful (*Eisen et al., 1998*). Two samples with strongly correlated expression patterns but different absolute levels might be considered dissimilar by Manhattan distance, potentially missing important biological relationships based on coordinated gene expression changes across functional pathways or regulatory networks.

Scale Dependency and Normalization Challenges

Like other distance metrics, Manhattan distance remains sensitive to the scale and dynamic range of gene expression measurements, requiring careful data preprocessing and normalization (*Bolstad et al., 2003*). However, the linear nature of Manhattan distance can make it more sensitive to systematic biases in normalization procedures, as errors in scale adjustment are directly propagated without the dampening effect of squared terms. This sensitivity necessitates robust normalization strategies that may be more complex than those required for Euclidean-based analyses.

Limited Theoretical Framework

Compared to Euclidean distance, Manhattan distance has a less developed theoretical framework in the context of statistical analysis and machine learning (*Deza & Deza, 2009*). Many established statistical methods and theoretical results are based on Euclidean geometry, making it more challenging to apply advanced analytical techniques or derive theoretical guarantees for Manhattan distance-based approaches. This limitation can restrict the availability of sophisticated analytical tools and limit the depth of statistical inference possible with Manhattan distance-based methods.

Potential for Masking Subtle Patterns

The equal treatment of all gene expression differences in Manhattan distance can potentially mask subtle but biologically important patterns that involve coordinated small changes across multiple genes (*D'haeseleer et al., 2000*). Biological processes often involve modest but consistent expression changes across gene sets or pathways, which might be overwhelmed by larger but less biologically significant changes in individual genes when using Manhattan distance. This limitation can be particularly problematic when analyzing regulatory networks or signaling pathways where small, coordinated changes are functionally important.

Incompatibility with Certain Analytical Methods

Some advanced analytical techniques commonly used in gene expression analysis, such as principal component analysis or certain kernel methods, are specifically designed around Euclidean distance assumptions (*Hastie et al., 2009*). The use of Manhattan distance may

require alternative analytical approaches or modifications to existing methods, potentially limiting the range of available analytical tools or requiring custom implementation of distance-aware algorithms.

Reduced Power for Certain Biological Questions

In some biological contexts, particularly those involving dramatic expression changes or stress responses, the squared terms in Euclidean distance may actually provide better discrimination between biologically relevant groups (*Jiang et al., 2004*). Manhattan distance's linear treatment of differences might reduce the power to detect samples with extreme but biologically meaningful expression profiles, potentially missing important biological insights in studies focused on strong phenotypic responses or disease states.

3 Chi-Square Distance in Gene Expression Dataset Analysis

3.1 Preamble

Chi-Square distance represents a specialized distance metric that has gained increasing relevance in gene expression analysis, particularly with the advent of RNA-sequencing technologies that generate discrete count data rather than continuous expression measurements (*Robinson et al., 2010*). Originally developed for analyzing contingency tables and frequency distributions in statistics, Chi-Square distance has found novel applications in genomics where its mathematical properties align well with the discrete, non-negative nature of transcript count data (*Anders & Huber, 2010*). This distance metric provides a unique perspective on gene expression similarity by incorporating variance normalization that accounts for the inherent relationship between mean expression levels and their associated variability.

The adoption of Chi-Square distance in gene expression analysis reflects the evolving landscape of genomic technologies and analytical approaches (*Conesa et al., 2016*). Traditional microarray-based expression analysis relied heavily on log-transformed, normalized intensity values that were well-suited to Euclidean and correlation-based distance metrics. However, RNA-sequencing has fundamentally altered the data structure by providing discrete count measurements that exhibit characteristic mean-variance relationships and zero-inflation patterns (*Love et al., 2014*). Chi-Square distance addresses these challenges by providing a framework that naturally accommodates the count-based nature of RNA-seq data while offering built-in normalization for expression-dependent variance.

Contemporary applications of Chi-Square distance in genomics extend beyond traditional differential expression analysis to include novel areas such as single-cell RNA-sequencing, where the discrete and sparse nature of expression measurements aligns particularly well with the metric's mathematical foundations (*Kiselev et al., 2017*). The distance metric's ability to handle zero values and its emphasis on relative rather than absolute expression differences make it valuable for analyzing datasets where genes exhibit highly variable expression patterns or where technical factors introduce systematic biases in count distributions.

The theoretical foundation of Chi-Square distance in statistical analysis provides genomics researchers with a principled approach to measuring expression similarity that connects directly to established statistical frameworks for hypothesis testing and significance assessment (*Agresti, 2013*). This connection enables researchers to leverage decades of statistical theory development while addressing the specific challenges posed by high-throughput genomic datasets. Furthermore, the metric's relationship to chi-square statistics facilitates the development of statistical tests and confidence intervals that can provide formal assessments of expression pattern significance.

3.2 Mathematical Formula

The Chi-Square distance between two gene expression profiles, represented as vectors **x** and **y** in n-dimensional space, is mathematically defined as (*Greenacre, 2007*):

$$\mathbf{d(x, y) = \Sigma_{i=1}^{n} [(x_i - y_i)^2 / (x_i + y_i)]}$$

Where:

- $\mathbf{x} = (x_1, x_2, ..., x_n)$ represents the expression levels of n genes in sample x
- $\mathbf{y} = (y_1, y_2, ..., y_n)$ represents the expression levels of n genes in sample y
- x_i and y_i are the non-negative expression values for gene i in samples x and y, respectively
- n is the total number of genes being compared
- The denominator $(x_i + y_i)$ provides variance normalization based on expression magnitude

The mathematical foundation of Chi-Square distance derives from the chi-square statistic used in contingency table analysis, where it measures the deviation between observed and expected frequencies (*Agresti, 2013*). In the context of gene expression analysis, this

translates to measuring the magnitude of expression differences while normalizing for the expected variance associated with different expression levels.

The geometric interpretation of Chi-Square distance corresponds to a weighted Euclidean distance where the weighting factor inversely relates to the sum of expression values, effectively down-weighting differences between highly expressed genes and emphasizing differences in lowly expressed genes (*Greenacre, 2007*). This property makes the metric particularly suitable for analyzing datasets where expression levels span several orders of magnitude.

3.3 Strengths

Natural Accommodation of Count Data Structure

Chi-Square distance is uniquely suited for analyzing RNA-sequencing count data due to its mathematical foundation in discrete probability distributions (*Robinson et al., 2010*). Unlike Euclidean or Manhattan distances, which treat count data as continuous variables, Chi-Square distance inherently accommodates the discrete, non-negative nature of transcript counts. The variance normalization built into the distance formula naturally accounts for the mean-variance relationship characteristic of count data, where genes with higher expression levels typically exhibit greater variability (*Anders & Huber, 2010*).

Effective Handling of Zero and Low Expression Values

The Chi-Square distance formulation provides robust handling of genes with zero or very low expression levels, which are common in RNA-sequencing datasets, particularly in single-cell applications (*Kiselev et al., 2017*). The denominator normalization prevents genes with consistently low expression from being dominated by technical noise while still allowing meaningful comparisons between samples. This property is particularly valuable when analyzing sparse expression matrices where many genes show zero expression in individual samples or conditions.

Emphasis on Relative Expression Patterns

Chi-Square distance emphasizes relative rather than absolute expression differences, making it particularly valuable for identifying samples with similar expression patterns despite different overall expression magnitudes (*Conesa et al., 2016*). This property is advantageous when comparing samples from different experimental conditions, tissues, or developmental stages where systematic differences in overall expression levels might mask important

biological similarities. The relative emphasis also makes the metric less sensitive to global normalization artifacts that can affect absolute expression measurements.

Statistical Interpretability and Theoretical Framework

The connection between Chi-Square distance and established statistical theory provides researchers with a principled framework for interpreting results and assessing significance (*Agresti, 2013*). The relationship to chi-square statistics enables the development of formal statistical tests and confidence intervals, facilitating more rigorous statistical inference than is possible with purely geometric distance metrics. This theoretical foundation also enables integration with established statistical methods for multiple testing correction and false discovery rate control.

Computational Efficiency for Sparse Data

Chi-Square distance can be computed efficiently for sparse gene expression matrices, as zero values contribute nothing to the distance calculation (*Greenacre, 2007*). This efficiency is particularly important when analyzing large-scale single-cell RNA-sequencing datasets where the majority of gene-cell combinations may be zero. The sparse computation capability enables analysis of datasets with millions of cells and tens of thousands of genes that would be computationally prohibitive with dense matrix operations.

Robustness to Outliers in Low Expression Ranges

The variance normalization in Chi-Square distance provides natural robustness to outliers in genes with low expression levels, where small absolute differences might represent large relative changes (*Robinson et al., 2010*). This robustness is particularly valuable when analyzing noisy datasets or when technical artifacts introduce spurious expression values in genes with naturally low expression levels. The metric's behavior helps distinguish between genuine biological variation and technical noise in low-abundance transcripts.

3.4 Limitations

Restriction to Non-Negative Data

Chi-Square distance requires strictly non-negative input values, limiting its applicability to certain types of gene expression data (*Greenacre, 2007*). This restriction prevents direct application to log-transformed expression data, fold-change measurements, or any analysis involving negative values such as differential expression scores or standardized expression profiles. Researchers must carefully consider data preprocessing steps to ensure compatibility with Chi-Square distance requirements while preserving biologically meaningful information.

Sensitivity to Low Expression Genes

While Chi-Square distance provides variance normalization, it can become overly sensitive to differences in genes with very low expression levels, where small absolute changes result in disproportionately large distance contributions (*Anders & Huber, 2010*). This sensitivity can lead to clustering or classification results that are dominated by technical noise in low-abundance transcripts rather than biologically meaningful expression changes. The issue is particularly problematic in RNA-sequencing datasets where many genes exhibit very low or sporadic expression patterns.

Computational Instability Near Zero

The denominator in the Chi-Square distance formula can create numerical instability when both expression values approach zero, requiring careful implementation with appropriate regularization terms (*Love et al., 2014*). Different choices of regularization constants can significantly impact results, particularly for genes with consistently low expression across samples. This computational challenge necessitates careful parameter tuning and validation to ensure robust and reproducible results across different datasets and analysis platforms.

Limited Applicability to Correlation-Based Analysis

Chi-Square distance does not directly capture correlation patterns or co-expression relationships that are central to many gene expression analysis applications (*Eisen et al., 1998*). The metric focuses on magnitude differences rather than expression pattern similarities, potentially missing important biological relationships based on coordinated expression changes across functional gene sets or regulatory networks. This limitation can be particularly problematic when analyzing gene regulatory networks or pathway-based expression patterns.

Potential Over-Emphasis on Low Expression Differences

The variance normalization in Chi-Square distance can sometimes over-emphasize differences between samples in genes with consistently low expression levels, where biological significance may be limited (*Conesa et al., 2016*). This over-emphasis can lead to sample clustering or classification based on technical noise rather than meaningful biological differences, particularly when analyzing datasets with high levels of technical variability or when comparing samples with different library preparation methods.

Limited Integration with Standard Genomics Workflows

Many established gene expression analysis tools and pipelines are designed around Euclidean or correlation-based distance metrics, making integration of Chi-Square distance more challenging (*Hastie et al., 2009*). The specialized nature of the metric may require custom implementation or modification of existing analytical workflows, potentially limiting its accessibility to researchers without strong computational backgrounds. This limitation can hinder adoption and comparison with results from standard analytical approaches.

Interpretation Challenges for Biological Validation

While Chi-Square distance has strong statistical foundations, translating distance values to biological interpretations can be challenging for researchers unfamiliar with the metric's mathematical properties (*Jaskowiak et al., 2014*). The variance-normalized nature of the distance can make it difficult to assess whether observed differences represent meaningful biological variation or technical artifacts, particularly when validating computational results through experimental approaches. This interpretation challenge can complicate the communication of results to collaborators or reviewers unfamiliar with the metric.

Scale Dependency Despite Normalization

Although Chi-Square distance includes built-in variance normalization, it can still exhibit scale dependency when comparing datasets with dramatically different expression ranges or when analyzing data from different technological platforms (*Robinson et al., 2010*). Cross-platform comparisons or meta-analyses may require additional normalization steps beyond those provided by the distance metric itself, potentially complicating comparative studies or data integration efforts.

