**Development of a data-container to efficiently manage subsets of scRNA-Seq data**

A thesis submitted by

**Irzam Sarfraz**

(19-NTU-MSCS-4008)

****

In partial fulfillment of the requirement for the degree of

Master of Science

in

Computer Science

Directorate of Graduate Studies & Research

Faculty of Science

**National Textile University, Faisalabad**

**December 2020**

**DEDICATION**

This modest effort is dedicated to my

Respected Teachers, Parents, Family Members,

&

Friends

# CERTIFICATE

**NATIONAL TEXTILE UNIVERSITY, FAISALABAD**

This thesis is written by Mr. Irzam Sarfraz (19-NTU-MSCS-4008) under the direction of his supervisors and approved by all the members of the thesis committee, has been presented to and accepted by the Director of Graduate Studies and Research, in fulfillment of the requirement of the degree of Master of Science in Software Engineering.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Supervisor Co-Supervisor

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Internal Examiner External Examiner

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Chairman, Director,

Department of Computer Science Graduate Studies & Research

Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_

# ACKNOWLEDGEMENT

Above everything else, I exalt and bow my head before “**ALLAH**” The most beneficent and merciful. I offer my humblest thanks to Him for bestowing upon me. His blessings with intellectual capability to work hard and successful accomplishment of my dissertation. Trembling lips and wet eyes praise for Holy Prophet Muhammad (SAW), who is forever a beacon of guidance and knowledge for humanity as a whole and enables us to recognize our creator.

I feel pleasure and honor of this available opportunity to offer sincerest thanks and gratitude to my worthy supervisor Dr. Muhammad Asif, Associate Professor, Department of Computer Science, National Textile University Faisalabad, for his kind and affectionate guidance and enlightened supervision during my research work. His wise kindness and help have been evident like the midday sun as a teacher as a supervisor. I also thankful, especially to my co-supervisor, Mr. Shahbaz Ahmad, Lectuere, Department of Computer Science, National Textile University Faisalabad, for the motivation and efforts of days and nights to accomplish this research work. A special thanks to Dr. Sohail Jabbar**,** Assistant Professor, Department of Computer Science, National Textile University Faisalabad, for their motivation, who build my research interests and skills, and I am also grateful for the supervisory committee.

I want to express my deep love, earnest thanks, and appreciate the moral support given by Parents for my success and for sacrificing their comforts to enable me to achieve this level of education.

I want to thank my sincere friend Umair Yousaf for every kind of practical support to make me capable of conducting various experiments.

I cannot forget the contributions of my friends, classmates, research fellows, and the people who were the part of my research. Their help and behavior will always be remembered and acknowledged.

*May Allah bless all these people with long, happy, and peaceful lives (Ameen)!*

***Noman Tahir***

***17-NTU-4147***

**TABLE OF CONTENTS**

**Description Page**

[DEDICATION ii](#_Toc53570916)

[CERTIFICATE iii](#_Toc53570917)

[ACKNOWLEDGEMENT iv](#_Toc53570918)

[LIST OF ABBREVIATIONS viii](#_Toc53570919)

[LIST OF TABLES ix](#_Toc53570920)

[LIST OF FIGURES x](#_Toc53570921)

[ABSTRACT xi](#_Toc53570922)

[Chapter 1 1](#_Toc53570923)

[Introduction 1](#_Toc53570924)

[1.1. Basic concepts and terminologies 2](#_Toc53570925)

[1.1.1. Keywords 2](#_Toc53570926)

[1.1.2. TF-IDF 3](#_Toc53570927)

[1.1.3. Machine learning 3](#_Toc53570928)

[1.1.4. Graphs 3](#_Toc53570929)

[1.2. Background 4](#_Toc53570930)

[1.3. Problem statement 5](#_Toc53570931)

[1.4. Research objectives 5](#_Toc53570932)

[1.5. Research questions 5](#_Toc53570933)

[1.6. Scope of the research 6](#_Toc53570934)

[1.7. Application of proposed research 6](#_Toc53570935)

[1.8. Thesis outline 6](#_Toc53570936)

[Chapter 2 7](#_Toc53570937)

[Literature review 7](#_Toc53570938)

[2.1 Non-graph-based (statistical) techniques 8](#_Toc53570939)

[2.2 Machine learning approaches 11](#_Toc53570940)

[2.3 Graph-based techniques 13](#_Toc53570941)

[Chapter 3 23](#_Toc53570942)

[Research methodology 23](#_Toc53570943)

[3.1 Dataset collection 25](#_Toc53570944)

[3.2 Plain text conversion 26](#_Toc53570945)

[3.3 Preprocessing 27](#_Toc53570946)

[*3.3.1 Tokenization 27*](#_Toc53570947)

[*3.3.2 Stop words removal 28*](#_Toc53570948)

[*3.3.3 Stemming 28*](#_Toc53570949)

[*3.3.4 Parts of speech (POS) tagging 29*](#_Toc53570950)

[3.4 N-grams generation 29](#_Toc53570951)

[3.5 Statistical / Non-graph-based approach 31](#_Toc53570952)

[3.5.1 *TF-IDF* 32](#_Toc53570953)

[3.6 Machine learning 32](#_Toc53570954)

[3.6.1 *Quality phrase mining* 33](#_Toc53570955)

[3.7 Building network (graph generation) 33](#_Toc53570956)

[3.8 State-of-the-art network-based approaches 35](#_Toc53570957)

[*3.8.1 HITS-I 35*](#_Toc53570958)

[*3.8.2 PageRank 37*](#_Toc53570959)

[*3.8.3 Proposed Method (Frequent Node method) 38*](#_Toc53570960)

