A Robust scRNA-seq Data Analysis Pipeline

for Measuring Gene Expression Noise

by

Parithi Balachandran

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved April 2017 by the
Graduate Supervisory Committee:

Xiao Wang, Chair
David Brafman
Thurmon Lockhart

ARIZONA STATE UNIVERSITY

May 2017

ABSTRACT

The past decade has seen a drastic increase in collaboration between Computer Science (CS) and Molecular Biology (MB). Current foci in CS such as deep learning require very large amounts of data, and MB research can often be rapidly advanced by analysis and models from CS. One of the places where CS could aid MB is during analysis of sequences to find binding sites, prediction of folding patterns of proteins.

Maintenance and replication of stem-like cells is possible for long terms as well as differentiation of these cells into various tissue types. These behaviors are possible by controlling the expression of specific genes. These genes then cascade into a network effect by either promoting or repressing downstream gene expression. The expression level of all gene transcripts within a single cell can be analyzed using single cell RNA sequencing (scRNA-seq). A significant portion of noise in scRNA-seq data are results of extrinsic factors and could only be removed by customized scRNA-seq analysis pipeline. scRNA-seq experiments utilize next-gen sequencing to measure genome scale gene expression levels with single cell resolution.

Almost every step during analysis and quantification requires the use of an often empirically determined threshold, which makes quantification of noise less accurate. In addition, each research group often develops their own data analysis pipeline making it impossible to compare data from different groups. To remedy this problem a streamlined and standardized scRNA-seq data analysis and normalization protocol was designed and developed.

After analyzing multiple experiments we identified the possible pipeline stages, and tools needed. Our pipeline is capable of handling data with adapters and barcodes,

which was not the case with pipelines from some experiments. Our pipeline can be used to analyze single experiment scRNA-seq data and also to compare scRNA-seq data across experiments. Various processes like data gathering, file conversion, and data merging were automated in the pipeline. The main focus was to standardize and normalize single-cell RNA-seq data to minimize technical noise introduced by disparate platforms.

ACKNOWLEDGMENTS

I would like to take this opportunity to thank my thesis advisor Dr. Xiao Wang for his continuous support throughout my time at Arizona State University. He guided me in the right direction whenever he thought I needed it. Without his encouragement this thesis would not have been possible.

I would like to acknowledge Dr. David Brafman and Dr. Thurmon Lockhart as the committee members for this thesis, and I am gratefully indebted to them for their very valuable comments on this thesis.

I would like to express my special gratitude and thanks to Christophe for allowing me to be his sidekick in his research. He has helped and guided me throughout the research. I am lucky to find a friend like him. I learned a lot about coding, research, games and chocolates from him.

I would also like to thank David, Hao and Kylie for being supportive and helpful when it needed the most. Thanks to all other members of Xiao Wang Lab for making my time in the lab memorable.

I am also grateful to all my friends and family members especially my Dad: Dr.K.Balachandran, Mom: Poongodi, Cousin: Yazhini and also my close friend: Boobalan who have supported me along the way.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

Computer Science in Molecular Biology

The advancements in Computer Science have provided us better tools and algorithms for analyzing, interpreting and validating data. Nowadays automation and optimization have become a part of the programming environment. Quantum algorithms, deep learning, genetic algorithms are some of the areas where the future is heading.

Most of the analysis can also be performed using microcomputers without many constraints. People have developed and will develop great tools and software packages that serve their purpose but get shelved once the users requirements are met. The biggest downsides with the software available for molecular biology are reproducibility and portability [1]. Over the recent years software have become more language, operating system and hardware specific. It’s hard to overcome these in the fast paced world we are living. Every field is growing at a rapid pace, so it’s up to us to whether use the advancement in other field to our advantage or not.

What Is Data Mining?

Data Mining can be considered as an interdisciplinary field, focused on extracting useful and meaning information from available data. The demand for the understanding about a disease and the biological process related with the disease in order to find a cure along with the advancements in the wireless communication, storage, processing power have lead to massive surge in the biological data being generated via sequencing, imaging, microarrays and more [1]. Biological data mining can be useful in Sequence

Analysis, Text Mining in Biomedicine and Healthcare, Network, 3D Medical Image analysis and in many other interesting and fascinating researches [2].

Most of the scRNA-seq data are publicly available and hosted online, which encourages mining information from them to discover meaningful knowledge. However, in order to extract meaningful knowledge the data should minimize extrinsic and technical noise. Technical noise is of particular concern when integrating data from various data sources; different sequencing platforms, different labs, and different time frames.

Brief Introduction about Molecular Biology

Studying the function and structure of proteins and nucleic acids leads us to a better understanding about life. The Genome constitutes all the genetic material of an organism. The main components of these genetic materials are deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). They are composed by four different monomers called nucleotides. Genes are sequences of nucleotides, typically DNA, that encode the information for the synthesis of a protein or RNA molecule. One of the important properties of DNA is replication, which is useful during cell division. The DNA of all organisms looks the same because they are mostly made up of the same nucleotides. The only way to differentiate the DNA of organisms is by identifying the order in which the nucleotides are arranged.

DNA is transcribed into RNA and then RNA is translated into proteins. Reverse transcription is possible in retrovirus and RNA Replication has also been seen in some viruses. This has led to the modified central dogma of molecular biology (coined by Francis Crick -1958), describes the flow of genetic information (Figure 1) [3]

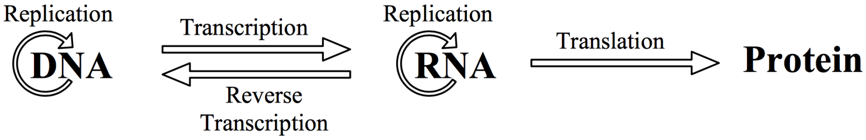


Figure 1: The Modified Centre Dogma of Molecular Biology[3]

What Is Sequencing?