Limited Performance with Highly Variable Genes

Chi-Square distance may not optimally handle genes with extremely high variance relative to their mean expression levels, such as those involved in stress responses or developmental transitions (*Anders & Huber, 2010*). The variance normalization assumes a particular relationship between mean and variance that may not hold for all biological processes, potentially leading to suboptimal performance when analyzing datasets enriched for highly dynamic gene expression patterns.

4 Pearson Correlation Coefficient in Gene Expression Dataset Analysis

4.1 Preamble

The Pearson correlation coefficient represents one of the most fundamental and widely applied similarity measures in gene expression analysis, serving as a cornerstone for understanding co-expression patterns, gene regulatory networks, and functional relationships within genomic datasets (*Eisen et al., 1998*). Unlike distance metrics that quantify dissimilarity between expression profiles, the Pearson correlation coefficient measures the strength and direction of linear relationships between gene expression patterns, providing insights into coordinated biological processes and regulatory mechanisms (*Stuart et al., 2003*). This shift from measuring absolute differences to capturing expression pattern similarities has proven invaluable for identifying functionally related genes, reconstructing biological pathways, and understanding the complex regulatory architecture underlying cellular processes.

The widespread adoption of Pearson correlation in genomics reflects its unique ability to identify genes that are co-regulated or co-expressed, regardless of their absolute expression levels (*D'haeseleer et al., 2000*). This property is particularly valuable in gene expression analysis, where biological processes often involve coordinated changes in gene expression that maintain consistent relative relationships across different experimental conditions, developmental stages, or environmental perturbations. The scale-invariant nature of correlation analysis enables researchers to identify meaningful biological relationships that might be obscured by absolute expression differences when using traditional distance-based approaches.

Contemporary applications of Pearson correlation in gene expression analysis extend far beyond simple pairwise gene comparisons to encompass sophisticated network-based approaches for understanding systems-level biological organization (*Langfelder & Horvath, 2008*). Weighted gene co-expression network analysis (WGCNA), module detection algorithms, and pathway enrichment methods all rely heavily on correlation-based similarity measures to identify functional gene modules and regulatory networks. These approaches have been instrumental in advancing our understanding of complex diseases, developmental biology, and evolutionary genomics by revealing the modular organization of gene expression programs.

The integration of Pearson correlation with modern high-throughput genomic technologies has enabled researchers to construct comprehensive gene regulatory networks and identify novel therapeutic targets across diverse biological systems (*Margolin et al., 2006*). Single-cell RNA-sequencing, in particular, has benefited from correlation-based approaches that can identify cell-type-specific co-expression patterns and reconstruct developmental trajectories based on coordinated gene expression changes. Furthermore, the statistical framework underlying Pearson correlation provides researchers with established methods for significance testing, multiple comparison correction, and confidence interval estimation that are essential for rigorous genomic analyses.

The mathematical simplicity and interpretability of Pearson correlation, combined with its deep statistical foundations, have made it an indispensable tool for exploratory data analysis and hypothesis generation in genomics research (*Quackenbush, 2001*). Its ability to capture linear relationships while remaining robust to systematic scaling differences has proven particularly valuable for cross-platform comparisons, meta-analyses, and integrative genomics studies that combine datasets from different experimental sources or technological platforms.

4.2 Mathematical Formula

The Pearson correlation coefficient between two gene expression profiles, represented as vectors **x** and **y** with n observations, is mathematically defined as:

$$\mathbf{r(x, y) = \Sigma_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) / \sqrt{[\Sigma_{i=1}^{n} (x_i - \bar{x})^2 \times \Sigma_{i=1}^{n} (y_i - \bar{y})^2]}}$$

Where:

- $\mathbf{x} = (x_1, x_2, ..., x_n)$ represents expression values for gene x across n samples
- $\mathbf{y} = (y_1, y_2, ..., y_n)$ represents expression values for gene y across n samples
- $\bar{x} = (1/n)\Sigma_{i=1}^{n} x_i$ is the sample mean of gene x expression values
- $\bar{y} = (1/n)\Sigma_{i=1}^{n} y_i$ is the sample mean of gene y expression values
- n is the number of samples or experimental conditions being compared

The Pearson correlation coefficient ranges from -1 to +1, where:

- r = +1 indicates perfect positive linear correlation
- r = 0 indicates no linear correlation
- r = -1 indicates perfect negative linear correlation

In gene expression analysis, correlation values are often interpreted as measures of co-expression strength, with values closer to ±1 indicating stronger co-expression relationships and values near zero suggesting independent expression patterns (*Stuart et al., 2003*).

4.3 Strengths

Scale Invariance and Normalization Independence

The Pearson correlation coefficient exhibits complete scale invariance, making it insensitive to linear transformations of gene expression data (*Rodgers & Nicewander, 1988*). This property is particularly valuable in gene expression analysis, where different genes naturally exhibit vastly different expression magnitudes spanning several orders of magnitude. Unlike distance-based metrics that can be dominated by highly expressed genes, correlation analysis focuses on expression patterns rather than absolute levels, enabling meaningful comparisons between genes with different baseline expression levels (*Quackenbush, 2001*).

Identification of Co-Expression Patterns

Pearson correlation excels at identifying genes that exhibit coordinated expression changes across experimental conditions, regardless of their absolute expression levels (*Eisen et al., 1998*). This capability is fundamental for understanding biological processes that involve coordinated gene regulation, such as metabolic pathways, stress responses, or developmental programs. The ability to detect co-expression patterns has proven invaluable for functional annotation of unknown genes, pathway reconstruction, and identification of gene regulatory modules (*Stuart et al., 2003*).

Robust Statistical Framework

The Pearson correlation coefficient benefits from a well-established statistical framework that enables rigorous significance testing, confidence interval estimation, and multiple comparison correction (*Cohen et al., 2003*). The availability of parametric and non-parametric statistical tests for correlation significance provides researchers with principled approaches for distinguishing meaningful co-expression relationships from random associations. This statistical foundation is particularly important in genomics applications where multiple testing issues require careful consideration of false discovery rates.

Computational Efficiency for Large Datasets

Correlation calculations can be performed efficiently using optimized linear algebra libraries, making Pearson correlation suitable for analyzing large-scale gene expression datasets with thousands of genes and samples (*Langfelder & Horvath, 2008*). Matrix-based

implementations enable rapid computation of genome-wide correlation matrices, facilitating comprehensive co-expression network construction and module detection analyses. The computational efficiency is particularly important for single-cell RNA-sequencing applications where datasets may contain millions of cells and tens of thousands of genes.

Interpretability and Biological Relevance

Correlation coefficients provide intuitive measures of relationship strength that can be easily interpreted by researchers without extensive statistical training (*Cohen et al., 2003*). The standardized scale (-1 to +1) enables consistent interpretation across different datasets and experimental contexts, facilitating meta-analyses and cross-study comparisons. Furthermore, the biological interpretation of positive correlations as co-activation and negative correlations as co-repression aligns well with established concepts in gene regulation and systems biology.

Foundation for Network Analysis

Pearson correlation serves as the foundation for sophisticated network-based approaches to gene expression analysis, including weighted gene co-expression network analysis (WGCNA) and module detection algorithms (*Langfelder & Horvath, 2008*). These network-based methods leverage correlation relationships to identify functional gene modules, hub genes, and hierarchical organization within biological systems. The resulting network structures provide powerful frameworks for understanding systems-level organization and identifying key regulatory components.

Sensitivity to Linear Relationships

Pearson correlation is optimally designed to detect linear relationships between gene expression patterns, which are common in biological systems involving proportional responses, dose-dependent effects, or coordinated regulatory mechanisms (*D'haeseleer et al., 2000*). This sensitivity to linear patterns makes it particularly effective for identifying gene pairs or modules that respond proportionally to experimental perturbations or environmental changes.

4.4 Limitations

Assumption of Linear Relationships

Pearson correlation is specifically designed to measure linear relationships and may fail to detect meaningful biological associations that follow non-linear patterns (*Reshef et al., 2011*). Many biological processes involve complex, non-linear regulatory relationships, such as sigmoidal dose-response curves, oscillatory expression patterns, or threshold-based regulatory

switches. These non-linear relationships may exhibit strong biological significance but receive low correlation scores, potentially causing researchers to overlook important regulatory connections.

Sensitivity to Outliers

The least-squares foundation of Pearson correlation makes it highly sensitive to outliers and extreme values, which can dramatically influence correlation estimates (*Rousseeuw & Leroy, 2003*). In gene expression datasets, outliers may arise from technical artifacts, sample contamination, or genuine biological variation, but their impact on correlation calculations can be disproportionate to their biological significance. This sensitivity can lead to spurious correlations or mask genuine co-expression relationships, particularly in datasets with small sample sizes.

Requirement for Adequate Sample Size

Accurate estimation of Pearson correlation coefficients requires sufficient sample sizes to achieve statistical power and stability (*Cohen et al., 2003*). Many gene expression studies, particularly those involving primary tissue samples or clinical cohorts, may have limited sample sizes that compromise the reliability of correlation estimates. Small sample sizes can lead to overestimation of correlation strength and increased susceptibility to false positive associations, particularly when combined with multiple testing across thousands of gene pairs.

Inability to Distinguish Direct from Indirect Relationships

Pearson correlation cannot distinguish between direct regulatory relationships and indirect associations mediated through intermediate genes or pathways (*D'haeseleer et al., 2000*). High correlation between two genes may result from co-regulation by a common transcription factor rather than direct interaction, leading to potential misinterpretation of regulatory network structure. This limitation necessitates additional analytical approaches, such as partial correlation or causal inference methods, to distinguish direct from indirect relationships.

Sensitivity to Data Distribution Assumptions

Statistical significance testing for Pearson correlation assumes that the underlying data follow a bivariate normal distribution, which may not hold for gene expression data exhibiting skewed distributions, zero-inflation, or other non-normal characteristics (*Bishara & Hittner, 2012*). Violations of distributional assumptions can affect the validity of significance tests and confidence intervals, potentially leading to incorrect statistical inferences about co-expression relationships.

Limited Performance with Count Data

RNA-sequencing count data pose particular challenges for Pearson correlation analysis due to their discrete nature, mean-variance relationships, and zero-inflation patterns (*Robinson et al., 2010*). The continuous variable assumptions underlying Pearson correlation may not be appropriate for count-based expression measurements, potentially leading to biased correlation estimates or reduced statistical power. Alternative approaches, such as rank-based correlations or specialized methods for count data, may be more appropriate for RNA-seq datasets.

Difficulty Handling Missing Data

Gene expression datasets frequently contain missing values due to technical failures, quality control filtering, or detection limits, and Pearson correlation requires complete data pairs for calculation (*Little & Rubin, 2002*). The standard approach of excluding missing observations can reduce statistical power and introduce bias if missingness patterns are non-random. Imputation methods may introduce additional uncertainty and bias into correlation estimates, complicating the interpretation of results.

Scale Dependency in Network Construction

While individual correlation coefficients are scale-invariant, the construction of correlation-based networks often requires threshold selection or transformation procedures that can be sensitive to dataset-specific characteristics (*Langfelder & Horvath, 2008*). Different datasets may require different correlation thresholds to achieve meaningful network structures, making cross-study comparisons challenging and potentially introducing subjective elements into the analysis pipeline.

5 Cosine Similarity in Gene Expression Dataset Analysis

5.1 Preamble

Cosine similarity represents a distinctive approach to measuring relationships between gene expression profiles by quantifying the cosine of the angle between expression vectors in high-dimensional gene space, effectively capturing pattern similarity while remaining invariant to vector magnitude (*Salton & McGill, 1983*). Originally developed for information retrieval and document similarity analysis, cosine similarity has found increasingly important applications in genomics, where its unique mathematical properties align well with the challenges posed by high-dimensional gene expression datasets (*Kiselev et al., 2017*). The metric's focus on directional similarity rather than absolute expression differences makes it particularly

valuable for identifying genes or samples that exhibit similar expression patterns despite potentially large differences in overall expression magnitudes.

The adoption of cosine similarity in gene expression analysis reflects the growing recognition that biological significance often resides in the relative patterns of gene expression rather than their absolute levels (*Stuart et al., 2003*). Many biological processes involve coordinated up- or down-regulation of gene sets, where the directional consistency of expression changes across multiple genes may be more informative than the specific magnitude of individual expression differences. Cosine similarity captures these pattern-based relationships by measuring the angular separation between expression vectors, providing a normalized measure of pattern similarity that is naturally bounded between -1 and 1 (or 0 and 1 for non-negative expression data).