[Chapter 4 40](#_Toc53570961)

[Results and discussions 40](#_Toc53570962)

[4.1 Precision and recall as a result evaluation parameter 40](#_Toc53570963)

[4.2 N-grams based results statistics: 41](#_Toc53570964)

[*4.2.1 Unigram 42*](#_Toc53570965)

[*4.2.2 Bigram 42*](#_Toc53570966)

[*4.2.3 Trigram 42*](#_Toc53570967)

[*4.2.4 Trigram 43*](#_Toc53570968)

[*4.2.5 Fourgram 44*](#_Toc53570969)

[*4.2.6 Machine Learning 45*](#_Toc53570970)

[4.3 Answer to research question 46](#_Toc53570971)

[Chapter 5 47](#_Toc53570972)

[Conclusion, limitations, and future work 47](#_Toc53570973)

[5.1 Limitation and Future Work: 47](#_Toc53570974)

[References 49](#_Toc53570975)

# LIST OF ABBREVIATIONS

**Abbreviations Descriptions**

HITS = Hypter Texts Induced Topic Search

IR = Information Retrieval

KEA = Keyword Extraction Algorithm

NLP = Natural Language Processing

QPM = Quality Phrase Mining

RNN = Recurrent Neural Network

SVM = Support Vector Machine

TF-IDF = Term Frequency Inverse Document Frequency

TKG = Twitter Keyword Graph

# LIST OF TABLES

[Table 2.1 Overview literature study 20](#_Toc59351890)

[Table 3.1 Overview of nodes file containing the node number and node name concatenation 36](#_Toc59351891)

[Table 3.2 Structural view of arcs file showing the connection building between nodes 37](#_Toc59351892)

# LIST OF FIGURES

Figure 1.1 Graph-based processing concept for key term identification ………………..………...4

Figure 2.1 Generic concept of statistical methods for key terms extraction ……………………... 9

Figure 2.2 Machine Learning approach’s conceptual overview ……………………………...…..12

Figure 2.3 Generic concept of graph-based methods for keyword extraction ………………...….13

Figure 2.4: Basics of HITS algorithm as a graph-based approach for keyword extraction………. 15

Figure 2.5: PageRank Algorithm process …………………………………………………………16

Figure 3.2: Research Methodology …………………………………………………………...…..23

Figure 3.3 Dataset statistics according to the no. of articles selected for specific years……........26

Figure 3.4: Workflow of Preprocessing…………………………………………………………... 27

Figure 3.5: POS Tagging through constraint of proposed method …………………………….…30

Figure 3.6: N-grams structure ……………………………………………………………….....…31

Figure 3.7: Node number assigning process ………………………………………………….…..34

Figure 3.8: Building Graph of Words……………………………………………………….….…35

Figure 3.9: Graph-based methods overview …………………………………………………...….35

Figure 3.10: The representation of the PageRank algorithm in terms of nodes and edges …….…38

Figure 3.11 FN method implementation based overview …………………………………………40

Figure 4.1 Graph of unigram results for all methods ………………………………………......…43

Figure 4.2 Graph of Bigram dataset results for all methods ………………………………………44

Figure 4.3 Graph of Trigram dataset results from all methods ……………………………..….…45

Figure 4.4 Fourgrams dataset results of all methods ………………………………………..….…46

Figure 4.5 Machine learning methods result ………………………………………………….…..47

Figure 4.6 Overall result statistics of all methods with bigram that is best in n-grams…………...48

# ABSTRACT

Keyword extraction is essential in determining influenced keywords from huge documents as the research repositories are becoming massive in volume day by day. The research community is drowning in data and starving for information. The keyword extraction plays a great role in determining the influenced keywords from massive documents. The keywrods are the words that describe the theme of the whole document in a precise way by consisting of just a couple of words. Furthermore, many state-of-the-art approaches are available for keyword extraction from a huge collection of documents and are classified into three types, the statistical approaches, the machine learning, and the graph-based methods. The machine learning approaches require a large training dataset that needs to be developed manually by domain experts, which sometimes is difficult to produce while determining influenced keywords. However, this research focused on enhancing state-of-the-art graph-based methods to extract keywords when the training dataset is unavailable. This research first converted the handcrafted dataset, collected from textiles and its application impact factor journals into n-grams combinations from unigram to pentagram and enhanced traditional graph-based approaches. The experiment was conducted on a handcrafted dataset, and all methods were applied to it. Domain experts performed the user study to evaluate the results. The results were observed from every method and were evaluated with the user study using precision and recall as evaluation matrices. The results showed that the proposed method performed well and scored near the machine learning approaches.

# Chapter 1

# Introduction

Over the past decade or so, due to the extensive availability of computational resources, computational approaches have contributed significantly to offer efficient and effective solutions to intricate research questions in domains that do not directly offshoot from traditional computer science paradigm. Computational Biology, Computational Biomedicine and Bioinformatics are some of the recently developed domains that have emerged directly from the amalgamation of Biology, Computer Science and Statistics due to a strenuous need of solving complex biological problems in the era of big data and informatics.

Some of the grand challenges that been worked on extensively over the past few decades include genomic sequence analysis, homology search and protein structure prediction. Overall, end goal of all of these challenges is to infer a better understanding of the underlying biological complexity and to somehow use this understanding to combat diseases that were previously uncurable or untreatable.

At the same time, the availability and the use of computational resources have increased by manifold over the last decade and therefore this provides an excellent opportunity to use these resources to solve the above-mentioned complex problems. Genomic sequence analysis is one the problems that is of particular interest to the researchers around the world due to its explanatory power of various diseases, their prognosis and outcomes.