The process of accurately identifying the nucleotides order: A (adenine), C (cytosine), G (guanine), T (thymine) or U (Uracil), in a nucleic acid molecule is known as sequencing. The purpose of sequencing is to unlock information, which is to find the order in which the nucleotides are arranged in the DNA. At present the sequencing is being widely used in the fields of medicine, genetics, environmental science and forensics. In medicinal field it can be used to provide more customized drug for patients or provide customized treatment/therapies based on their DNA profile. In genetics on of use for sequencing is to identify the genes in the sequence responsible for production of proteins and to facilitate the manipulation of those genes. In environmental science it's useful in analyzing bacteria and virus from soil, air and water and estimate the degree of damage caused by pollution [4,5]. In forensics, once the suspects DNA sample is obtained it can be sequenced and with the information acquired a model of a suspect can be built with features like the eye/hair/skin color, height [4, 5, 6]. It's also useful in trying to compare similarities between two samples DNA.

In the past decade sequencing has seen an astonishing growth from trying to identify the nucleotides of a single gene to whole genome sequencing. It has also evolved from being an expensive and time-consuming process into a rapid and broadly available process. In 1953, James Watson and Francis Crick were able to identify the molecular structure of the DNA with help of the crystallographic data from Rosalind Franklin and Maurice Wilkins [6,7]. They visualized the 3-D structure of the DNA but were not able to read the contents of the DNA. Initial stages of sequencing were done using the single stranded transfer RNA. The first protein-coding sequence (coat protein of bacteriophage MS2) was produced in Walter Fiers' Laboratory using 2-D fractionation [9].

First Generation Sequencing

Sequencing was initially performed over microbial mRNA and tRNA because they were short and single stranded and they could be produced in bulk quantity. Large oligonucleotides were partially degraded using snake venom phosphodiesterase and terminal nucleosides were noted from the resulting product [8]. This was the first technique used in finding an order of oligonucleotides in a sequence. The era of sequencing started with two influential protocols in 1975, the "plus and minus" system of Sanger and Alan Coulson and "chemical cleavage sequencing" technique by Allen Maxam and Walter Gilbert. Later in 1977 Sanger developed the much-improved version of sequencing known as the "chain-termination" technique [9].

Maxam and Gilbert's chemical sequencing method [9]: Fragments are obtained by chemically treating radiolabelled DNA. These chemical treatments break the chain at specific bases. For example Hydrazine removes bases from pyrimidine but in presence of high salt concentration it can only remove cytosine. The Fragments are visualized through electrophoresis on a polyacrylamide gel. After visualizing, these methods need one additional step to determine the sequence due to the presence of dual bands in certain lanes. 4 lanes used in this method are G, A+G, C and C+T. One or two bands can be present for each position.

Sanger's chain-termination sequencing method [9]: Fragments are obtained by mixing radiolabelled Dideoxyribonucleotides (ddNTPs) into the DNA polymerization regular. These ddNTPs prevent further addition of deoxyribonucleotides (dNTPs) and cause sequence elongation to terminate. This step is carried out for each of the nucleotide bases. Electrophoresis was used to visualize the lanes. Unlike in the previous method only one band would be appear in each position. After few years the radiolabelling was replaced with fluorescent tags, this improvement helped in designing automated sequencing machines.

First generation sequencing machines were able to produce reads under one kilobase (kb) in length. Longer lengths were obtained by using shotgun sequencing and with the advancements in extraction methods in order to obtain high quality sequence for sequencing.

Microarray Technique

Microarray can be used to measure the differences in mRNA expression [10,11]. mRNA is removed from the organism and placed in a test tube containing more mRNA. Fluorescent tags are added into the test tube and they get attached to the mRNA. The microarray contains thousands of spot with each spot representing a particular gene of the organism. The tagged mRNA from the test tube is then added to the microarray. Each tagged mRNA stick to only one complimentary DNA in the microarray. The fluorescent intensity is recorded across the microarray. This helps researchers in identify when genes expressed more and when ones expressed less.

Second Generation Sequencing

Pyrosequencing became the first commercial Next Generation Sequencing (NGS) technology to be successful. Pyrosequencing was about measuring pyrophosphate synthesis by luminescent methods and amplification was done in parallel [13]. They also avoided the use of modified ddNTPs and made observation possible in real time. The bridge amplification technique became famous as it enabled production of clusters of identical molecules. These clusters are visualized with help of fluorescence-based reversible-terminator dNTPs. The workflow of Illumina, Ion Torrent and SOLiD sequencing techniques are shown in Figure 2 [12].

Illumina Sequencing [12]: The prepared RNA-seq library gets attached to the flow with the help of adapters. Adapters are added during the library preparation and its compliment version is found along the flow cell, which causes the sequence to attach along the flow cell. A complimentary version of this cDNA is created by the DNA polymerase, which then undergoes several rounds of PCR amplification. Quality control of library is critical as it aids in optimizing the cluster density and it reduces over-clustering. Fluorescence-based dNTPS are used during the synthesis process. Multiple cycles are run and only one nucleotide gets attached during each cycle (Figure 2a). This helps us in capturing an image of which base gets attached during each cycle and reduced the error rate (< .2%)

Ion Torrent Technique [12]: Emulsion based PCR is used during library preparation. Then beads with complimentary oligonucleotides are attached to each fragment and are then amplified. They are then placed on a semiconductor chip, which detects the change in pH level caused by the release of protons (H + ions) during polymerization (Figure 2b). These chips use complementary metal-oxide-semiconductor (CMOS) technology. Since there is no need for any optical detection Ion Torrent sequences much faster than Illumina. Monomers are detected based on number of H + ions released but it gets tougher to detect homopolymers of length greater than 7.