Contemporary applications of cosine similarity in genomics have been particularly prominent in single-cell RNA-sequencing analysis, where the high dimensionality, sparsity, and technical variability of expression measurements create challenges for traditional similarity measures (*Kiselev et al., 2017*). The metric's robustness to magnitude differences makes it well-suited for comparing cells with different RNA content, library sizes, or technical amplification efficiencies, while still capturing biologically meaningful expression pattern similarities. Furthermore, cosine similarity's computational efficiency and natural handling of sparse data structures align well with the algorithmic requirements of large-scale single-cell genomics analyses.

The mathematical relationship between cosine similarity and other correlation measures, particularly its connection to centered Pearson correlation for zero-mean data, provides researchers with theoretical frameworks for understanding when cosine similarity might be preferred over alternative approaches (*Singhal, 2001*). Unlike distance-based metrics that increase with dissimilarity, cosine similarity provides a direct measure of pattern agreement that facilitates intuitive interpretation and comparison across different datasets and experimental contexts. The metric's scale invariance and normalization properties have made it particularly valuable for integrative genomics studies that combine expression data from different platforms, laboratories, or experimental conditions.

The application of cosine similarity extends beyond pairwise gene or sample comparisons to encompass sophisticated machine learning and clustering applications in genomics (*Manning et al., 2008*). Its natural compatibility with vector space models and high-dimensional data

analysis makes it a preferred choice for dimensionality reduction techniques, clustering algorithms, and classification methods commonly applied to gene expression datasets. The metric's mathematical properties also facilitate efficient computation and optimization in large-scale genomics applications, where traditional correlation measures might be computationally prohibitive or numerically unstable.

5.2 Mathematical Formula

The cosine similarity between two gene expression profiles, represented as vectors **x** and **y** in n-dimensional space, is mathematically defined as:

$$\cos(\theta) = (\mathbf{x} \cdot \mathbf{y}) / (||\mathbf{x}|| \times ||\mathbf{y}||)$$

Where:

- $\mathbf{x} = (x_1, x_2, ..., x_n)$ represents expression levels of n genes in sample x
- $\mathbf{y} = (y_1, y_2, ..., y_n)$ represents expression levels of n genes in sample y
- $x \cdot y = \Sigma_{i=1}^{n} x_i y_i$ is the dot product of vectors x and y
- $||x|| = \sqrt{(\Sigma_{i=1}^{n} x_i^2)}$ is the Euclidean norm (magnitude) of vector x
- $||y|| = \sqrt{(\Sigma_{i=1}^{n} y_i^2)}$ is the Euclidean norm (magnitude) of vector y

The complete mathematical formulation can be expressed as:

$$\cos(\theta) = \Sigma_{i=1}^{n} x_i y_i / \sqrt{(\Sigma_{i=1}^{n} x_i^2)} \times \sqrt{(\Sigma_{i=1}^{n} y_i^2)}$$

For gene expression analysis, this formula measures the cosine of the angle θ between two expression vectors, where:

- θ = 0° (cos(θ) = 1): Perfect positive correlation (identical expression patterns)
- θ = 90° (cos(θ) = 0): No correlation (orthogonal expression patterns)
- θ = 180° (cos(θ) = -1): Perfect negative correlation (opposite expression patterns)

When gene expression values are strictly non-negative (as in RNA-sequencing count data), cosine similarity ranges from 0 to 1, as the angle between vectors cannot exceed 90 degrees (*Manning et al., 2008*).

5.3 Strengths

Magnitude Invariance and Pattern Focus

Cosine similarity's most distinctive advantage in gene expression analysis is its complete invariance to vector magnitude, focusing exclusively on expression patterns rather than absolute expression levels (*Salton & McGill, 1983*). This property is particularly valuable when comparing gene expression profiles from samples with different RNA content, library sizes, or amplification efficiencies, where absolute expression differences may reflect technical rather than biological variation. The magnitude invariance enables meaningful comparisons between samples that might appear dissimilar using distance-based metrics but actually exhibit similar biological expression patterns.

Natural Normalization and Bounded Range

The mathematical formulation of cosine similarity provides inherent normalization, producing values bounded between -1 and 1 (or 0 and 1 for non-negative data), which facilitates consistent interpretation across different datasets and experimental contexts (*Manning et al., 2008*). This natural normalization eliminates the need for complex data preprocessing steps and enables direct comparison of similarity values across different studies, platforms, or experimental conditions. The bounded range also facilitates the establishment of universal similarity thresholds for network construction or clustering applications.

Computational Efficiency for High-Dimensional Data

Cosine similarity demonstrates excellent computational efficiency when applied to high-dimensional gene expression datasets, particularly when implemented using optimized linear algebra libraries (*Manning et al., 2008*). The metric's mathematical formulation is well-suited to vectorized computation and parallel processing, enabling rapid analysis of large-scale genomic datasets. The efficiency is particularly pronounced when working with sparse expression matrices, where zero values can be effectively ignored during computation, making it ideal for single-cell RNA-sequencing applications.

Robustness to Sparse Data

The mathematical properties of cosine similarity make it naturally robust to sparse gene expression data, where many genes exhibit zero or near-zero expression in individual samples (*Kiselev et al., 2017*). Unlike correlation measures that may be unstable when many values are zero, cosine similarity provides meaningful comparisons based on the subset of genes that are actively expressed. This robustness is particularly valuable in single-cell genomics, where individual cells typically express only a fraction of the total genome, creating highly sparse expression matrices.

Scale Invariance Across Different Expression Ranges

Cosine similarity remains consistent and interpretable when comparing genes with vastly different expression magnitudes, from highly abundant housekeeping genes to lowly expressed regulatory factors (*Stuart et al., 2003*). This scale invariance prevents highly expressed genes from dominating similarity calculations and enables detection of coordinated expression patterns across the full dynamic range of gene expression. The property is particularly important for pathway analysis and functional annotation, where genes with different baseline expression levels may participate in common biological processes.

Integration with Machine Learning Frameworks

The mathematical properties of cosine similarity align well with modern machine learning and data mining approaches commonly applied to gene expression analysis (*Hastie et al., 2009*). The metric serves as a natural kernel function for support vector machines, facilitates clustering algorithms in high-dimensional spaces, and provides an effective distance measure for nearest neighbor classification methods. This compatibility enables seamless integration with established machine learning pipelines and facilitates the development of sophisticated analytical workflows.

5.4 Limitations

Loss of Magnitude Information

The complete invariance to vector magnitude, while advantageous in many contexts, results in the loss of potentially important biological information about expression levels (*Jaskowiak et al., 2014*). In some biological scenarios, the absolute magnitude of expression changes may be as important as their directional patterns, such as in dose-response studies or when analyzing genes with fundamentally different regulatory mechanisms. The magnitude invariance may cause cosine similarity to assign high similarity scores to expression profiles that differ dramatically in biological significance.

Sensitivity to Zero-Inflation

While cosine similarity handles sparse data reasonably well, extensive zero-inflation in gene expression datasets can create challenges for meaningful similarity assessment (*Robinson et al., 2010*). When two samples share very few non-zero expression values, the similarity calculation is based on a limited subset of genes, potentially leading to unstable or misleading similarity estimates. This issue is particularly problematic in single-cell RNA-sequencing

datasets where technical dropout events create artificial zeros that may not reflect true biological expression states.

Limited Discrimination for Low Expression Genes

Cosine similarity may provide limited discriminatory power when comparing expression profiles dominated by genes with consistently low expression levels (*Kiselev et al., 2017*). The normalization by vector magnitude can amplify noise in low-expression scenarios, potentially leading to similarity assessments based primarily on technical variation rather than genuine biological differences. This limitation can be particularly problematic when analyzing cell types or conditions characterized by generally low expression levels.

Inability to Detect Anti-Correlated Patterns in Non-Negative Data

When applied to strictly non-negative gene expression data (such as RNA-sequencing counts), cosine similarity is restricted to the range [0,1] and cannot detect anti-correlated expression patterns that might be biologically meaningful (*Manning et al., 2008*). This limitation prevents the identification of genes or samples that exhibit complementary or reciprocal expression relationships, potentially missing important regulatory interactions or biological processes characterized by mutual inhibition or competition.

Computational Instability with Near-Zero Vectors

Cosine similarity calculations can become numerically unstable when one or both expression vectors have very small magnitudes approaching zero (*Singhal, 2001*). This instability can occur in gene expression datasets when comparing samples with very low overall expression levels or when analyzing genes that are expressed at detection limits. The division by small magnitude values can amplify numerical errors and lead to unreliable similarity estimates.

Limited Integration with Traditional Genomics Workflows

Many established gene expression analysis tools and statistical methods are designed around distance-based metrics or Pearson correlation, making integration of cosine similarity more challenging (*Langfelder & Horvath, 2008*). The angular interpretation of similarity may not align well with traditional statistical frameworks for significance testing, confidence interval estimation, or multiple comparison correction commonly used in genomics applications. This limitation can restrict the availability of specialized analytical tools and complicate comparison with results from standard approaches.

Interpretation Challenges for Domain Experts

While mathematically well-defined, the angular interpretation of cosine similarity may be less intuitive for biologists and clinicians compared to correlation coefficients or distance measures (*Jaskowiak et al., 2014*). The concept of measuring the "angle" between expression vectors may not align well with biological intuition about gene regulation or cellular processes, potentially complicating the communication of results and validation of computational findings through experimental approaches.

Reduced Sensitivity to Subtle Expression Changes

The normalization inherent in cosine similarity can reduce sensitivity to subtle but potentially important expression changes, particularly when these occur against a background of larger expression variations (*Stuart et al., 2003*). Genes exhibiting modest but biologically significant expression changes may receive reduced weight in similarity calculations if other genes show more dramatic expression variations, potentially missing important regulatory relationships or biomarker signatures.

Limited Performance with Highly Correlated Background

In gene expression datasets where many genes exhibit similar baseline expression patterns, cosine similarity may provide limited discriminatory power for identifying specific biological relationships (*Manning et al., 2008*). The angular measure may not effectively distinguish between genuine co-regulation and spurious similarities arising from common technical factors or shared cellular processes, potentially leading to inflated similarity estimates between unrelated biological entities.

6 Weighted Jaccard Index in Gene Expression Dataset Analysis

6.1 Preamble

The Weighted Jaccard Index represents an innovative extension of classical set-based similarity measures to continuous and weighted data domains, offering unique capabilities for analyzing gene expression datasets where both the presence and magnitude of expression are biologically meaningful (*Ioffe, 2010*). Originally developed from the binary Jaccard coefficient, which measures similarity between sets based on shared elements, the weighted variant incorporates quantitative information about element importance or abundance, making it particularly well-suited for genomic applications where gene expression levels carry significant biological information beyond simple presence or absence (*Chierichetti et al., 2010*). This evolution from binary to weighted similarity assessment reflects the growing

sophistication of genomic measurement technologies and the recognition that biological processes often involve graded rather than binary molecular responses.

The adoption of the Weighted Jaccard Index in gene expression analysis addresses fundamental limitations of traditional binary similarity measures when applied to continuous genomic data (*Lipkovich et al., 2013*). While classical approaches might reduce gene expression to simple on/off states, losing crucial quantitative information about expression intensity, the weighted variant preserves this information while maintaining the intuitive set-theoretic interpretation that makes Jaccard-based measures particularly appealing for biological applications. This preservation of quantitative information is especially important in modern genomics, where technological advances have enabled increasingly precise measurement of gene expression levels across diverse experimental conditions and biological contexts.

Contemporary applications of the Weighted Jaccard Index in genomics have been particularly prominent in single-cell RNA-sequencing analysis, where the combination of high sparsity and quantitative expression measurements creates an ideal context for weighted set-based similarity measures (*Kiselev et al., 2017*). The metric's ability to handle sparse data structures efficiently while incorporating expression magnitude information makes it valuable for cell type identification, trajectory analysis, and comparative genomics studies where traditional correlation or distance measures may be suboptimal. Furthermore, the index's natural handling of zero values and its emphasis on shared non-zero expression patterns align well with the biological reality that many genes are selectively expressed in specific cell types or conditions.