Genomic sequence extraction technologies have improved drastically since the beginning of this millennium and have only continued to grow. These technologies offer extraction of genetic data at very high resolution (even possible at cellular level) and the analysis of this data can result in extremely useful biological insights. For example, Precision Medicine is an emerging domain that combines genetic data extracted through these technologies with the patient clinical data to predict patient outcome.

These computational analyses that lead to useful biological insights are often quite complex, use large amounts of data and consume vast computational resources. It is therefore, utmost important, to tackle these analyses by developing methods, tools and data structures that can conserve one or more of these resources. Additionally, because most of these analyses are quite complex, lengthy and require the use of supplementary data (in addition to the primary genomics data), it becomes quite tricky to keep track of the different shapes of the data that have emerged over the analysis pipeline. Therefore, in this study, we focus on developing a data-structure that efficiently manages subsets of high-throughput genomic sequence data by eliminating the need to redundantly store additional data which is a common occurrence in many of the computational analyses of biological high-throughput data and somehow deliver features for provenance tracking for easier tracking different shapes of the data throughout an analysis workflow.

# Basic concepts and terminologies

As the proposed study/project has applications in Bioinformatics, a brief introduction and explanation of the key concepts and terminologies is required. Some of these are given below.

# *High-throughput genomic data*

. show pic etc.

High-throughput genomic technologies such as microarrays and sequencing allow rapid generation of vast amounts of data from both DNA and RNA of an origanism that is both large in size and high-dimensional. These high-throughput genome extraction technologies are widely used to profile the genome of an organism, to better understand the genetic structure of an organism or even to identify and address the causes of certain diseases and defects. In contrast, the traditional PCR-based approaches are more concerned with amplification and detection of a few transcripts of interest as compared to hundreds of thousands in high-throughput approaches. Most of the times, the output of high-throughput sequeuncing experiments are counts matrices where each column of the matrix represents a particular sample and the row represents the quantified presence of a particular gene or transcript in that sample. These matrices form the basis of a large amounts of statistical analyses that can be performed to gain insight about the underlying biology of the samples.

# *Assay*

. pic etc.

An assay (or a data assay) is the actual form of the data that can be used with a particular statistical analysis method. Initially, this assay is the counts matrix which is output from a sequencing experiment but can undergo several transformations throughout its lifetime during an analysis workflow. For example, the initial assays (counts matrix) can undergo normalization, log-transformation, scaling, trimming among others and are stored as separate assays which be called and used during the analysis workflow as per requirements of the workflow and the method being used. However, it must be understood that these multiple assays belong to the same experiment but only represent the multiple transformations that have been performed on the original assay so as to have original assay in all forms. The assays always have dimensions of m\*n where m is the number of genes (rows) and n is the number of samples (columns).

# *colData*

. pic/figure

The ‘colData’ is short for column data and is always the metadata associated with the samples (columns) of the dataset (columns of an assay). The ‘colData’ has dimensions n\*q where n is the number of samples and q >= 0 based on the number of metadata columns (for samples) available for a particular dataset. The ‘colData’ is often referred to as the phenotype data as it is associated with the phenotypic characteristics (patient data) of the dataset. For example, the ‘colData’ may have the disease condition of each sample in the dataset (tumor or normal) or age of the patient that the sample belongs to. This ‘colData’ may be used in many downstream analysis methods or visualizations for labeling of samples or even in statistical methods to separate the samples based on a phenotype, e.g., a linear model may be used between normal and tumor tissues (referenced from colData) to identify significant genes based on the quantified assay data.

# *rowData*

. pic/figure

The ‘rowData’ as obvious is the metadata associated with the genes (rows) of the dataset (rows of the an assay). The ‘rowData’ has dimensions m\*q where m is the number of genes and q>=0 abased on the number of metadata columns (for genes) available for a particular dataset. The ‘rowData’ is often used to store the computed statistics for the genes in the dataset. For example, during the highly variable genes computation, the variance and the mean computed for each gene is stored in separate columns of the ‘rowData’ which may be used at a later stage to identify the most highly variable genes from these statistics. Additionally, ‘rowData’ may be used to store anything that may be related to the genes of the dataset, e.g., genes are often referenced by several different systems of naming conventions (ENSEMBLE, Entrez or Gene Symbols) which can be stored in the ‘rowData’.

# *Bioconductor & CRAN*

. figure

Packages or libraries built for R environment are available for download and use from multiple sources including R CRAN, Bioconductor or even directly from Github. The Comprehensive R Archive Network (CRAN) is the largest repository for R packages maintained by R developers, while Bioconductor is a community maintained (R core) repository (as well as a database for sample datasets) and specializes in packages/libraries with applications in biology. Submission to both of these repositories undergo strenuous review (including novelty of the package, code quality and testing) before it is accepted and becomes available for users to download and use. Packages from both of these resources once available can easily be downloaded from within R console environment.

# *Bioconductor Experiment objects*

. add figs

As previously explained that Biocodunctor is a database/repository for R packages and datasets having applications in biology, data-containers or data-structures are required to particularly handle the such high-dimensional datasets which can manage multiple data assays and the associated phenotype data (colData) and the gene metadata. While, this can be managed by using native R data-structures like matrix and dataframes, it is much more intuitive to handle this multi-data datasets using Experiment classes that are designed specifically for this purpose.

These Experiment classes including SummarizedExperiment and SingleCellExperiment are designed in such a way as to allow the users to use the objects of these classes in a more biologically and statistically relevant manner especially considering the workflow of such datasets. For example, these classes allow storage of multiple assays, colData, rowData as well as other types of data such as dimensionality reduction results within the same object as all of these data types are used within an analysis workflow from the same experiment.