Sequencing by Oligonucleotide Ligation and Detection (SOLiD) system [12]: Libraries are prepared in the same way as of Ion Torrent sequencing. Sequencing is done by ligation. 16 fluorescently labeled 8-mer oligonucleotides are used. 1-2 position in the 8-mers is represented by all combination of 2-mer and the last 3 positions are filled with 4 different colored fluorescent labels. The middle parts of the kmer are unknown. Since 5mers get attached after each cycle the length of the sequence attached gets long and also we have fluorescent measurement for every 5th base, so a reset is done after every 5-7 cycles. This helps us to decode using the two base color encoding (Figure 2c). Since each nucleotide is read twice and this reduces the error rate much further (<0.1%) but produces shorter reads.

The biggest advantage of second generation sequencers are that these processes can be run in parallel, meaning that despite lower read lengths much higher throughput can be achieved. Micro-fabrication and high-resolution imaging also played a big part in the success of second-generation sequencers.

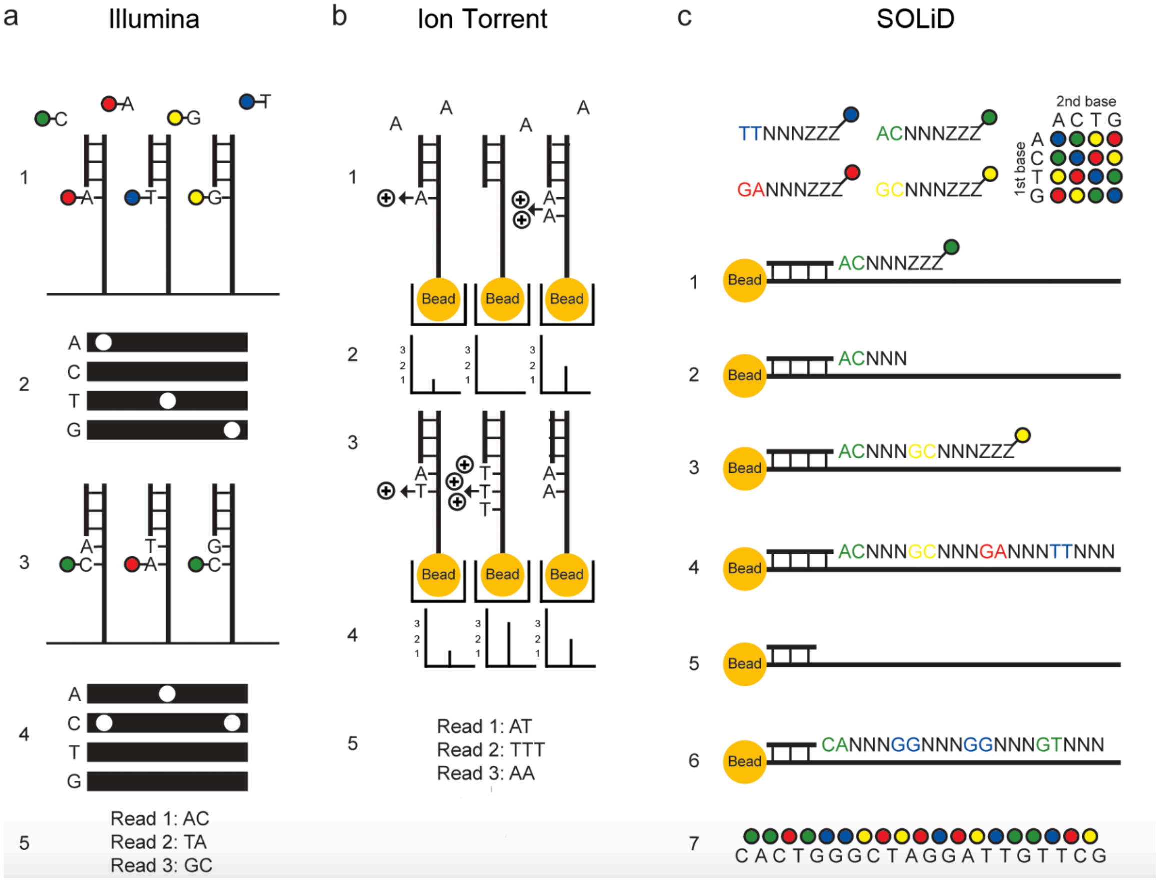


Figure 2: Second Generation Sequencing Techniques a. Illumina detects using fluorescence-based reversible terminator dNTPs. b. Ion Torrent detects by the release of hydrogen ions (⊕). c. SOLiD sequence detects by ligating fluorescently labeled oligonucleotides[12].

Third Generation Sequencing

The primary focus of third generation sequencing is sequencing single molecule and neglect the need for DNA amplification. Zero-mode waveguides (ZMW) and Nanopore sequencing are the two successful techniques in this generation with the later most attention due its portable nature [12]. Third generation sequencing techniques are capable of producing long reads (upwards of 880kb [13]). ZMW has a DNA polymerase attached at the bottom of the device. Then the target sequence is added along with the fluorescent nucleotides. Nucleotides are added to during the DNA polymerization at the base of the ZMW, which emit a short bust of light through its narrow ZMWs providing real-time sequencing. Pacific Biosciences develop single molecule real time (SMRT) platforms.

Nanopore sequencing is a concept older than Sanger's method but was not technologically viable until recently, and was brought back into the field of sequencing by Oxford Nanopore Technology (ONT). Voltage is applied to a nanopore embedded in a synthetic membrane through which a denatured single strand of DNA is passed. Each nucleotide base has its unique way of altering the ion flow, reducing the current measured at the nanopore for a length of the time. Advantages of Nanopore sequencing are that it requires very minimal amount of samples and that leads to less time being spent on library preparation. Once the required resources are made ready all that's left is a nanopore channel, steady flow of current and way to measure the fluctuation in current. These requirements were built into a small device (ONT's MinION) containing a flow cell, which is powered via USB 3.0 cable [14]. Multiple experiments can be run in sequence in a single flow cell in a single device. The Portable nature of these nanopore sequencers has drastically changed the way sequencers are being used, as sequencing can now be done in the lab or home or even in the field. This has eliminated the need for a lab infrastructure to sequence data and has minimized the human work involved to get the sample sequenced.