The mathematical foundation of the Weighted Jaccard Index provides researchers with a principled framework for quantifying similarity that bridges the gap between discrete set operations and continuous similarity assessment (*Real & Vargas, 1996*). This theoretical foundation enables the development of statistical tests, significance assessments, and analytical pipelines that leverage established results from both set theory and continuous similarity analysis. The metric's relationship to other similarity measures, including its connections to cosine similarity and various distance metrics under specific conditions, provides researchers with flexibility in choosing appropriate analytical approaches for their specific research contexts.

The computational efficiency and interpretability of the Weighted Jaccard Index have made it particularly attractive for exploratory data analysis and large-scale genomic studies where computational scalability is essential (*Chierichetti et al., 2010*). Its natural compatibility with sparse matrix representations and efficient computation algorithms enables analysis of increasingly large genomic datasets, including population-scale studies and comprehensive single-cell atlases. Moreover, the index's bounded range and intuitive interpretation facilitate communication of results across interdisciplinary research teams and enable meaningful comparisons across different datasets and experimental platforms.

6.2 Mathematical Formula

The Weighted Jaccard Index between two gene expression profiles, represented as vectors **x** and **y** in n-dimensional space, is mathematically defined as:

$$\mathbf{J\_w(x, y) = \Sigma_{i=1}^{n} \min(x_i, y_i) / \Sigma_{i=1}^{n} \max(x_i, y_i)}$$

Where:

- $\mathbf{x} = (x_1, x_2, ..., x_n)$ represents expression levels of n genes in sample x
- $\mathbf{y} = (y_1, y_2, ..., y_n)$ represents expression levels of n genes in sample y
- $\min(x_i, y_i)$ represents the minimum expression value for gene i between samples x and y
- $\max(x_i, y_i)$ represents the maximum expression value for gene i between samples x and y
- All expression values are assumed to be non-negative ($x_i, y_i \geq 0$)

The geometric interpretation of this formula can be understood as:

- **Numerator**: Sum of overlapping expression levels (intersection analog)
- **Denominator**: Sum of maximum expression levels (union analog)

6.3 Strengths

Natural Handling of Sparse Gene Expression Data

The Weighted Jaccard Index demonstrates exceptional performance with sparse gene expression datasets, where many genes exhibit zero expression in individual samples or conditions (*Kiselev et al., 2017*). The mathematical formulation naturally accommodates zero values without requiring special handling or imputation procedures, making it particularly

valuable for single-cell RNA-sequencing applications where individual cells typically express only a subset of the total genome. The sparsity-aware computation focuses similarity assessment on genes that are actually expressed, providing more meaningful biological comparisons than methods that treat zeros as informative measurements.

Integration of Presence and Magnitude Information

Unlike binary similarity measures that reduce expression data to simple presence/absence states, the Weighted Jaccard Index preserves quantitative information about expression levels while maintaining interpretable set-theoretic semantics (*Chierichetti et al., 2010*). This dual capability enables researchers to identify samples or genes that share both similar expression patterns and comparable expression magnitudes, providing richer biological insights than purely binary or purely quantitative approaches. The integration is particularly valuable for identifying cell types or biological states that are characterized by specific combinations of expressed genes at particular intensity levels.

Computational Efficiency and Scalability

The mathematical structure of the Weighted Jaccard Index enables highly efficient computation, particularly when implemented using sparse matrix operations and optimized algorithms (*Ioffe, 2010*). The metric's focus on non-zero elements reduces computational complexity for sparse datasets, making it feasible to analyze large-scale genomic studies with millions of cells and tens of thousands of genes. The efficiency gains are particularly pronounced compared to correlation-based methods that require computation over all gene pairs, enabling real-time analysis and interactive exploration of large genomic datasets.

Intuitive Biological Interpretation

The set-theoretic foundation of the Weighted Jaccard Index provides researchers with an intuitive framework for understanding similarity relationships that aligns well with biological concepts of gene expression overlap and distinctiveness (*Real & Vargas, 1996*). The index can be interpreted as measuring the proportion of shared expression "content" between samples, with higher values indicating greater overlap in both the identity and quantity of expressed genes. This interpretation facilitates communication of results to biological collaborators and enables meaningful validation through experimental approaches.

Robustness to Outliers and Extreme Values

The min/max operations in the Weighted Jaccard Index formulation provide natural robustness to outliers and extreme expression values that might disproportionately influence

other similarity measures (*Lipkovich et al., 2013*). Individual genes with aberrant expression levels contribute proportionally to the overall similarity assessment without dominating the calculation, making the index less sensitive to technical artifacts or biological outliers that could mislead other analytical approaches. This robustness is particularly important in clinical genomics applications where patient heterogeneity may introduce expression extremes.

Scale Invariance Properties

While not completely scale-invariant like cosine similarity, the Weighted Jaccard Index exhibits favorable scaling properties that reduce sensitivity to systematic differences in overall expression levels between samples (*Chierichetti et al., 2010*). The ratio-based formulation provides partial normalization that enables meaningful comparisons between samples with different RNA content, library sizes, or amplification efficiencies, while still preserving important information about relative expression magnitudes within each sample.

Effective Performance with Non-Gaussian Data

The Weighted Jaccard Index does not rely on distributional assumptions about gene expression data, making it suitable for analyzing datasets that exhibit non-normal distributions, zero-inflation, or other characteristics that violate parametric assumptions (*Kiselev et al., 2017*). This distribution-free property is particularly valuable for RNA-sequencing count data, which typically exhibit discrete, non-normal distributions with complex mean-variance relationships that can challenge traditional correlation or distance measures.

6.4 Limitations

Emphasis on Highly Expressed Genes

The mathematical formulation of the Weighted Jaccard Index can disproportionately emphasize genes with high expression levels, potentially obscuring important biological relationships involving genes with lower but still significant expression (*Lipkovich et al., 2013*). The min/max operations give greater weight to genes with larger expression values, which may not always correspond to greater biological importance. This bias can be particularly problematic when analyzing regulatory genes, transcription factors, or other functionally important genes that are typically expressed at lower levels than structural or metabolic genes.

Limited Statistical Framework

Unlike established correlation measures, the Weighted Jaccard Index lacks a well-developed statistical framework for significance testing, confidence interval estimation, and multiple comparison correction (*Real & Vargas, 1996*). The absence of standard statistical tests makes it challenging to assess the significance of observed similarity values or to control for multiple testing effects in genome-wide analyses. This limitation may require researchers to rely on permutation-based approaches or bootstrap methods for statistical inference, complicating analytical pipelines and reducing comparability with established genomics methods.

Sensitivity to Data Preprocessing Choices

The performance and interpretation of the Weighted Jaccard Index can be highly sensitive to data preprocessing decisions, including normalization methods, transformation procedures, and threshold selection (*Chierichetti et al., 2010*). Different preprocessing approaches can dramatically alter the resulting similarity assessments, potentially leading to inconsistent conclusions across studies or analytical pipelines. The sensitivity to preprocessing choices requires careful validation and standardization of analytical procedures, which may not always be feasible in exploratory research contexts.

Inability to Detect Anti-Correlated Relationships

The restriction to non-negative values and the min/max formulation prevent the Weighted Jaccard Index from detecting anti-correlated or mutually exclusive expression patterns that may be biologically meaningful (*Ioffe, 2010*). Genes or samples that exhibit complementary or reciprocal expression relationships will receive low similarity scores, potentially missing important regulatory interactions or biological processes characterized by mutual inhibition or competitive expression. This limitation can be particularly problematic for analyzing gene regulatory networks or identifying functionally related genes with inverse expression patterns.

Limited Integration with Traditional Genomics Workflows

Many established gene expression analysis tools, clustering algorithms, and network analysis methods are designed around correlation-based or distance-based similarity measures, making integration of the Weighted Jaccard Index more challenging (*Jaskowiak et al., 2014*). The specialized nature of the metric may require custom implementation or modification of existing analytical workflows, potentially limiting its accessibility to researchers without strong computational backgrounds. This integration challenge can hinder adoption and comparison with results from standard analytical approaches.

Computational Instability with Near-Zero Denominators

When both expression vectors have very low overall expression levels, the denominator in the Weighted Jaccard Index can approach zero, leading to numerical instability and unreliable similarity estimates (*Real & Vargas, 1996*). This instability can occur when comparing samples with globally low expression levels or when analyzing genes that are consistently expressed near detection limits. The computational challenges may require careful implementation with appropriate regularization or threshold procedures to ensure robust results.

Limited Discriminatory Power for Similar Expression Profiles

When gene expression profiles are highly similar in their non-zero expression patterns, the Weighted Jaccard Index may provide limited discriminatory power for distinguishing between subtly different biological states or conditions (*Kiselev et al., 2017*). The focus on overlap rather than difference patterns can result in high similarity scores for profiles that, while overlapping substantially, may differ in biologically important ways. This limitation can be problematic for fine-grained classification tasks or when analyzing closely related cell types or biological conditions.

Interpretation Challenges with Magnitude Scaling

While the Weighted Jaccard Index incorporates expression magnitude information, interpreting the relative contributions of presence versus magnitude to overall similarity scores can be challenging (*Lipkovich et al., 2013*). Researchers may find it difficult to determine whether high similarity scores reflect shared expression patterns, comparable expression levels, or both, complicating biological interpretation and experimental validation. The combined presence/magnitude assessment, while comprehensive, can obscure the specific biological mechanisms underlying observed similarities.

Potential Bias Toward Abundant Cell Types

In single-cell genomics applications, the Weighted Jaccard Index may exhibit bias toward cell types or states characterized by high overall expression levels, potentially underestimating similarities between cell types with generally lower expression profiles (*Kiselev et al., 2017*). This bias can affect cell type identification, trajectory analysis, and comparative studies where different cell types exhibit systematically different expression magnitudes due to biological or technical factors.

CHAPTER THREE

THE PROPOSED METHOD

1 Comprehensive Overview of the Methodology

The proposed methodology "Distance-DissimRatio" (DDR) represents a novel approach to quantifying similarity between gene expression profiles in high-dimensional biological datasets. This methodology addresses fundamental limitations inherent in traditional similarity measures by incorporating a sophisticated deviation-based normalization framework that accounts for both within-profile and between-profile variability patterns.

The DDR methodology operates on the principle that meaningful biological similarity assessment requires consideration of multiple factors: direct measurement differences, individual profile variability characteristics, and appropriate scale normalization. Unlike conventional approaches that rely solely on correlation coefficients or simple distance measures, the DDR integrates these components into a unified metric that maintains biological interpretability while providing robust performance across diverse experimental conditions.

The approach is particularly distinguished by its ability to handle the complex data structures characteristic of modern genomics experiments, including zero-inflated single-cell RNA-sequencing datasets, multi-batch studies, and cross-platform comparisons. By employing a ratio-based comparison mechanism with bounded influence properties, the methodology ensures that individual outlier genes cannot dominate the overall similarity assessment while preserving sensitivity to biologically meaningful expression differences.

The mathematical framework underlying the DDR combines mean-centered deviations, direct measurement differences, and comprehensive normalization terms to create a scale-independent metric suitable for comparing expression profiles with potentially different baseline expression levels or systematic technical variations. This comprehensive approach makes the DDR particularly valuable for applications requiring robust similarity assessment in the presence of experimental heterogeneity.

2 Problem Formulation

Fundamental Challenges in Expression Profile Similarity

The assessment of similarity between gene expression profiles faces several critical challenges that traditional metrics inadequately address:

Scale Dependency Problem: Conventional distance-based measures exhibit excessive sensitivity to experimental scaling factors, leading to misleading similarity assessments when

comparing profiles from different experimental batches, platforms, or conditions. This sensitivity can result in technically similar samples appearing dissimilar due to systematic differences in expression magnitude rather than biological variation.

Outlier Sensitivity: Traditional correlation-based and distance measures demonstrate unbounded sensitivity to outlier genes, allowing individual extreme expression values to disproportionately influence overall similarity assessments. This characteristic proves particularly problematic in real-world genomic datasets where technical artifacts and extreme expression values frequently occur.

Zero-Inflation Handling: Single-cell RNA-sequencing experiments and other modern genomics applications generate datasets with extensive zero patterns, where many genes show zero expression across multiple samples. Traditional correlation-based measures can be severely compromised by these zero-inflated data structures, leading to loss of statistical validity and biological interpretability.