Moreover, because these data-containers are specialized, they acquire multiple features necessary for such high-dimensional data such as use of dgCMatrix (compressed sparse matrices) for assay data, dataframes for colData/rowData and SimpleList for dimensionality reduction results, all of which have additional advantages instead of using a single base data-stucture for all types of data.

# Background

High-throughput experiments like microarrays and RNA sequencing allow simultaneous extraction of data from thousands of genes at the same time, thus enabling researchers and scientists to study the effects of mutations (change in gene structure) and expression changes (quantifying the presence or absence of a gene) in genes that result in phenotype variations [1]. These changes often help us in understanding the underlying causes of numerous diseases, particularly heterogeneous diseases such as cancer, where different cells may represent individual changes at the genetic level [2]. Single Cell RNA Sequencing (scRNA-Seq) is effectively a recent advancement that allows data extraction at a very high resolution i.e. at the cellular level and reveals expression changes in genes against each individual cell in the sample space. In contrast, microarrays [3], and bulk RNA sequencing [1] both average the expression values extracted from the tissue of interest. This high-resolution capture of genomic data at cellular level (scRNA-Seq) has enabled researchers belonging to a broad range of scientific domains to come together and collaborate across the board to solve biological problems including but not limited to drug discovery [4], artificial intelligence in precision medicine [5], targeted patient therapies and biological data mining to better understand the disease progression and outcomes [6]. All of the previously mentioned biological problems require a computational analysis of some sort to reach an insightful conclusion particularly owing to the fact that these analyses often run on scRNA-Seq datasets that are generally of the order of hundreds if not thousands of gigabytes [7]. These computational analyses can range from a simple visualization of data for quality control analysis or differential expression to identify interesting genes and not to mention even for the understanding of cell trajectories over pseudotime [8], which is a rather complex task than the former two. To design and run these analyses, several computer scripting languages such as R [9] and Python [10] support manipulation of scRNA-Seq data due to the extensive availability of relevant packages, libraries, and toolkits. Bioconductor [11] platform serves as a central repository for all such tools and packages to analyze high-throughput genomic data in general for scripting in R language. The first and the foremost task of any such analysis is the choice of a data structure or more appropriately, a container to hold the scRNA-Seq data for manipulation during all subsequent tasks of an analysis. This is relatively a tricky decision because most of the toolkits and analytical pipelines for analysis of scRNA-Seq data support only a few of the proposed containers and occasionally provide their internal objects to allow easier manipulation within the toolkit. However, issues arise when a range of analytical methods are integrated in a single toolkit or an analytical pipeline, as is the case with many of the proposed command line toolkits or Shiny web-interfaces [12]. One issue that is of particular interest, is the generation and management of subsets of scRNA-Seq data, the manipulation of which is abundant in many of the common tasks of analysis pipeline such as during the identification of highly variable genes [13] or the integration of multiple assays [14] or even basic filtering steps. In such cases, the available choice of containers does not provide flexibility in terms of storage of such subsets and the consequent usage in the downstream analysis. In many cases where there are several libraries and methods integrated together to form a streamlined analysis pipeline, as in the case of many toolkits and Shiny applications [15], multiple objects from these containers at times hold the same data albeit in subset form. This has widespread consequences in terms of increased code complexity against the underlying analysis, data redundancy, and increased memory consumption as well as complexity that arises from the handling of multiple objects created due to the subsets of the data. Furthermore, because of the creation of multiple objects that essentially point to the same data, provenance tracking is often overlooked. Typically, in such analyses, the utmost concern of the researcher is to apply the appropriate statistical technique to gain useful biological insights. Alternatively, in the case of a toolkit programmer, the main concern is the addition of functionality to support such analyses. In both scenarios, data handling and management is ultimately at the bottom of the workflow stack. Therefore, the complexity and redundancy that emerges from the creation of a collection of objects due to the nested nature of the subsets that are essentially pointing to the same data, results in disastrous implications as far as time and memory consumption are concerned. These factors become especially significant when datasets are enormous [7] and the corresponding analyses are tricky, as is the case with scRNA-Seq datasets.

# Problem statement

The available data-structure classes for manipulation of high-throughput sequencing data do not support storage, management and provenance tracking of subsets of the data which are quite common in a typical analysis workflow. Therefore, there is a need to develop a data-structure class that can efficiently manage subsets of such data while maintaining data provenance.

# Research objectives

* Development of a data container that supports storage of subsets of sequencing data and allows efficient manipulation of subsets while significantly reducing the memory footprint.
* Interface to the user must be same as offered by SingleCellExperiment and SummarizedExperiment classes to ensure easy replacement of existing code with the new package.
* Ensure conservation of memory by eliminating the need to store redundant data values against all the data assays, rowData and colData, by using pointers to the existing data whenever and wherever possible.
* Provenance tracking to ensure that the origin of the data can easily be tracked.
* Apply this data container to commonly used analysis workflows and show how efficiently it manages subsets of data from within a single object while saving memory.

# Research questions

RQ 1: How can we develop a data-structure that can efficiently manage subsets of high-throughput genomic data while preserving data provenance and ensuring no additional redundant data is stored?

# Scope of the research

i. The proposed data-container shall be available as an R package through the Bioconductor repository accompanying a vignette that illustrates the description and common usage of the available methods with examples. ii. Inheritance from the commonly used data-containers shall ensure that interface is offered to the users for manipulation is same as the input class. iii. Single Cell RNA Sequencing data is the primary data supported by the package, yet all Bioconductor experiment datasets should work fine if they follow the experiment design of the SingleCellExperiment and SummarizedExperiment classes i.e., numeric matrices for assay data, and numeric/character data frames for rowData and colData. iv. Package shall be platform independent and should work with all operating systems supported by R language.