Data Storage for Sequencing Data

The three members of the International Nucleotide Sequence Database Collaboration (INSDC): GenBank (Bethesda, Maryland USA), European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-Bank in Hinxton, UK), and the DNA Data Bank of Japan (Mishima, Japan) together hold close to 2000 gigabases of sequence data including data from Whole Genome Sequencing (WGS)[15]. Each experiment measures gene activity levels or genome sequence, and could represent data from a single cell or a population. Unrestricted access is provided to these data, which helps scientists all over the globe to study and compare data. This allows collaborative research to flourish.

GenBank is maintained by the National Center for Biotechnology Information (NCBI), a part of the National Library of Medicine, National Institutes of Health. Figure 3[3,15] shows the surge in the number of sequences being stored under GeneBank since 1982. Around 3 million new sequences are being added every month to the database. It is amazing how the INSDC has led to the open exchange of biological information.

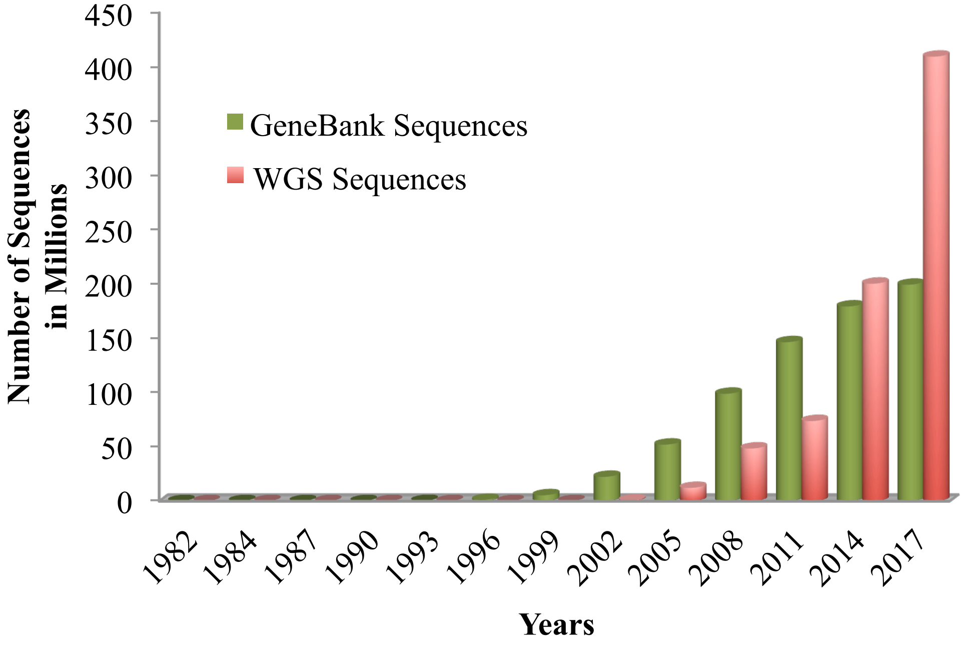


Figure 3: Surge in the Total Number of Sequences in Genebank[15]

CHAPTER 2

RNA SEQUENCING

What Is a Transcriptome?

During the process of transcription gene readouts are transcribed into the RNA. These gene readouts are called transcripts and the collection of all possible gene transcripts are called as transcriptome. In many ways the transcriptome is a subset of the genome, however RNA post-processing like alternative splicing mean that the transcriptome contains some sequences not present in the genome [16].

How Important Can Analyzing Transcriptome Be?

Analyzing the collection of RNA sequences in a cell it is possible to determine when and where a gene gets turned on or off. Gene activity can be measured with the help of number of transcripts present. This is also called as gene expression. Analyzing transcriptome will help us gain insight on how a particular type of cell functions, and how change in normal level of gene activity can relate to disease. They also help in finding which genes are active in a cell at a given time [16]. One of the important things to consider while analyzing transcriptome is the noise present because ignoring noise can give raise to false positives. Some of the genes functions are still unknown. If a gene is expressed more in cancer cells than normal cells, then there is possibility that the gene can be related to cancer.

Human biology can be studied effectively using rat and mouse models. Apart from physical and environment advantage of using mice, they have long been used as an animal model for understanding of cancer and other disease. They are also used in therapeutics experiments to check the safety of drugs [17]. NIH-supported initiative, The Mouse Transcriptome Project [18] generated a free, public database of gene transcripts for many mouse tissues. The main goal for this project was to gather information and build a database of expression of every gene. This would be help is minimizing duplication and speed up the process of extracting meaningful genomic information from sequences.

Why RNA Sequencing?

RNA Sequencing (RNA-seq) is considered as highly accurate and sensitive tool for measuring expression across the transcriptome. Expressions can even be measured for cells that undergo certain therapy. The beauty about RNA-seq when compared to previous techniques like micro-array is that it captures both known and unknown information. RNA-seq can be performed for organisms without reference and lower cost compared to other methods in measuring gene expression [19].

Why Now?

Earlier when Sanger sequencing was used for sequencing of cDNA the results had low throughput, high cost and was less qualitative. Serial analysis of gene expression (SAGE), cap analysis of gene expression (CAGE) and massively parallel signature sequencing (MPSS) were some of the tag-based approach used to get precise and high throughput gene expression levels [19]. Since all used Sanger method they were not able to completely analyze the transcripts and isoforms. Second and Third generation sequencing methods have helped overcome most of the above said limitations. These developments in high-throughput sequencing have lead to improve the process of mapping and quantifying transcriptomes.

What Is RNA-seq?

High-throughput sequencer (Illumina IG, Roche's 454 Pyrosequencing and Thermo Fisher's SOLiD and Ion Torrent) is used to obtain short sequences and then they are either aligned to the reference genome or assembled de novo without genomic sequence. Expression levels of each gene can then be approximated based on the sequenced population [12].