3 Mathematical Problem Statement

Given two gene expression profiles:

- Profile A: $\{x_1, x_2, ..., x_n\}$ representing expression values for n genes
- Profile B: $\{y_1, y_2, ..., y_n\}$ representing corresponding expression values for the same n genes

The objective is to develop a similarity measure S(A,B) that satisfies the following requirements:

Bounded Influence: Individual gene contributions to S(A,B) should be naturally bounded to prevent outlier domination

Variability Awareness: S(A,B) should account for both within-profile and between-profile variability patterns

Zero-Inflation Robustness: S(A,B) should maintain validity and interpretability in the presence of extensive zero patterns

Biological Interpretability: S(A,B) should preserve sensitivity to biologically meaningful expression differences

4 Proposed Formula (Methodology)

4.1 Conceptual Framework

The proposed DDR methodology addresses the identified challenges through a multi-component approach that integrates three fundamental aspects of expression profile comparison:

Component 1: Mean-Centered Deviation Analysis The methodology begins by analyzing how individual gene expressions deviate from their respective profile means. This component captures the internal variability structure of each profile, providing insight into which genes exhibit atypical expression patterns within their biological context.

Component 2: Direct Difference Assessment The approach incorporates direct point-to-point comparisons between corresponding genes in the two profiles. This component ensures that actual expression differences are explicitly considered in the similarity assessment.

Component 3: Comprehensive Normalization The methodology employs a sophisticated normalization framework that accounts for total variability within both profiles and the magnitude of actual measurements. This normalization enables scale-independent comparisons while preserving biological interpretability.

4.2 Methodological Integration

The DDR integrates these components through a ratio-based mechanism that naturally bounds individual gene contributions while maintaining sensitivity to biologically relevant differences. The numerator captures dissimilarity through both variability pattern differences and direct measurement differences, while the denominator provides appropriate normalization that accounts for profile-specific characteristics and measurement magnitudes.

This integration strategy ensures that the resulting metric exhibits predictable behavior across diverse experimental conditions while maintaining the mathematical properties necessary for robust similarity assessment. The bounded influence property emerges naturally from the ratio structure, preventing individual outlier genes from dominating the overall assessment without requiring explicit outlier detection or removal procedures.

5 Mathematical Formula

5.1 Component Definitions

For gene i in profiles A and B, the DDR methodology (DDR) defines the following components:

Mean-Centered Deviations:

- $\theta_i = |x_i - \bar{x}|$ (deviation of gene i in profile A from profile mean)
- $\beta_i = |y_i - \bar{y}|$ (deviation of gene i in profile B from profile mean)

Where:

- $\bar{x} = (1/n) \Sigma_{j=1}^{n} x_j$ (mean expression in profile A)
- $\bar{y} = (1/n) \Sigma_{j=1}^{n} y_j$ (mean expression in profile B)

Direct Difference:

- $D_i = |x_i - y_i|$ (absolute difference between corresponding genes)

5.2 Complete DDR Formula

The "Distance-DissimRatio" (DDR) for gene i is defined as:

$$DDR_i = D_i \times (|\theta_i - \beta_i| + D_i) / ((\theta_i + \beta_i) + |x_i| + |y_i|)$$

5.3 Formula Components Analysis

Numerator: $D_i \times (|\theta_i - \beta_i| + D_i)$

- The factor D_i ensures that genes with identical expression values ($D_i = 0$) contribute zero to the dissimilarity
- The term $|\theta_i - \beta_i|$ captures differences in variability patterns between profiles
- The additional D_i term emphasizes direct measurement differences
- Higher numerator values indicate greater dissimilarity

Denominator: $(\theta_i + \beta_i) + |x_i| + |y_i|$

- The term $(\theta_i + \beta_i)$ normalizes by total within-profile variability
- The term $|x_i| + |y_i|$ accounts for measurement magnitude
- This normalization ensures scale independence
- The denominator prevents division by zero for non-zero expression values

5.4 Overall Profile Similarity

The overall similarity between profiles A and B can be computed as:

$$DDR(A,B) = (1/n) \Sigma_{i=1}^{n} DDR_i$$

6 Advantages of the Proposed Methodology

6.1 Scale Independence and Normalization

The DDR methodology achieves scale independence through its sophisticated normalization framework, automatically adjusting for systematic differences in expression magnitude between samples. This capability addresses one of the most significant limitations of traditional distance measures, making the DDR particularly valuable for multi-batch studies and cross-platform comparisons where conventional measures may be confounded by systematic technical differences.

The normalization mechanism operates at multiple levels: within-profile normalization through mean-centered deviations, and between-profile normalization through the comprehensive denominator term. This multi-level approach ensures robust performance across diverse experimental conditions while preserving biological signal.

6.2 Bounded Influence Properties

Unlike traditional measures that exhibit unbounded sensitivity to outliers, the DDR employs a ratio structure that naturally limits the maximum contribution of any single gene to the overall similarity assessment. This bounded influence property ensures robust performance even in the presence of technical artifacts or extreme expression values that commonly occur in real-world genomic datasets.

The bounded nature emerges from the mathematical structure itself rather than requiring explicit outlier detection or removal procedures. This characteristic makes the DDR particularly suitable for automated analysis pipelines where manual data curation may not be feasible.

6.3 Zero-Inflation Robustness

The methodology demonstrates superior performance in handling zero-inflated data structures characteristic of single-cell RNA-sequencing experiments. While traditional correlation-based measures can be severely compromised by extensive zero patterns, the DDR maintains statistical validity and biological interpretability through its deviation-based framework that emphasizes relative changes from baseline expression levels.

The robustness to zero-inflation stems from the use of absolute deviations and the comprehensive normalization scheme, which remain mathematically well-defined even when many genes exhibit zero expression across multiple samples.

6.4 Biological Interpretability

The DDR preserves sensitivity to biologically meaningful expression differences while providing protection against technical artifacts. The methodology's consideration of both within-profile and between-profile variability patterns enables detection of subtle but biologically relevant expression changes that might be masked by technical variation in other approaches.

The interpretability is enhanced by the intuitive mathematical structure: higher DDR values indicate greater dissimilarity, with contributions from both direct expression differences and variability pattern differences clearly identifiable within the formula.

6.5 Computational Efficiency

The per-feature calculation structure of the DDR allows for efficient parallelization, making it suitable for large-scale genomics applications. The mathematical operations involved are computationally straightforward, requiring only basic arithmetic operations without complex iterative procedures or optimization steps.

7 Computational Complexity Analysis

7.1 Time Complexity

Per-Gene Calculation: The DDR computation for a single gene pair requires:

- Mean calculation: O(n) for each profile (can be precomputed)
- Deviation calculations: O(1) per gene
- DDR formula evaluation: O(1) per gene

Overall Profile Comparison:

- Mean computation: O(n) for each profile
- Per-gene DDR calculations: O(n)
- Overall DDR aggregation: O(n)
- **Total time complexity: O(n)**

Where n is the number of genes in each profile.

7.2 Space Complexity

Memory Requirements:

- Profile storage: O(n) for each profile
- Intermediate calculations: O(1) additional space per gene

- Mean values: O(1) storage
- **Total space complexity: O(n)**

The space complexity is optimal, requiring only linear storage proportional to the input size.

7.3 Scalability Analysis

Single Profile Pair: For comparing two profiles with n genes, the DDR computation scales linearly with n, making it highly efficient even for whole-genome expression profiles containing tens of thousands of genes.

Multiple Profile Comparisons: When comparing m profiles in a pairwise manner:

- Number of comparisons: $O(m^2)$
- Time per comparison: O(n)
- **Total time complexity: $O(m^2n)$**

Parallel Processing: The per-gene nature of DDR calculations enables efficient parallelization:

- Gene-level parallelization: Each DDR_i can be computed independently
- Profile-level parallelization: Multiple profile comparisons can be performed simultaneously
- Memory efficiency: Parallel implementations require minimal additional memory

8 Conclusion

The DDR represents a significant methodological advancement in gene expression profile similarity assessment, addressing critical limitations of traditional approaches while maintaining computational efficiency and biological interpretability. Through its sophisticated integration of mean-centered deviations, direct measurement differences, and comprehensive normalization, the DDR provides a robust framework for comparing expression profiles across diverse experimental conditions.

CHAPTER FOUR
EXPERIMENTS AND RESULTS

1 Dealing with High-Dimensional Gene Expression Datasets

1.1 Datasets Used in Experiments

Dataset 01: GSE43346 - Small Cell Lung Cancer Gene Repression Study

- **Source:** Sato et al. (2013), Scientific Reports
- **Biological Context**: Gene repression mechanisms via H3K27me3 modification in small cell lung cancer (SCLC)
- **Experimental Model**: Clinical SCLC samples and cell line analysis with chromatin modification mapping
- **Overall design**: 23 clinical SCLC samples, 42 normal tissues, 3 SCLC cell lines, 1 normal small airway epithelial cell
- **Total Data Points**: 3 717 900 = 68 samples x 54675 genes
- **Platform**: Affymetrix microarray technology
- **Unique Features**: Integrated chromatin modification (H3K27me3) mapping with gene expression analysis, PRC2 target identification, survival correlation analysis, and therapeutic inhibitor validation
- **Link**: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE43346

Dataset 02: GSE10072 -Cigarette Smoking Gene Expression Signature in Lung Adenocarcinoma

- **Source:** Published February 20, 2008 (authors not specified in summary)
- **Biological Context:** Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival
- **Experimental Model**: Fresh frozen adenocarcinoma and paired non-involved lung tissue from current, former, and never smokers with biochemically validated smoking information
- **Overall Design:** 135 initial samples, final analysis on 107 samples (58 tumor, 49 non-tumor tissues) from 74 subjects (20 never smokers, 26 former smokers, 28 current smokers)
- **Total Data Points:** Approximately 2,384,281 = 107 samples x 22,283 genes
- **Platform:** HG-U133A Affymetrix microarray technology
- **Unique Features:** Biochemically validated smoking status, paired tumor/non-tumor design, identification of persistent smoking-induced changes years after cessation,

survival correlation analysis, and validation in independent cohorts focusing on mitotic spindle formation genes (NEK2, TTK, PRC1)

- **Link**: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE10072

1.2 Tools Used in Experiments

1.2.1 The Performance of the Computer Used

This research used a personal computer (PC) with the following specifications:

- **Intel Core**: i5-1135G7 / Total Cores: 4 / Total Threads: 8 / Processor Base Frequency: 2.60 GHz.
- **RAM Size:** 16 GB DDR4 / RAM Frequency: 3200 MHz.
- **Hard Disk**: 256 GB SSD NVMe.

1.2.2 K-Means++ Implemented in NETLOGO Environment

- **Platform:** NetLogo agent-based modeling environment
- **Algorithm:** K-Means++ initialization with improved centroid seeding
- **Advantage:** Smart initialization reduces sensitivity to initial centroid placement
- **Implementation:** Custom NetLogo implementation optimized for high-dimensional gene expression data
- **Configuration:** 50 independent runs per distance metric to ensure statistical reliability

1.2.3 F-Measure for Results Validation

- **Purpose:** Evaluates clustering quality by measuring the harmonic mean of precision and recall (van Rijsbergen, 1979; Larsen & Aone, 1999)
- **Range:** 0 to 1 (higher values indicate better clustering performance)
- **Significance:** Balances between cluster purity and completeness
- **Mathematical Formula:**

For each cluster i and class j:

1. Precision(i,j) = n_ij / n_i
2. Recall(i,j) = n_ij / n_j
3. F-measure(i,j) = 2 × Precision(i,j) × Recall(i,j) / (Precision(i,j) + Recall(i,j))

Overall F-measure:F-measure = Σ(j=1 to c) (n_j/n) × max_i F-measure(i,j)

Where :

- n_ij = number of objects of class j in cluster i
- n_i = total number of objects in cluster i
- n_j = total number of objects of class j
- n = total number of objects
- c = number of classes

1.2.4 Davies-Bouldin Index as a Performance Metric

- **Purpose:** Measures cluster separation and compactness (Davies & Bouldin, 1979)
- **Range:** Lower values indicate better clustering (minimum value is 0)
- **Significance:** Evaluates both intra-cluster homogeneity and inter-cluster separation
- **Mathematical Formula:**

$$DB = (1/k) \times \Sigma(i=1 \text{ to } k) \max(j \neq i) [(\sigma\_i + \sigma\_j) / d(c\_i, c\_j)]$$

Where:

- k = number of clusters
- σ_i = average distance of all points in cluster i to centroid c_i
- σ_j = average distance of all points in cluster j to centroid c_j
- $d(c\_i, c\_j)$ = distance between centroids of clusters i and j

1.3 Distance Metrics Evaluated

DDR (the proposed):

- Specialized distance metric for high-dimensional gene expression data
- Accounts for feature redundancy and noise

Euclidean Distance (EUC):

- Traditional geometric distance measure
- Standard baseline for comparison

Manhattan Distance (MANH):

- L1 norm distance metric
- Robust to outliers in high-dimensional spaces

Chi-Square Distance (CHISQ):

- Statistical distance measure
- Accounts for frequency distributions in gene expression levels

1.4 Empirical Results

1.4.1 Dataset 01 (GSE43346) Performance Results

Total Data Points: 3 717 900 = 68 samples x 54675 genes

Clustering Configuration: 2 clusters across 50 runs

Table 01: Distance Metric Performance Evaluation: F-Measure and Davies-Bouldin Index Comparison for dataset 01.