# Application of proposed research

As the direct beneficiaries of our research, the analysts, researchers and developers, can indeed build analysis pipelines and workflows for genetic sequencing data in a much better, convenient and efficient manner by using our approach as a drop-in replacement and a building block for already available containers for such data that often include subsets of such data. Indirectly, it is the patients and the general public that is the recipient of the outcomes of the research of this advancement and consequent improvements in such approaches, including ours that impacts the disease prognosis, its outcome and its treatment. While the abstract utility of our proposal lies in Bioinformatics (particularly in biological data mining and analysis), the proposed implementation and the direct application indeed acquires knowledge from a range of multidisciplinary domains (data mining, computational biology and statistics) all of which have a common root in Computer Science.

# Thesis outline

Chapter 1: Deals with the introduction, basic terminologies, background knowledge, problem statement, research objectives, research question and scope of the research (given in the Section 1.1,1.2,1.3,1.3,1.5, and 1.6 respectively).

Chapter 2: Discusses the literature review according to the statistical, machine learning, and graph-based approaches (given in the Section 2.1, 2.2 and2.3, respectively).

Chapter 3: Deals with the detailed description of proposed research methodology from data collection to implementation (given in the Section 3.1 to 3.8, respectively).

Chapter 4: Deals with the representation and discussion of results (given in the Section 4).

Chapter 5: Discusses the conclusion, limitation, and future work (given is the Section 5)

# Chapter 2

# Literature review

In a typical genomic experiment, a matrix-type data-structure represents the data extracted from the experiment, where the rows of this matrix-type data-structure represent the genes or features and columns represent the samples. The row-column intersection of such a matrix records the data-value for that gene in the row against a sample. The data-value can be either the expression value of a gene in the case of microarray data or counts of a gene if the data is from a sequencing experiment. Additionally, these values may also represent transformed data, such as when the matrix undergoes log-transformation, normalization, or scaling as per the requirements of the downstream analysis. SummarizedExperiment [16] originally proposed to serve as an all-in-one container for sequencing data supports the storage and manipulation of multiple data assays as well as rowData and colData. SummarizedExperiment follows the ExpressionSet [17] class available for microarray experiments, technically following a similar structure but allows more flexible manipulation in terms of management of additional assays and feature information within a single object. Built on top of SummarizedExperiment (more accurately the RangedSummarizedExperiment class) is the SingleCellExperiment [18] class that in addition to the features offered by the former, allows the storage of dimensionality reduction computations. Moreover, SingleCellExperiment class offers storage of alternate experiments such as in the case of spike-in transcripts [19] that have different dimensions from the original assays. The design of both SingleCellExperiment and SummarizedExperiment mandates the use of the object from these classes as a container for a single experiment that can have multiple data assays. For example, PBMC3K [7] dataset has 13714 features (genes) and 2700 samples (cells) available for use and the data values are represented by a counts matrix, each cell of which contains the number of molecules identified against a feature and a sample. This represents a single experiment and all relevant data for this experiment i.e. the actual data values as represented by the counts matrix (assay), feature metadata (rowData) and sample metadata (colData), group together in a single SingleCellExperiment or SummarizedExperiment object. These classes allow the storage of multiple assays i.e. the transformed versions of counts matrix such as the scaled counts matrix or log transformed counts matrix, within a single object as they correspond to the same experiment. In doing so, both classes impose restrictions on the dimension of the objects and the data supported by these objects. For example, in the case of the generation of a subset of a counts matrix, as is the case of many analysis tasks such as during variable genes identification, these classes do not allow the storage of the subset counts matrix corresponding only to the variable genes back into the original objects. Therefore, for this purpose, only a new object can handle this subset data which again is not only the subset counts matrix, but also the corresponding rowData and colData. When an experiment constitutes multiple observations having varying dimensions, MultiAssayExperiment [20] offers integration of these observations through the sampleMap function available with the package. More intuitively, integration of datasets within a single object as offered by MultiAssayExperiment is more oriented towards selection of features represented by a specific condition. For example, consider a scenario where huge amounts of data are available against a particular disease, yet it is extracted by different protocols thus resulting in non-uniform datasets as far as features, possible sub-types of diseases and the associated origin of the samples are concerned. Consider cancer for example [21], a heterogenous disease that results in different progression and prognosis based upon different factors with geography being one of them. In this case, MultiAssayExperiment offers easy integration of this diverse data with respect to a defined condition, such as the patient geography and allows usage of features only represented by this condition. In addition to the above packages that primarily serve as a data-structure or a data-container for holding and manipulating genomic data such as scRNA-Seq data, some specific pipelines and toolkits developed for analysis of scRNA-Seq data also offer native objects for the same purpose. A widely used toolkit for analysis of scRNA-Seq data i.e. Seurat [14][22], provides a native Seurat object for manipulation of data within the toolkit. While it offers flexibility in terms of storage of multiple assays without imposing restrictions on assay dimensions, little support outside of the toolkit makes it a less desirable option when compared with the dedicated containers such as the widely used SingleCellExperiment container. All the previously discussed packages and containers support the common R paradigm for subsetting data, either with built-in functions or through standard R syntax for subsetting matrices and data frames. However, storing back these subsets into the original objects and then using them for further data manipulation or transformation while keeping the original data and the subset data does not fall into the design considerations of these packages.