Advantages of Using RNA-seq

When comparing with other transcriptomes methods (Tiling microarray, EST sequencing) RNA sequencing has advantages such as single base resolution, high throughput, relatively low background noise, ability to distinguish different isoforms, low quantity of RNA required for sequencing because no cloning is needed and, it has relatively low cost for mapping transcriptome of large genomes. Additionally, when compared with microarray it allows a researcher to examine the transcripts that are present, rather than checking if specific transcripts are present.

Single Cell Sequencing:

A single cell contains a vast amount of coding information that can be extracted using deep sequencing. Analyzing single cell sequences have provided us with information that cell populations are made up of many individual cells with heterogeneous cell states capable of producing system level functions [20,21]. Cell states help us understand about cell function and dysfunction. Degeneration of certain cell around normal cells and how cells respond to drugs can be studied with single cell analysis.

The standard workflow for single cell RNA sequencing (scRNA-seq), shown in Figure 4 [22], involves 4 major stages. First stage is to isolate the single cell for the tissue, then during the second stage RNA is extracted and convert that into a cDNA (reverse transcription), third stage would be to amplify the cDNAs and the fourth stage is to generate the library and sequence it. During these stages in the scRNA-seq workflow noises and bias get introduced.

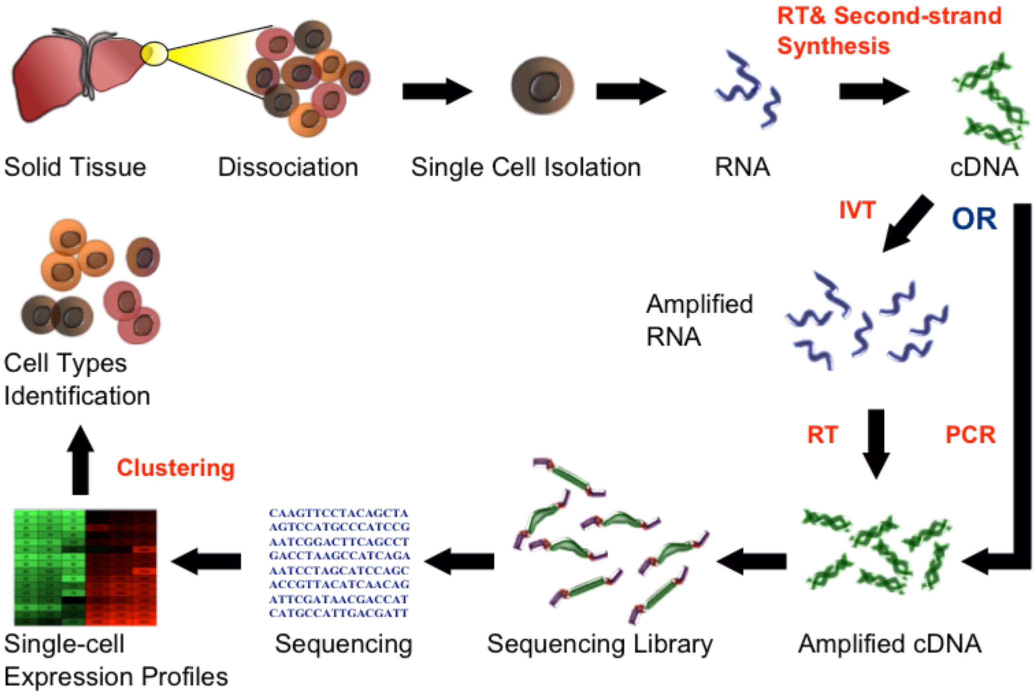


Figure 4: scRNA-seq Workflow[22]

CHAPTER 3

ANALYSIS AND DESIGN

Analysis

Design of the pipeline must be simple and similar to the standard experimental design of any scRNA-seq experiments. The final pipeline was prepared after studying multiple experiments from NCBI, and analyzing the presence of absence of various steps. But in order to test the functionality of this pipeline, data was necessary. All data used were obtained from experiments that used second-generation sequencing. The main reasons for going with second-generation sequencing data was because of their high quality and large volume. The sequencing data for each experiment considered were produced from Illumina sequencers.

Through analysis of experiments present in the publicly available NCBI GEO repository it can be noted that experiments were being published with single-ended 25 base pair reads, while the most prolific sequencers can offer paired-end 100 base pair reads, and newer second-generation sequencers like Illumina's MiSeq offers double-ended 300 base pair reads [23, 24]. While this development increases data quality, it does mean that assumptions made by existing sequence alignment pipelines might not be valid for the newest sets of data or sequencers. For example, most aligners dealing with paired-end reads include parameters for the minimum and maximum gap between paired ends. These parameters can increase the mapping quality by ensuring that the gap between the paired ends is a length reasonable for an mRNA fragment. One major problem arising with sequencing and alignment is that mRNA transcripts samples are preprocessed (typically through sonication/shearing or a Nextera transposon library-prep), meaning that fragments have an average size set external to the sequencer itself, and with a distribution heavily dependent on the techniques used. This becomes problematic if, for example, one has sequence fragments for a paired-end 100 base run (~250 bp) but instead the data is sequenced on a machine giving 200 bp reads. In this case the inter-read spacing becomes negative, and for some reads (that fall below 200bp) collected reads will include both the original transcript, and the adapter sequence at a high confidence score, such that it cannot be easily removed solely through quality trimming [23]. While there are efforts to counter this through selection for longer fragments, over 30% of reads containing a copy of the adapter sequence (GSE52583 [44])

Data Retrieval

All data used for this thesis have been retrieved from NCBI. Gene Expression Omnibus (GEO) is a public repository that contains genomic data submitted by researchers. The Entrez Programming Utilities (eUtils) is a program, which acts as an interface into the database available at NCBI [25]. The way eUtils works is that it uses fixed URL syntax to search and retrieve data from the database. Data retrieval can be streamlined into a pipeline with the help of eUtils components and languages that can send a URL to the eUtils server and interpret the XML response. Python has been used for the mentioned purposes.