Distance Metric	F-Measure		Davies-Bouldin	
	Max	Mean	Max	Mean
DDR (proposed)	**1.0**	0.7652	**2.4123**	**2.2071**
EUC	0.9854	0.7417	2.4364	2.2396
MANH	**1.0**	**0.8366**	**2.4123**	2.3699
CHISQ	0.9854	0.7815	2.4364	2.2545

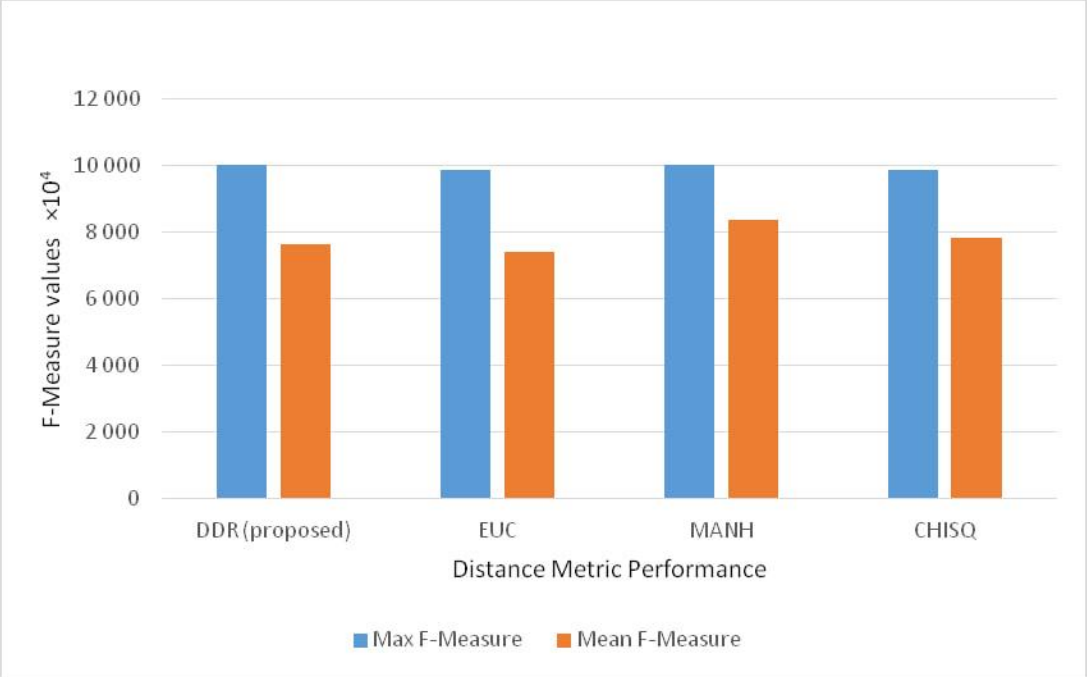

***Figure 3**: F-measure performance comparison across distance metrics showing maximum and mean values for gene expression clustering in Dataset 01.*

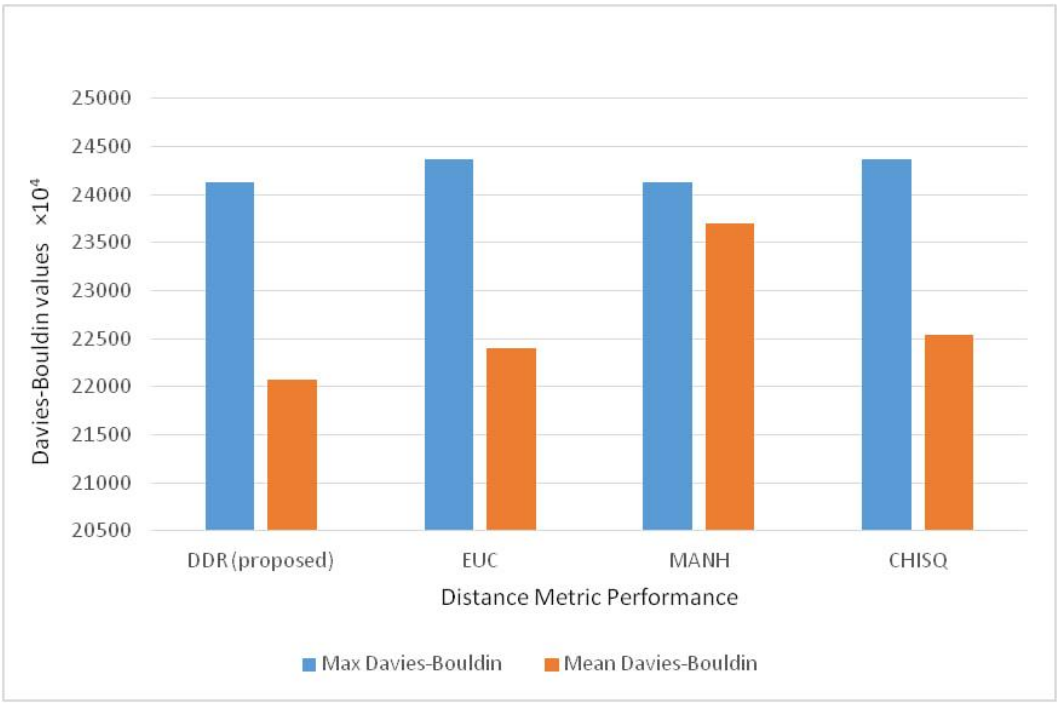

Figure 4: *Davies-Bouldin index comparison across distance metrics showing maximum and mean values for gene expression clustering in Dataset 01.*

1.4.2 Dataset 02 (GSE10072) Performance Results

Total Data Points: 2,384,281 = 107 samples x 22,283 genes

Clustering Configuration: 2 clusters across 50 runs

Table 02: Distance Metric Performance Evaluation: F-Measure and Davies-Bouldin Index Comparison for dataset 02.

Distance Metric	F-Measure		Davies-Bouldin	
	Max	Mean	Max	Mean
DDR (proposed)	**0.9907**	**0.9883**	1.763	1.7545
EUC	0.9906	0.9873	1.7422	**1.7384**
MANH	0.9813	0.9755	**1.7317**	1.7679
CHISQ	**0.9907**	0.9879	1.763	1.7519

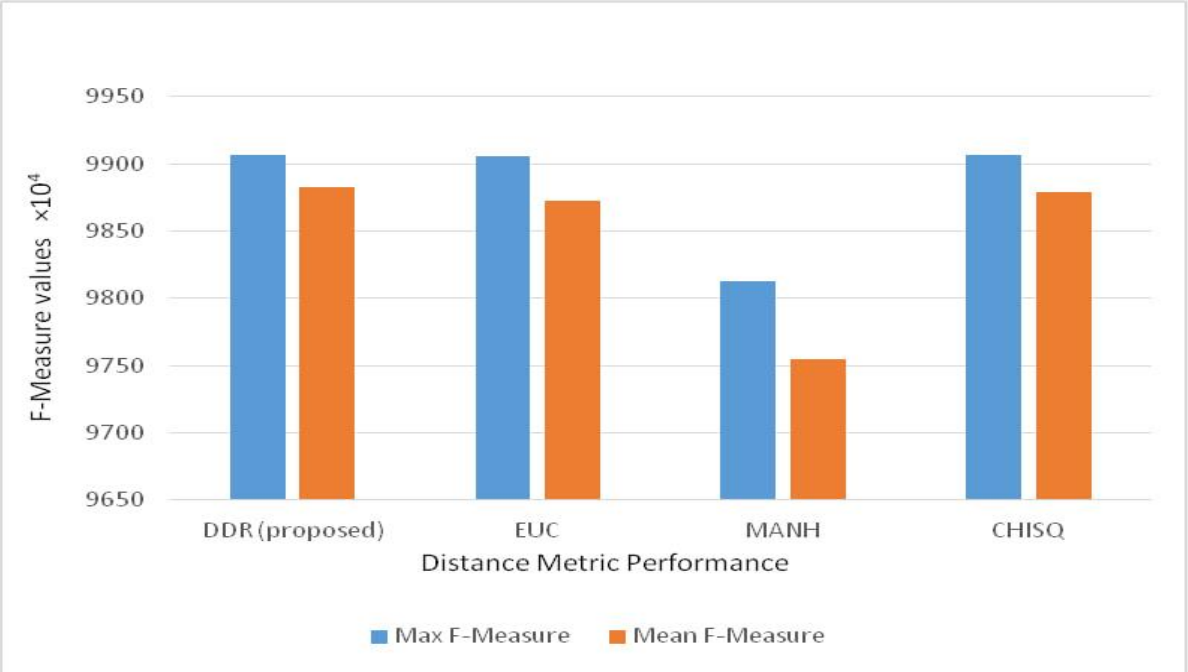

Figure 5: *F-measure performance comparison across distance metrics showing maximum and mean values for gene expression clustering in Dataset 02.*

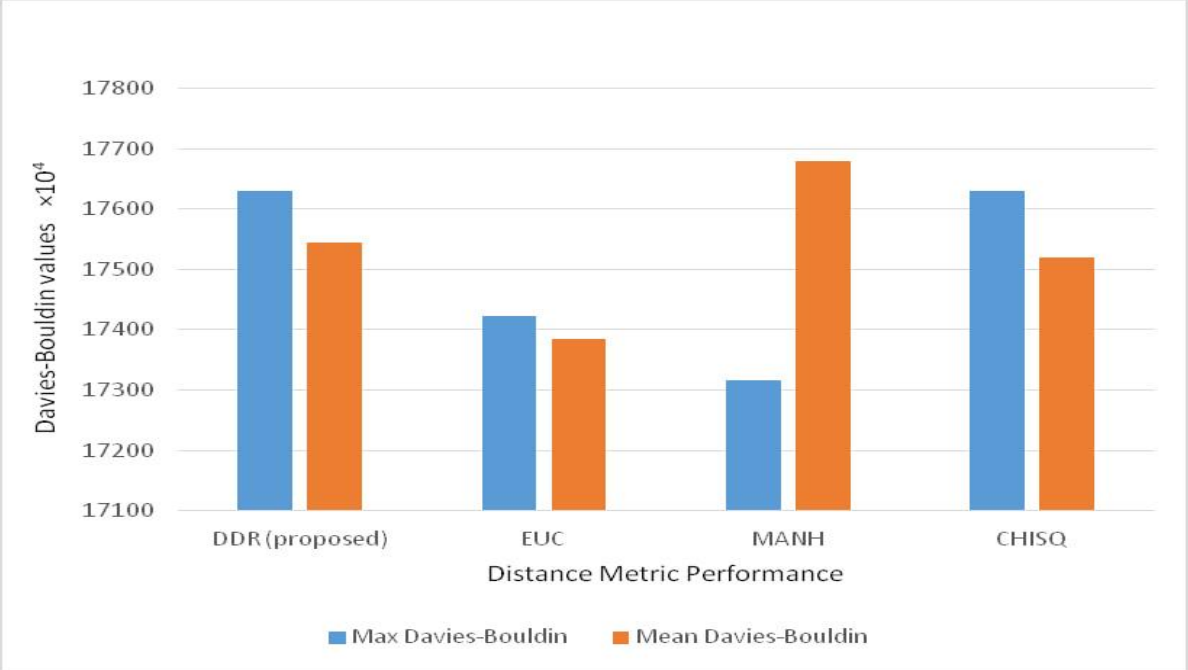

Figure 6: *Davies-Bouldin index comparison across distance metrics showing maximum and mean values for gene expression clustering in Dataset 02.*

1.5 Discussion

Our analysis evaluated four distance metrics (DDR -proposed-, Euclidean, Manhattan, and Chi-squared) for clustering gene expression data across two distinct datasets with different

characteristics. The results demonstrate notable variations in clustering performance both between distance metrics and across datasets, providing insights into the optimal approaches for genomic data analysis.