# SummarizedExperiment

SummarizedExperiment is an S4-based R class to serve as a data-container for sequencing experiments with properities similar to the ExpressionSet class historically used for microarray experiments. SummarizedExperiment offers a all-in-one solution to store and coordinate expression measurements from sequencing experiments with support for many other statistical functions and methods that can directly work on SummarizedExperiment objects.

The slots available with SummarizedExperiment objects include assay slot to hold multiple data assays (using SimpleList), colData slot to hold column metadata (dataframe), rowData slot to hold row metadata (datafrane) and metadata slot to hold general metadata (list) about the dataset. Common methods provided by SummarizedExperiment class include assay() method to get or set an assay, colData() method to get or set colData, rowData() method to get or set rowData() and a metadata() method to get or set metadata to SummarizedExperiment object.

. give diagram

However, SummarizedExperiment objects have a limitation of not letting users to store assays of different sizes within a single object. This is because of its uniform structure that coordinates the assays with rowData/colData by keeping a object-level row/column size of the overall object to avoid complications. As a result, when assays are subsetted as in many cases during an analysis workflow, they cannot be stored back into the original object.

# SingleCellExperiment

SingleCellExperiment is a data-container specifically built for storage and manipulation of single-cell data in contrast to the SummarizedExperiment class which has applications throughout gene expression measurements regardless of the data extraction protocol. As a result, SingleCellExperiment has additional slots for single-cell data manipulation while keeping all of the slots of SummarizedExperiment intact by using a direct inheritance approach.

The additional slots that have been added to SingleCellExperiment class are the ‘reducedDims’ for storage and manipulation of dimensionality reduction results and ‘altExps’ for storage of additional alternative experiments. Both of these slots have been added keeping in mind the specific needs of single-cell data, which due to its extraction protocol requires the addition of spike-in genes to the data assays to somewhat transform the data in a shape better suited for downstream analysis by adjusting for excess zeros which is a major characteristic of single-cell data.

While the `altExps` slot allows the storage of a complete experiment object of different dimensions, it is originally meant for the manipulation of spike-in genes and therefore lacks coordination of data between multiple dimensions. As a result, it does not support direct manipulation of subsets of data similar to the SummarizedExperiment class.

. give diagram

# MultiAssayExperiment

MultiAssayExperiment is another Experiment class for storage of expression data, but allows integration of data from multiple experiments and provides a common interface to this integrated data to the user. It is different from both SummarizedExperiment and SingleCellExperiment classes in its overall objective, which is the integration of datasets through a common variable instead of acting as a basic container for day-to-day manipulation.

For example, if multiple datasets are available for lung cancer, they can be integrated together through a sampleMap() function provided by the class that uses an anchor column for phenotype data to merge and integrate together these many datasets having different characteristics such as age, type of lung cancer, site of tumor tissue and many others.

The MultiAssayExperiment allows the storage of multiple datasets having different dimensions but it is more oriented towards integration of data and therefore does not provide a straightforward approach towards subsetting of single experiments.

. give diagram

. add one table that shows comparison of these classes

# Chapter 3

# Research methodology

The overall research methodology for this study includes the package design (high-level logical design to support subsets in existing \*Experiment packages), package implementation (development and deployment of package in R environment, including package documentation) and the validation of the efficiency of the package on different datasets (results). The research methodology can be perceived easily from the Figure.

Graphical user interface, text, application, chat or text message

Description automatically generated

## Package Design

Keeping in view the objectives and scope of the package, an abstract logical design of the package utility is presented in the figure.

Diagram

Description automatically generated

In the figure above, an abstract design for the utlitiy of the package is presented. The goal of the package is to provide the ability to create subsets from available main data assays (counts, logCounts etc. from the figure) and link them in a hierarchical order. This linking of assays allows the creation of hierarchical subsets which are common during large and complex analysis workflows. Essentially, subsets can be created from either the main assays, or other subsets, where the the newly created subset can have a subset of rows or columns or a combination of both.

A picture containing diagram

Description automatically generated

The figure above demonstrates the general class structure of the package.

Timeline

Description automatically generated with medium confidence

The figure above represents the class diagram of the new proposed class to manage subsets of data.

## Package Implementation

The R programming language supports multiple types of object-oriented class systems including S3, S4, R5 and R6 (give comparison). For the implementation of the said package, we use the S4 class system because of the following two reasons:

1. S4 class system is more OOP oriented than other classes, where you can have setter/getter methods with a single generic.
2. Experiment classes that we wish to inherit are also S4-based.

Keeping in view the above arguments, we use the S4 class system to implement our class by inheriting from other Experiment classes.

A brief diagram of implementing our package (and the class) is presented in the figure below.

Graphical user interface

Description automatically generated with medium confidenceA picture containing graphical user interface

Description automatically generated

### Setup R environment

1. R environment for package build is setup using RStudio software and a github repository is linked for version control.
2. R supports multiple system of classes where each serves a specific purpose and needs. A comparison of these classes is provided in the table below.
3. As the requirements of our package is to serve as a drop-in replacement class for other Experiment classes, we use the same class system as these Experiment classes i.e., S4 class system that uses a systematic way of managing setters and getters and allows inclusion of slots.
4. The drop-in replacement feature would allow our package and classes to be used interchangeably with the existing methods that utilize other Experiment classes without having the need to convert between these.
5. As there are 5 Experiment classes namely SummarizedExperiment, RangedSummarizedExperiment, SingleCellExperiment, TreeSummarizedExperiment and SpatialExperiment, we inherit from each of these separately to build new subset classes that we name as SubsetSummarizedExperiment, SubsetRangedSummarizedExperiment, SusbetSingleCellExperiment, SubsetTreeSummarizedExperiment and SubsetSpatialExperiment. All five of these new Subset classes inherit directly from their parent Experiment classes so as to allow these subset classes to work exactly as intended by their original parent Experiment classes as a drop-in replacement with existing methods but adding subset support to these new Subset classes.