Unique record identifier (UID) acts as a primary key for the database in NCBI. Each record is assigned with an UID. Entrez GEO DataSets (gds) are interfaced using an GEO accession number with a prefix GSE. GEO accession number along with the prefix acts as the UID while retrieving a sample or the entire study. There are many API endpoints in eUtils to retrieve data from the database. All the URLs constructed for data retrieval need to be in the format of the endpoint used.

Each API endpoint retrieves different kind of information. ESummary, ESearch and EFetch are the most commonly used eUtils pipelines. Queries are constructed with fields based on the data needed. Various fields are available which helps us in narrowing our search.

Sample Query:

- Base Version: This was used in the NCBI's gds webpage.

  "single cell"[All Fields] AND ("rna"[MeSH Terms] OR RNA[All Fields]) AND seq[All Fields]) AND "Mus musculus"[porgn]

- eUtils Version: This was the URL to retrieve the data.

  http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=gds&term=single+cell%5BAll%20fields%5D+AND+RNA%5BAll%20Fields%5D+seq%5BAll%20Fields%5D+AND+Mus+musculus%5Borgn%5D&retmax=1000&usehistory=y

Once the query is run, all the experiments and their summary (Title, Organism, Description, ID, etc.) are returned and can be analyzed. Once the desired studies have been identified their ID can be obtained and the download process can be started. The pseudo code of the program designed and developed for this thesis is outlined below.

1. Create an Excel document containing the Column (IDs, Size and Status: Finished/Null/Incorrect).
2. Enter the IDs of the experiments whose data needs to be downloaded into the excel document.

(Step 1 and Step 2 are manual process but can be automated if needed)

3. Start the python script

   3.1. Load the excel and check the Status of each experiment.

      3.1.1. If Status = Finished, move to next experiment.

      3.1.2. If Status = Incorrect, remove the experiment ID.

      3.1.3. If Status = Null, Check the validity of the ID.

      3.1.4. If valid proceed to Step 3.2 else change the status to Incorrect.

   3.2. Use EFetch to retrieve the NCBI internal ftp path.

   3.3. Login to the server

   3.4. Get the list of files available

   3.5. Calculate the size of the entire experiment

   3.6. If no files available down the data directly using URL obtained from EFetch

   3.7. Navigate into a sample folder and download the sequence read archive (sra) file.

   3.8. Repeat Step 3.6 till all samples are downloaded.

   3.9. Compare the total size of the downloaded file with the size from Step 3.5. If equal, update the status else check the size of each sample sra file downloaded with the same file in the server and re-download the ones that don't match.

4. Exit

Design

Any scRNA-seq data analysis pipelines consist of multiple steps, all with specific functions (Figure. 4). After raw reads have been received from the sequencer, there is usually some cleanup to be done. Specifically, the adapter sequences that allow fragments to be read are often included in the reads themselves, which should be trimmed to maximize the chance of alignment as most aligners only allow for a very small number of errors. Next, sequencers cannot be certain about which bases are present in a sequence and instead it provides its best guess, with a Phred (probability) score. In long sequences, and with older sequencers there are often many bases that have a low Phred (<99% confidence) and as such they are often removed. After the quality control checks the data can be aligned to a reference genome or transcriptome, allowing one to get a rough idea of the activity levels of various genes. Following alignment, there is a required quantification step to determine actual gene activity levels. This step typically accounts for factors such as gene length (long genes at the same activity level will have more fragments sequenced), multi-matched reads (either split evenly between reads, or more commonly distributed according to an EM-maximization), and scaling expression levels out of a raw count and into a ratio measurement. Because of the large number of steps involved there are multiple ways to take the same input data and produce differing quantifications.

The flowchart representation of the design is shown in Figure 5. Each stage is crucial and can introduce bias if not handled properly.

- Sequence: The output from the sequencer. For our pipeline, this is considered as the raw reads.
- QC-check: This step is used to identify low quality data and also to trim data with low confidence score. Some experiments might contain Spike-in (used for assessing the quality and success of your process) that can be removed in this step. Examples: FastQC [26], PRINSEQ [27]
- Adapter Trim: This step is used to remove/trim the adapter sequence if required. Examples: btrim [28], Flexbar [29], Sickle [30]

- Reference Genome: A reference genome is a representation of the genes or transcripts of an organism. Aligning to different genome can affect the outcome of an experiment to certain extent [31]. The mm9 and mm10 are two versions of the mouse genome (Mus musculus) popularly used. Genome Reference Consortium (GRC) is constantly updating mm10 [32]. Total number of bases in last version of mm10 is provided in Table 1.

Table 1: Number of Bases in Mus musculus Reference Genome

	mm9	**mm10**
Total Bases	2,745,142,291	2,807,715,301
Total Non-N Bases	2,648,522,751	2,728,358,445

- Annotation: Process of identifying coding and non-coding part of the sequence and further labeling the coding part with the gene it corresponds to.
- Indexing: An index is created for the reference genome. Indexing reduces the time taken for searching in the reference genome. It is similar to Appendix/Index in a textbook. Examples: Sailfish [33], bowtie [34], STAR [35]
- Alignment: The goal of this step is to find the position of the pattern (short-end) in a large text (genome). Aligning reads are equivalent to string matching. Examples: Tophat [36], STAR [35], HISAT [37], Sailfish [33], BWT [38]
- Expression: Calculates the amount of gene expressed in each sample. Reads Per Kilobase Million (RPKM) and Transcripts Per Million (TPM) were two commonly used units. The main difference between these two are the way the values are normalized. TPM normalized for the length of the gene and then for

sequencing depth where as RPKM did the inverse [39]. Examples: Cufflinks [40], RSEM [41]

In this thesis the following were used, mm10 [32], STAR (indexing) [35], Flexbar (trimming) [29], STAR (Aligning) [35], Samtools (for file conversion) [42] and Cufflinks (Expression) [40].