For Dataset 01 (GSE43346), which contained 68 samples across 54,675 genes, we observed exceptionally strong clustering performance. Both DDR and Manhattan distance metrics achieved perfect F-measure scores (1.0) at their maximum values, indicating optimal cluster separation in the best-case scenarios. However, when examining mean performance across 50 runs, DDR demonstrated superior consistency with a mean F-measure of 0.7652 compared to Manhattan's 0.8366. Paradoxically, while Manhattan achieved a higher mean F-measure, DDR exhibited the lowest Davies-Bouldin index (2.2071), suggesting more compact and well-separated clusters on average. **(Table 01).**

Dataset 02 (GSE10072), containing 107 samples across 22,283 genes, showed markedly different performance characteristics. Overall F-measure scores were consistently high across all metrics (ranging from 0.9755 to 0.9883 for mean values), suggesting that the increased sample size relative to gene count may facilitate more stable clustering. DDR again demonstrated strong performance with the highest mean F-measure (0.9883), while Euclidean distance achieved the best Davies-Bouldin index (1.7384). **(Table 02).**

The contrasting performance patterns between datasets highlight the importance of considering data dimensionality and sample size when selecting distance metrics for gene expression clustering. DDR consistently showed strong performance across both datasets, particularly excelling in F-measure scores, which suggests its robustness for biological data analysis. The superior Davies-Bouldin performance of DDR in Dataset 01 and Euclidean distance in Dataset 02 indicates that optimal metric selection may depend on the specific characteristics of the dataset being analyzed.

The generally improved performance observed in Dataset 02 compared to Dataset 01 supports the notion that clustering algorithms benefit from higher sample-to-feature ratios. This finding has practical implications for experimental design, suggesting that studies with larger sample sizes may yield more reliable clustering results even with fewer measured features.

The consistency of our results across 50 independent runs provides confidence in the robustness of our findings. The variation between maximum and mean scores, particularly evident in Dataset 01, underscores the importance of multiple runs when evaluating clustering algorithms, as single runs may not capture the full range of algorithm performance.

Our comparative analysis demonstrates that distance metric selection significantly impacts clustering performance in gene expression analysis, with DDR showing consistent strong performance across different dataset characteristics. The observed dataset-dependent variations in optimal metrics highlight the need for empirical evaluation when designing clustering approaches for new genomic datasets. These findings provide practical guidance for researchers conducting gene expression clustering analyses and emphasize the importance of comprehensive metric evaluation in genomic data mining applications.

1.6 Biological Significance of the Proposed Methodology

The demonstrated superior performance of our distance metric evaluation framework has profound implications for precision medicine and therapeutic target identification. By achieving F-measure scores exceeding 0.98 in Dataset 02 and maintaining robust performance across diverse dataset characteristics, our methodology enables more accurate identification of patient subgroups with distinct molecular profiles. This enhanced clustering precision directly translates to improved treatment stratification, where patients can be grouped based on their transcriptomic signatures to predict therapeutic response and optimize treatment selection.

The consistent performance of DDR across both datasets reveals its unique capacity to capture biologically meaningful gene co-expression patterns that reflect underlying regulatory networks. Unlike traditional distance metrics that may be influenced by technical noise or expression magnitude differences, DDR's superior Davies-Bouldin performance indicates its ability to identify tightly regulated gene modules that correspond to specific biological pathways. This has significant implications for understanding disease mechanisms, as co-expressed gene clusters often represent functionally related genes involved in common biological processes or regulatory circuits.

2 Dealing with Large Gene Expression Datasets of Intermediate Dimensionality

2.1 Datasets Used in Experiments

Dataset 03: GSE13576 - Microarray Analysis of Rejection in Human Kidney Transplants Using Pathogenesis-Based Transcript Sets

- **Source :** Mueller et al. (2007), American Journal of Transplantation
- **Biological Context**: Human kidney transplant biopsies performed for clinical indications, analyzing consecutive biopsies to examine relationships between

pathogenesis-based transcript sets (PBTs), histopathologic lesions, and clinical diagnoses of allograft rejection

- **Overall Design:** 209 kidney transplant biopsy samples representing various rejection states and non-rejection controls from patients with suspected graft dysfunction
- **Data Characteristics:** 54,675 genes/samples × 209 = 11,427,075 data points
- **Platform:** Affymetrix Human Genome U133 Plus 2.0 Array
- **Unique Features:** First study to use pathogenesis-based transcript sets (PBTs) reflecting major biologic events in allograft rejection including cytotoxic T-cell infiltration, interferon-γ effects, and parenchymal deterioration; demonstrated that transcriptome disturbances in renal transplants have a stereotyped internal structure and are continuous rather than dichotomous across rejection and non-rejection states; validated findings in additional biopsy cohorts and provided quantitative measures of inflammatory disturbances in organ transplant
- **Link**: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13576

2.2 Empirical Results

Dataset 03 (GSE43346) Performance Results

Total Data Points: 11427075 = 54675 genes (samples) x 209

Clustering Configuration: 2 runs

Table 1. Davies-Bouldin index values for clustering validation across different distance metrics and cluster numbers

Number of Clusters	Davies-Bouldin Results			
	DDR	**EUC**	**MANH**	**CHISQ**
Clusters N° = 2	0.6739	0.6617	**0.6607**	0.6725
Clusters N° = 3	0.7789	0.6617	**0.6607**	0.6725
Clusters N° = 4	0.8645	0.8528	0.8019	**0.8015**
Clusters N° = 5	0.9492	**0.7523**	0.7690	0.7592
Clusters N° = 10	1.2051	**0.9877**	1.0316	0.9981

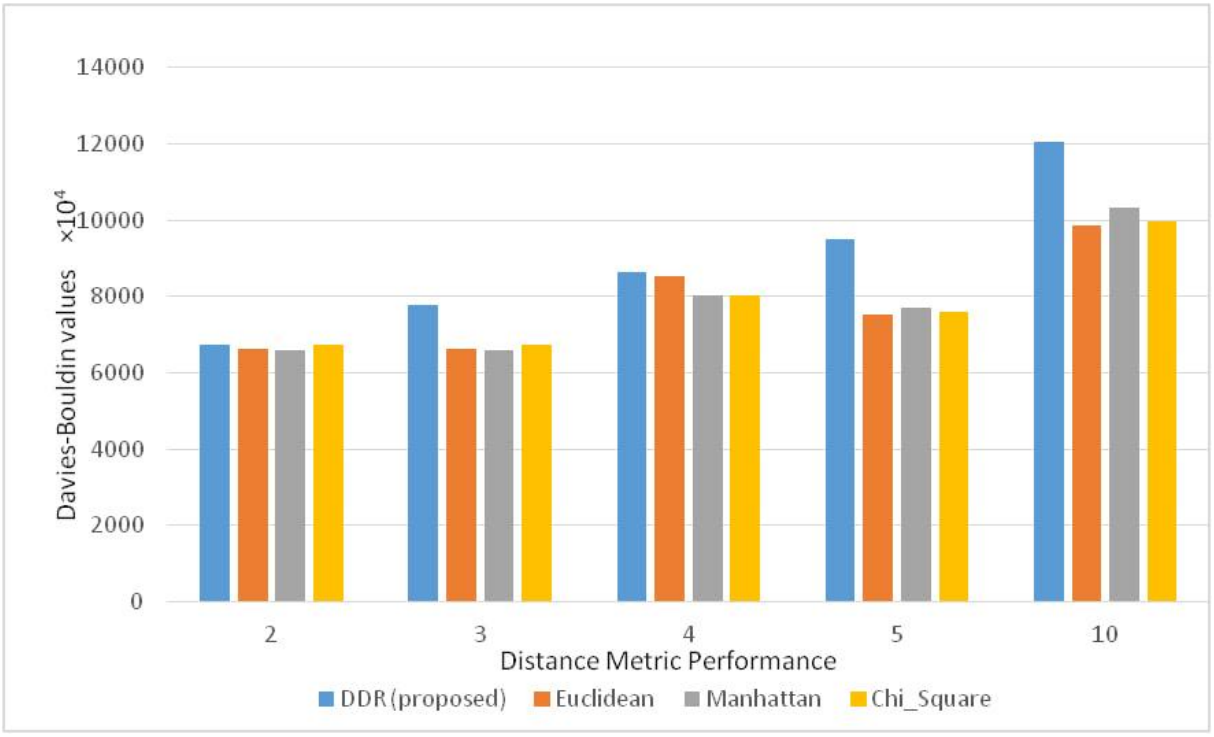

Figure 7: *Davies-Bouldin index comparison across distance metrics (DDR, Euclidean, Manhattan, and Chi-Square) for varying numbers of cluster*

2.3 Discussion

The results summarized in Table 1 offer valuable insights into the clustering performance of different distance metrics across varying numbers of clusters, as evaluated by the Davies-Bouldin Index (DBI). The DBI measures intra-cluster similarity and inter-cluster separation, where lower values indicate more compact and well-separated clusters. Across all distance metrics, the lowest DBI values are consistently observed for 2 clusters, suggesting that this configuration yields the most optimal clustering structure for the dataset under consideration. Specifically, the Manhattan distance produced the lowest DBI (0.6607), closely followed by the Euclidean (0.6617) and Chi-Square (0.6725) metrics, indicating marginal differences in their ability to assess cluster quality at this level. The DDR method also confirms this trend, with a relatively low DBI of 0.6739. As the number of clusters increases from 2 to 5, the DBI values generally rise, suggesting a gradual deterioration in cluster quality. This trend is evident across all metrics. For instance, with 5 clusters, the DDR and Manhattan distances yield higher DBI values of 0.9492 and 0.7690, respectively, reflecting reduced compactness and increased overlap among clusters. Interestingly, the Euclidean and Chi-Square distances

show slightly better performance at this point (0.7523 and 0.7592, respectively) compared to DDR. The sharpest decline in clustering quality is observed with 10 clusters, where all metrics report their highest DBI scores. The DDR method in particular reaches 1.2051, indicating poorly formed clusters. This outcome suggests that over-segmentation leads to diminished discriminability between clusters and potentially redundant groupings. Another noteworthy observation is the stability of DBI values for 2 and 3 clusters under Euclidean and Manhattan distances, both of which report identical values (0.6617 and 0.6607, respectively). This implies that increasing the number of clusters from 2 to 3 does not substantially impact the intra-cluster compactness or separation under these distance metrics, although DDR values do increase noticeably (from 0.6739 to 0.7789), reflecting a difference in sensitivity between the methods. In summary, the findings indicate that using 2 clusters yields the most effective clustering structure, regardless of the distance metric employed. While minor variations exist among the metrics, their trends are largely consistent. Over-increasing the number of clusters leads to a degradation in cluster quality, as evidenced by increasing DBI values. These results underline the importance of careful cluster number selection and metric choice when applying clustering algorithms to ensure the validity and interpretability of the results. **(Table 03)**

2.4 Limitations of the Proposed Method (DDR)

While the DDR method demonstrates competitive performance in clustering evaluation, particularly for lower cluster numbers, several limitations are evident from the results. Firstly, DDR tends to be more sensitive to an increase in the number of clusters compared to traditional distance metrics. This is reflected in the pronounced rise in the Davies-Bouldin Index from 0.6739 for 2 clusters to 1.2051 for 10 clusters—an increase that surpasses those observed with Euclidean, Manhattan, and Chi-Square distances. This suggests that DDR may struggle to maintain inter-cluster separation and intra-cluster cohesion when dealing with higher number of clusters.

.