. put comparison table of R class systems

### Implement Constructor

1. To enable easy conversion from other Experiment classes, we implement a simple constructor function ExperimentSubset() that takes input an object that belongs to any of the other Experiment classes and converts it into a Subset class appropriate to that object. This allows easy use of the object as a drop-in replacement for other Experiment classes in downstream analysis.
2. The constructor method also supports quick subsetting if the creation of a subset is required at the time of the creation of the object. For example, in case when other Experiment objects are being used in an analysis workflow and subset support is needed, users can immediately convert to a Subset class and create a specified subset from within the constructor method. For this purpose we have optional parameters in the constructor method that allow this creation of a subset.

### Internal class AssaySubset

1. To store subset information in our package we create another internal class called AssaySubset which serves not only for storing subset information but also as a conveninent way to link these subsets to the original parent data and other subsets.
2. The AssaySubset class is also an S4 class but it is not exported which means it cannot be directly used by the users. Instead, the constructor method of this class is called in our wrapper method createSubset() that internally uses this constructor method to initialize an object of class AssaySubset and stores the relevant subset information and characteristics in this AssaySubset object. This AssaySubset object that stores the subset information is saved inside the subset slot of Subset class which uses this slot and the information inside this slot to coordinate the subsets with the parent data.
3. Each AssaySubset object has further multiple slots that store the name of the subset, name of the parent assay, row indices from the parent assay that are part of this subset and column indices from the parent assay that are part of this subset. Additionally, each AssaySubset object stores a empty object of the original class to store non-redundant transformed data matrices within this subset. The row/column indices work similar to pointers and by using the parent assay name these indices are used to reference the redundant data from the original parent assay. This strategy eliminates the need to store any redundant data thus saves considerable amount of memory. The overall structure of this AssaySubset class and how the member slots of this class are used to reference data from parent assays is shown in the figure.

. put diagram too with member variables and their functioning and how row/col indices are used to reference data

. update the main class diagram in methods with assaysubset too

### Implement subset manipulation methods

For the manipulation of subsets, we specifically create and implement the following methods:

1. createSubset() method is implemented for the creation of a subset. The input to this method is an ExperimentSubset object and specific parameters that specify the subset name, row/column indices of the parent assay and the name of the parent assay to which this subset belongs. The result of using this function is the creation of a new AssaySubset class object that stores the information of the new subset which is ultimately stored as an element in the input ExperimentSubset object.
2. getSubsetAssay() and setSubsetAssay() methods are implemented to get or set an assay specifically to a subset. S4 method approach is not used here since both getter and setter methods work differently and require different number of methods.
3. subsetColData() S4 method is implemented that allows setting and getting of colData specifically from a subset.
4. subsetRowData() S4 method is implemented that allows setting and getting of rowData specifically from a subset.
5. susbetColnames() S4 method is implemented that allows setting and getting of colnames specially from a subset.
6. susbetRownames() S4 method is implemented that allows setting and getting of rownames specially from a subset.
7. subsetAssayCount() method returns a total count of the subsets and the internal assays in these subsets.
8. subsetCount() method returns count of just the subsets from the input object.
9. subsetDim() method returns the dimensions of a particular subset.
10. subsetAssayNames() method returns the names of the subsets and the assays inside these subsets collectively.
11. subsetNames() method returns only the names of the subsets stored inside an input ExperimentSubset object.
12. subsetParent() method retrieves a complete subset to parent link of a specified subset. This method is particularly helpful in understanding the provenance of the data, i.e., how this particular subset originated and how it has been transformed over time.
13. subsetSummary() method is similary to the R base summary() method but it specifically provides a summary of an input ExperimentSubset object and the current state of the subsets. Specifically, it describes the size of the overall object, all subsets, their sizes, their internal assays and if any additional reducedDims have been stored in this subsets. Additionally, it shows the subset-parent hierahrical link of each subset retrieved from the subsetParent() method.
14. subsetSpatialData() method from SpatialExperiment class to get or set spatialData to specifically a subset.
15. subsetSpatialCoords() method from SpatialExperiment class to get or set spatialCoords to specifically a subset.
16. subsetRowLinks() method from TreeSummarizedExperiment class to get or set rowLinks to specifically a subset.
17. subsetColLinks() method from TreeSummarizedExperiment class to get or set colLinks to specifically a subset.