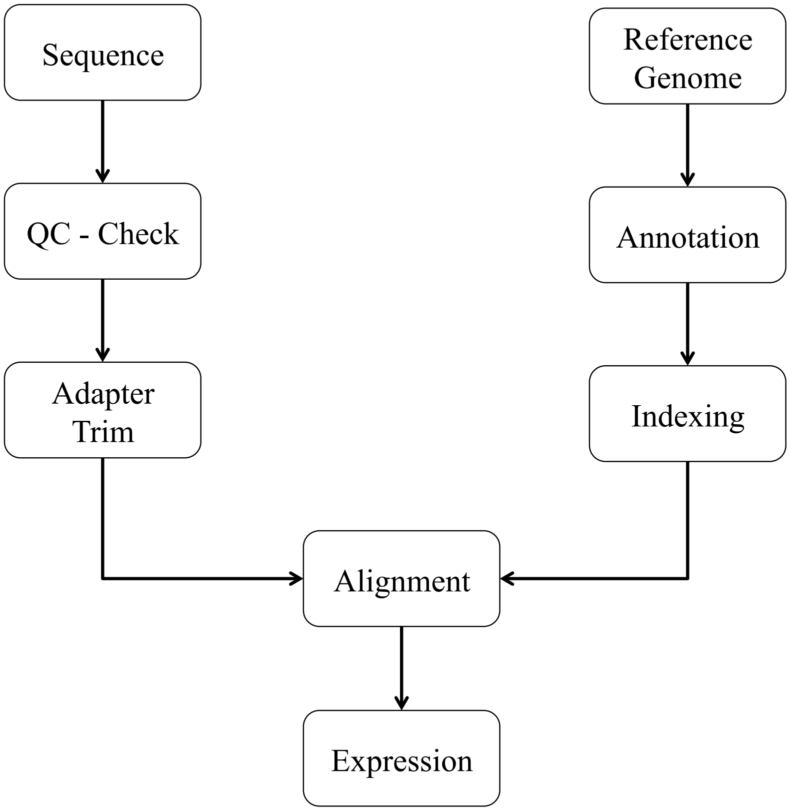


Figure 5: General Flowchart Representation of Sequencing Pipeline

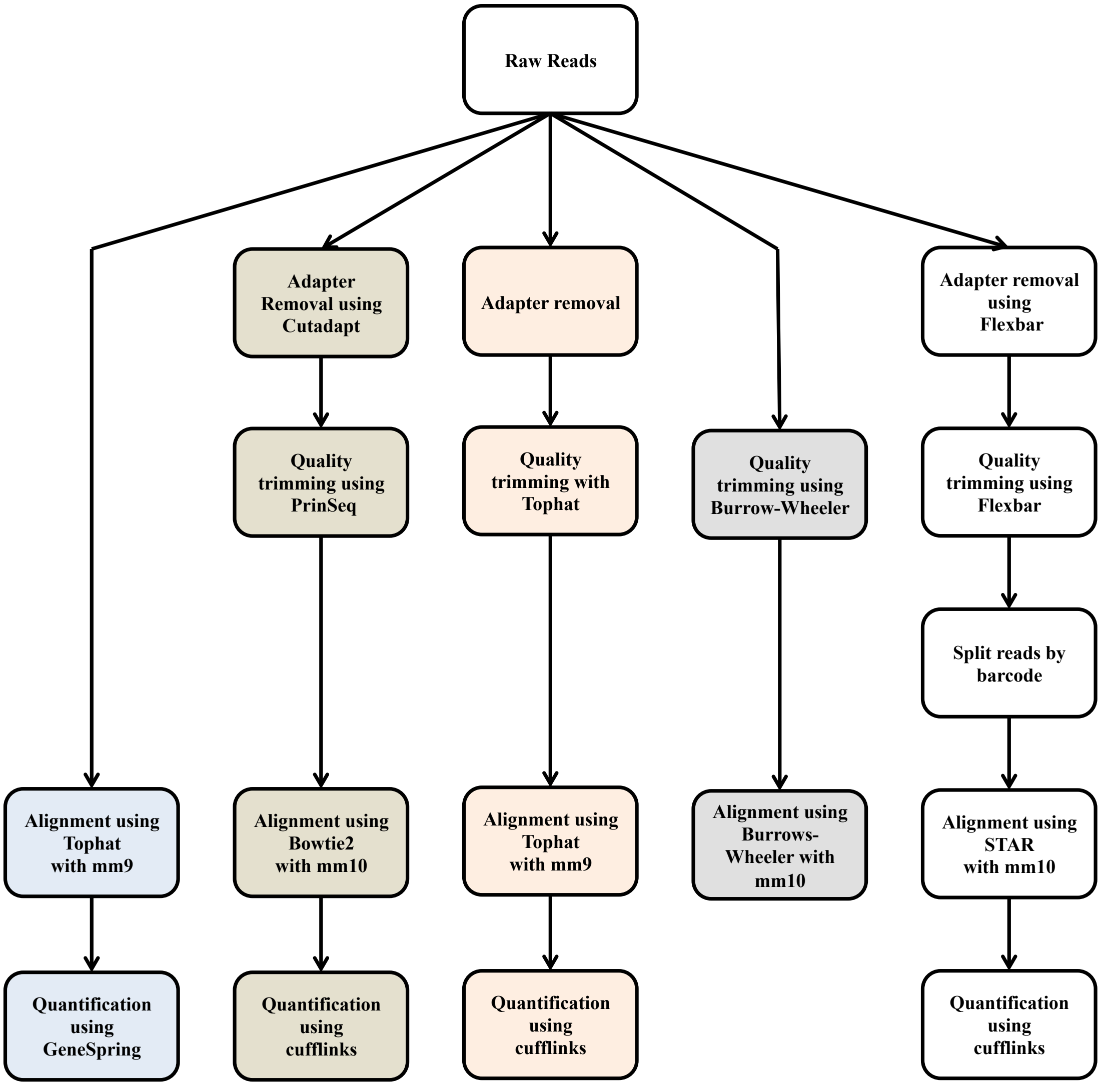
CHAPTER 4

RESULTS AND CONCLUSION

Figure 6 summarizes five different scRNA-seq analysis pipelines. Most labs follow similar pipelines, although often with different tools to accomplish the tasks required. The rightmost pipeline in Figure 6 is our customized pipeline. The other pipelines mentioned in Figure 6 are (leftmost to right) GSE59127 and GSE59129 [43], GSE52583 [44], GSE47835 [45] and GSE55291 [46] respectively. Table 2 was constructed as to infer type of tools, number of sample, experiment size and reference genome used across all the experiment used in this thesis.