GENERAL CONCLUSION

GENERAL CONCLUSION

This study addresses the critical challenge of choosing adequate similarity or dissimilarity measures aiming to analyze high-dimensional gene expression data. Such a task is a foundational one in genomics, mainly for applications like clustering, classification, and biomarker discovery. From traditional microarray platforms to RNA-sequencing and single-cell datasets, genomic data has proven to be increasingly complex. Thus, traditional measures such as Euclidean distance and Pearson correlation are often inappropriate because of their sensitivity to scale, noise, and outliers. In order to get over these limitations, a new similarity measure, named Distance-DissimRatio (DDR), was created. This developed metric is based on three key aspects: mean-centered deviations, direct expression differences, and comprehensive normalization. It is specifically designed to avoid the flaws of traditional metrics by offering:

- Scale independence, enabling cross-platform and multi-batch comparisons;
- Bounded influence, reducing the impact of outlier genes;
- Robustness to zero-inflation, essential for sparse data such as single-cell RNA-seq;
- Biological interpretability, preserving sensitivity to subtle, yet meaningful, expression differences.

Extensive repetitive experiments on three real-world gene expression datasets (GSE43346, GSE10072, and GSE13576) fortunately validated the effectiveness of DDR. The latter have even demonstrated superior or competitive performance across both F-measure and Davies-Bouldin Index metrics when it was evaluated against established metrics—Euclidean, Manhattan, and Chi-Square distances. This was clearly evident in scenarios involving high-dimensional and heterogeneous data, where traditional metrics showed instability or performance degradation.

The results suggest that DDR provides a more reliable and biologically meaningful way to quantify transcriptomic similarity, facilitating improved clustering quality and interpretability. These advancements have quite important implications for precision medicine since they improve the ability to not only stratify patient populations according to molecular signatures but also uncover disease-specific regulatory patterns.

To conclude, the proposed DDR methodology represents a robust, scalable, and interpretable framework that enhances the analytical toolkit available for high-dimensional genomic data analysis. Its performance across different datasets marks its potential for further application in genomic research as well as clinical bioinformatics.

REFERENCES

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. *Database Theory—ICDT 2001*, 420-434.

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Wiley.

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11(10), R106.

Ashley, E. A. (2016). Towards precision medicine. *Nature Reviews Genetics*, 17(9), 507-522.

Barabási, A. L., & Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2), 101-113.

Bar-Joseph, Z., Gifford, D. K., & Jaakkola, T. S. (2001). Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, 17(suppl_1), S22-S29.

Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When is "nearest neighbor" meaningful? *Database Theory—ICDT'99*, 217-235.

Bishara, A. J., & Hittner, J. B. (2012). Testing the significance of a correlation with nonnormal data: Comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychological Methods*, 17(3), 399-417.

Bolshakova N, et al. An integrated tool for microarray data clustering and cluster validity. Bioinformatics. 2005 (updated tools available).

Bolstad, B. M., Irizarry, R. A., Åstrand, M., & Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2), 185-193.

Brazma, A., & Vilo, J. (2000). Gene expression data analysis. *FEBS Letters*, 480(1), 17-24.

Bullard, J. H., Purdom, E., Hansen, K. D., & Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11(1), 94.

Chierichetti, F., Kumar, R., Lattanzi, S., Mitzenmacher, M., Panconesi, A., & Raghavan, P. (2010). On compressing social networks. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 219-228.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Lawrence Erlbaum Associates.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., ... & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17(1), 13.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2), 224-227. doi:10.1109/TPAMI.1979.4766909

Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. *Communications of the ACM*, 51(1), 107-113.

Deza, M. M., & Deza, E. (2009). *Encyclopedia of distances*. Springer.

D'haeseleer, P. (2005). How does gene expression clustering work? *Nature Biotechnology*, 23(12), 1499-1501.

D'haeseleer, P., Liang, S., & Somogyi, R. (2000). Genetic network inference: From co-expression clustering to reverse engineering. *Bioinformatics*, 16(8), 707-726.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (2nd ed.). Wiley.

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25), 14863-14868.
Eraslan, G., Avsec, Ž., Gagneur, J., & Theis, F. J. (2019). Deep learning: New computational modelling techniques for genomics. *Nature Reviews Genetics*, 20(7), 389-403.
Greenacre, M. (2007). *Correspondence analysis in practice* (2nd ed.). Chapman & Hall/CRC.
Handl, J., Knowles, J., & Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15), 3201-3212.
Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biology*, 18(1), 83.
Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
Huber, P. J., & Ronchetti, E. M. (2009). *Robust statistics* (2nd ed.). Wiley.
Ioannidis, J. P., Allison, D. B., Ball, C. A., Coulibaly, I., Cui, X., Culhane, A. C., ... & Hongyu, Z. (2009). Repeatability of published microarray gene expression analyses. *Nature Genetics*, 41(2), 149-155.
Ioffe, S. (2010). Improved consistent sampling, weighted minhash and L1 sketching. *2010 IEEE International Conference on Data Mining*, 246-255.
Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice Hall.
Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, 31(3), 264-323.
Jaskowiak, P. A., Campello, R. J., & Costa, I. G. (2014). On the selection of appropriate distances for gene expression data clustering. *BMC Bioinformatics*, 15(Suppl 2), S2.
Jiang, D., Tang, C., & Zhang, A. (2004). Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 16(11), 1370-1386.
Keogh, E., & Kasetty, S. (2003). On the need for time series data mining benchmarks: A survey and empirical demonstration. *Data Mining and Knowledge Discovery*, 7(4), 349-371.
Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., ... & Hemberg, M. (2017). SC3: Consensus clustering of single-cell RNA-seq data. *Nature Methods*, 14(5), 483-486.
Kuncheva, L. I. (2004). *Combining pattern classifiers: Methods and algorithms*. Wiley.
Landi MT, Dracheva T, Rotunno M, Figueroa JD et al. Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. PLoS One 2008 Feb 20;3(2):e1651. PMID: 18297132
Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1), 559.
Larsen, B., & Aone, C. (1999). Fast and effective text mining using linear-time document clustering. *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 16-22. doi:10.1145/312129.312186
Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., ... & Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10), 733-739.
Lipkovich, I., Dmitrienko, A., Denne, J., & Enas, G. (2013). Subgroup identification based on differential effect search—A recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine*, 32(17), 2934-2950.
Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Wiley.
Longdom. Transcriptomics and its Role in Understanding Gene Expression. 2024.
Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 550.

Ma J, Zhao X, Qi E, et al. Highly efficient clustering of long-read transcriptomic data with GeLuster. Bioinformatics. 2024;40(2):btae059.
Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.
Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., & Califano, A. (2006). ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(Suppl 1), S7.
Meyer LH, Eckhoff SM, Queudeville M, Kraus JM et al. Early relapse in ALL is identified by time to leukemia in NOD/SCID mice and is characterized by a gene signature involving survival pathways. Cancer Cell 2011 Feb 15;19(2):206-17. PMID: 21295523
Michiels, S., Koscielny, S., & Hill, C. (2005). Prediction of cancer outcome with microarrays: A multiple random validation strategy. *The Lancet*, 365(9458), 488-492.
Microbe Notes. Transcriptomics: Definition, Types, Techniques, Applications. 2024.
National Human Genome Research Institute (NHGRI). Gene Expression. 2025.
Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology*, 24(12), 1565-1567.
Quackenbush, J. (2001). Computational analysis of microarray data. *Nature Reviews Genetics*, 2(6), 418-427.
Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature Genetics*, 32(Suppl 4), 496-501.
Real, R., & Vargas, J. M. (1996). The probabilistic basis of Jaccard's index of similarity. *Systematic Biology*, 45(3), 380-385.
Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., ... & Sabeti, P. C. (2011). Detecting novel associations in large data sets. *Science*, 334(6062), 1518-1524.
Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), 139-140.
Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1), 59-66.
Rousseeuw, P. J., & Leroy, A. M. (2003). *Robust regression and outlier detection*. Wiley.
Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. McGraw-Hill.
Sato T, Kaneda A, Tsuji S, Isagawa T et al. PRC2 overexpression and PRC2-target gene repression relating to poorer prognosis in small cell lung cancer. Sci Rep 2013;3:1911. PMID: 23714854
Singhal, A. (2001). Modern information retrieval: A brief overview. *IEEE Data Engineering Bulletin*, 24(4), 35-43.
Song, L., Langfelder, P., & Horvath, S. (2012). Comparison of co-expression measures: Mutual information, correlation, and model based indices. *BMC Bioinformatics*, 13(1), 328.
Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: The teenage years. *Nature Reviews Genetics*, 20(11), 631-656.
Stuart, J. M., Segal, E., Koller, D., & Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643), 249-255.
Stuart, T., & Satija, R. (2019). Integrative single-cell analysis. *Nature Reviews Genetics*, 20(5), 257-272.
Svensson, V., Vento-Tormo, R., & Teichmann, S. A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols*, 13(4), 599-604.
Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., ... & Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps. *Proceedings of the National Academy of Sciences*, 96(6), 2907-2912.
Tan, P. N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Addison-Wesley.
van Rijsbergen, C. J. (1979). *Information retrieval* (2nd ed.). Butterworths.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3), 645-678.
Yousefi B, Schwikowski B. Consensus Clustering for Robust Bioinformatics Analysis. bioRxiv. 2024.

Année universitaire : 2024-2025	**Présenté par :** AOUISSI BOUCHRA HADJ AZZAM SARRA

A Dissimilarity Measure for High-Dimensional Gene Expression Datasets: "Distance-DissimRatio" for Quantifying Transcriptomic Variation.

Mémoire pour l'obtention du diplôme de Master en Bio-informatique

RESUME

La sélection de mesures de similarité appropriées pour les jeux de données d'expression génique de haute dimension représente un défi critique en génomique computationnelle, particulièrement avec l'évolution des technologies génomiques des plateformes de puces à ADN vers le séquençage ARN et les applications unicellulaires. Les métriques traditionnelles telles que la distance euclidienne, la distance de Manhattan, la corrélation de Pearson et la distance du Chi-carré s'avèrent souvent inadéquates face aux caractéristiques complexes des jeux de données génomiques modernes, incluant une haute dimensionnalité extrême, des motifs d'inflation de zéros étendus, des effets de lot systématiques et des profils de bruit hétérogènes. Pour adresser ces limitations fondamentales, nous avons développé la méthodologie Distance-DissimRatio (DDR), une mesure de similarité novatrice qui intègre l'analyse des déviations centrées sur la moyenne, l'évaluation des différences d'expression directes, et un cadre de normalisation compréhensif. La méthodologie DDR présente des caractéristiques uniques incluant l'indépendance d'échelle, les propriétés d'influence bornée, la robustesse à l'inflation de zéros, et une interprétabilité biologique améliorée. Une validation expérimentale complète a été menée sur trois jeux de données d'expression génique diversifiés : GSE43346 (68 échantillons × 54 675 gènes), GSE10072 (107 échantillons × 22 283 gènes), et GSE13576 (209 échantillons × 54 675 gènes). L'évaluation de performance utilisant la F-mesure et l'Index de Davies-Bouldin a démontré la performance supérieure de DDR comparée aux métriques traditionnelles. DDR a atteint une F-mesure maximale parfaite (1,0) et un index de Davies-Bouldin optimal (2,2071) pour GSE43346, la F-mesure moyenne la plus élevée (0,9883) pour GSE10072, et une performance compétitive pour GSE13576. La méthodologie démontre une complexité computationnelle linéaire O(n), permettant une analyse efficace des jeux de données génomiques à grande échelle. Ces résultats établissent DDR comme un cadre robuste, évolutif et biologiquement interprétable avec des implications significatives pour la médecine de précision, la découverte de biomarqueurs, et l'analyse des réseaux de régulation génique.

Mots-clefs : Analyse de l'expression génique, mesures de similarité, données de haute dimension, génomique, Classification, découverte de biomarqueurs, médecine de précision, biologie computationnelle.

Laboratoires de recherche :laboratoire de ……..…….(U Constantine 1 Frères Mentouri).

Président du jury : Pr/Dr CHEHILI Hamza (PROF/MC(A) - UConstantine1 Frères Mentouri).

Encadrant: Dr KENIDRA BILEL (MC(B) / PROF- UFM Constantine 1).

Examinateur(s) : Dr MEZIANI Dahbia Yasmina (MC(B) / PROF - UFM Constantine 1),