. some pic/fig

### Override methods

1. To support subset capability with existing methods from all Experiment classes, we override these methods and add subset support to them made possible by use of direct inheritance from these classes.
2. For methods that exist with multiple classes, we create a single function and then call this function in the generic function of each of the class method to minimize the code redundancy and to increase code reusability.
3. The methods inherited from the Experiment classes to which subset support have been added through the strategy described above are specified below:
   1. show()
      1. The show() is a base R (methods class) function that prints the object summary, visualizes a plot or prints other necessary information which is suitable for that particular class of objects. Here, we override the show() function and mimic the output of other Experiment classes but additionally add printing of subsets and subset assays for easier understanding of the current object structure.
   2. assay()
      1. An S4 method that can set or get an assay (data) depending upon on which side of the assignment operator (<-) it is used. In our package, we override this method to support setting or getting an assay from a subset aswell by using an additional parameter called ‘subsetName’ in which you can specify the name of the subset to which or from which subset you want to set or extract an assay.
   3. rowData()
      1. An S4 method that can set or get rowData (row or genes or features metadata) associated with the current object. In this package, we additionally add support to set or get rowData specifically from a subset by specifying the name of the subset in the additional ‘subsetName’ parameter. This data occasionally includes additional information abou the genes or features in the dataset and is also used to store computations limited to the genes for example, statistics computed for variable genes are stored in the rowData of the object.
   4. colData()
      1. An S4 method that can setor get colData (columns or cells or samples metadata) associated with the current object. In our package, we additionall add support to set or get colData specifically from a subset by specifying the name of the subset in the additional ‘subsetName’ parameter. This data may also include additional information about the samples such as computed clusters or sample information.
   5. metadata()
      1. A general S4 method to store additional metadata about the dataset which cannot be included in the rowData or colData. This slot generally includes information about how the data was extracted by a particular protocol. Here, we allow the users to use the additional ‘subsetName’ parameter to additionally store metadata for just a particular subset of interest.
   6. reducedDim()
      1. The reducedDim S4 method sets or gets the dimensionality reduction results to or from an Experiment object. Additional support for storage and retrieval of dimensionality reduction results for subsets is added through the ‘subsetName’ parameter.
   7. reducedDims()
      1. The reducedDims S4 method sets or gets multiple reducedDims to or from an Experiment object. Additional support for doing so with subsets is added through the `subsetName` parameter.
   8. reducedDimNames()
      1. The reducedDimNames S4 method sets or gets the names of the stored dimensionality reduction results in an Experiment object. By using the ‘subsetName’ parameter, names of the reducedDims can be retrieved or modified from a subset.
   9. altExp()
      1. The altExp S4 method sets or gets additional alternative experiment objects to the parent Experiment object. We have added support to add or remove (or retrieve) alternative experiments to each specific subset by using the ‘subsetName’ parameter.
   10. altExps()
       1. The altExps S4 method sets or gets multiple alternative experiment objects to the parent Experiment object. We have added support for adding, removing or retrieving of multiple alternative experiments from a specific subset by using the ‘subsetName’ parameter.
   11. altExpNames()
       1. The altExpNames S4 method gets or sets names of the stored alternative experiments in an Experiment object. Similarly, names of the stored alternative experiments can be set or get from subsets using the additional ‘subsetName’ parameter.
   12. spatialData()
       1. The spatialData S4 method is explicitly available in the ‘SpatialExperiment’ class to set or get the spatial data from an object. We have added subset support to set or get spatial data to or from a subset using the additional ‘subsetName’ parameter as long as the object is inherited from the parent ‘SpatialExperiment’ class.
   13. spatialCoords()
       1. The spatialCoords getter is a complimentary method to the spatialData method and retrives only the spatial coordinates from the overall spatial data. We have added subset support to this method which allows the users to retrieve the spatial coordinates just for a particular subset by specifying the name of the subset in the additional ‘subsetName’ parameter.
   14. rowLinks()
       1. The rowLinks accessor method retrieves the information of the rows linked with the row tree in the ‘TreeSummarizedExperiment’ class. Here we have added support for subsets using the ‘subsetName’ parameter which lets us retrieve the rowLinks for just a particular subset as long as the object is inherited from the same parent class.
   15. colLinks()
       1. The colLinks accessor method retrieves the information of the columns linked with the col tree in the ‘TreeSummarizedExperiment’ class. Here we have added support for subsets using the ‘subsetName’ parameter which lets us retrieve the colLinks for just a particular subset as long as the object is inherited from the same parent class.

### Document package

R packages are generally documented in two standard ways, one through ‘roxygen2’ package that creates a documentation page for each function or method and secondly through a package vignette which is more of a user guide. Following both of these standards, we used the ‘roxygen2’ R package to create documentation (Rd files) for each of the function/method in our package that describes the function/method name, short description, input/output of the function/method and a runnable example to show the working of the function/method. These ‘.Rd’ files generated by ‘roxygen2’ then become available online repositories for users to view such as the ‘rddr.io’.

Additionally, to document the usage of the package, we used Rmarkdown to create a vignette to serve as a user manual. In the vignette, we describe the overall purpose/motivation of the package, the structure of the package, available functions/methods and a sample toy example. Moreover, we have added a sample workflow using the ‘pbmc’ dataset that illustrates the usefulness of our package. This vignette is available with the package and can either be viewed from the Bioconductor landing page or from within the R console.

### Test package

To ensure the correctness of the our package throughout its development and post-development life-cycle, we make use of unit testing through the R ‘testthat’ package to test our package in small units, particularly each function and method in our package. We particularly use unit-testing because the github travis build allows us to automatically run the unit tests for each commit/update and ensures that with each update none of the functionality breaks. For each function or method that we have either created or overrided from the other Experiment classes, we specify a set of input and outputs and test these at each iteration of the development life-cycle and later with the updates and ensure that the output returned by these functions are exactly the ones that are expected with the functionality.

The ‘testthat’ package allows us to specify all unit tests within a single ‘testthat.R’ file with their corresponding inputs and the expected outputs.

# Chapter 4

# Results and Discussion

To overcome the challenges

. tell about the analysis

. how different objects were used

. tell about results and compare them

. discuss from paper

Validation:

Move this after results (or at the end of results)

. first get results, then validate them using below:

. talk about validating package correctness

. talk about validating package efficiency

. talk about validating package provenance tracking

## Answer to research question

. tell that this is a great alternative to manually using subsets by creating multiple obejcts due to procenanc tracking, efficiency and less code

# Chapter 5

# Conclusion, limitations, and future work

. conclusion from paper (zada kar kay)

. limitation (rownames/colnames are redundant, find more?

. keep update with SPE and coassayexperiment­

# References