Table 2: Comparison of Experiments

Experiment ID	Sample Count (Size in GB)	Cell Types	Reference Genome	Library Prep	Aligner/ Quantifier
GSE47835	103(71.1)	Embryonic Stem cells	mm10	Illumina	Burrows – Wheeler
GSE55291	95(69.1)	Induced pluripotent stem cells	mm9	NEBNext	TopHat Bowtie cufflink
GSE52583	201(65.8)	Lung	mm10	Nextera	Bowtie2 Tophat cufflinks
GSE59127	86(12.3)	Kidney	mm9	Nextera	TopHat GeneSpring 12.6
GSE59129	49(6.2)	Kidney	mm9	Nextera	TopHat GeneSpring 12.6



Raw Reads
Adapter Removal using Cutadapt
Adapter removal
Adapter removal using Flexbar
Quality trimming using PrinSeq
Quality trimming with Tophat
Quality trimming using Burrow-Wheeler
Quality trimming using Flexbar
Split reads by barcode
Alignment using Tophat with mm9
Alignment using Bowtie2 with mm10
Alignment using Tophat with mm9
Alignment using Burrows-Wheeler with mm10
Alignment using STAR with mm10
Quantification using GeneSpring
Quantification using cufflinks
Quantification using cufflinks
Quantification using cufflinks

Figure 6: Comparison of Pipelines

After the pipeline was developed we validated it by running the raw data obtained from the experiments mentioned in Table 2. The metric used for the process of validation in this thesis is Coefficient of Variation (CV). CV is a standard metric that is used to describe the amount of variability among similar experiments where Standard Deviation

(SD) fails to identify the variation precisely [47]. CV is calculated as the ration of SD to mean for the given set of data.

After the raw data was run through the pipeline, the estimates of gene activity levels are obtained in FPKM. In Figure 7 the CVs of genes from authors data was grouped along with the CVs from our pipeline in form of histograms. Histograms in Figure 7 were created with bin count as 50. The Figure 7 clearly shows that our pipeline yield similar results to that of the authors. The note worthy point is that we used the same pipeline for data from different cell types and were able to get results in agreement with the authors. Experiments GSE52583[44] and GSE59127 [43] were used for the Figure 7.

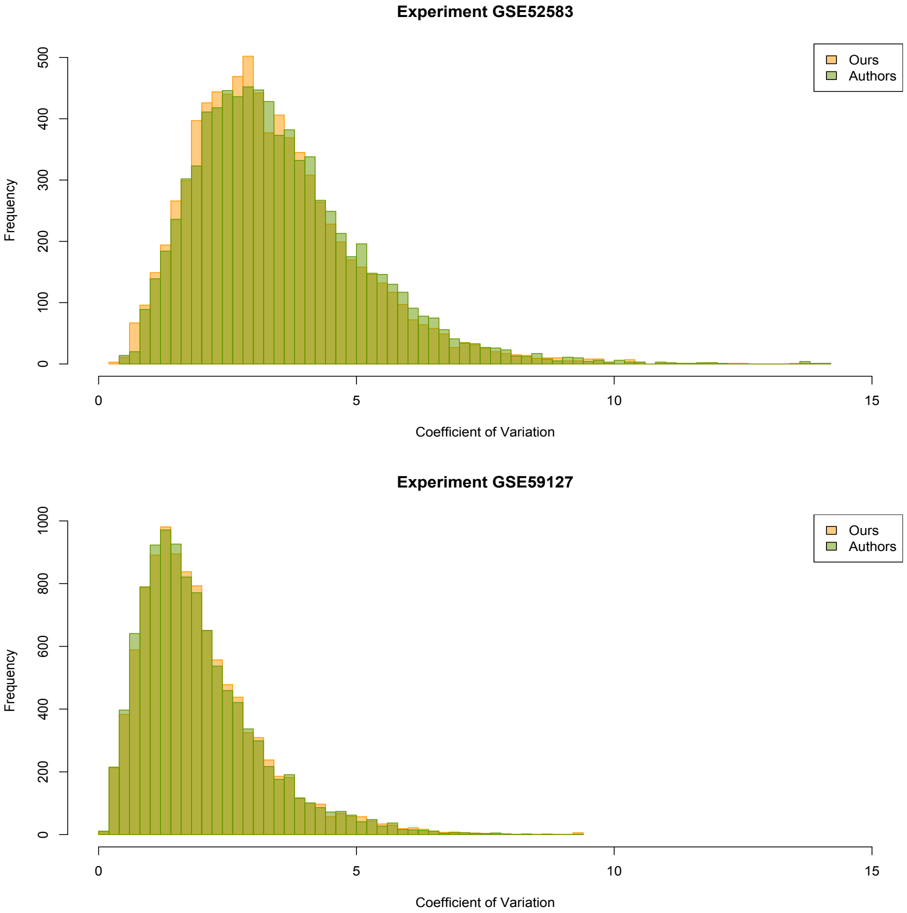


Figure 7: Comparison of Coefficient of Variation of Genes

Aggregating the data from all the experiments mentioned in Table 2 and then running through our pipeline generated the Figure 8. Our pipeline performs better when the data obtained are from different cell types as there is clear increase in the frequency of genes with low CV.

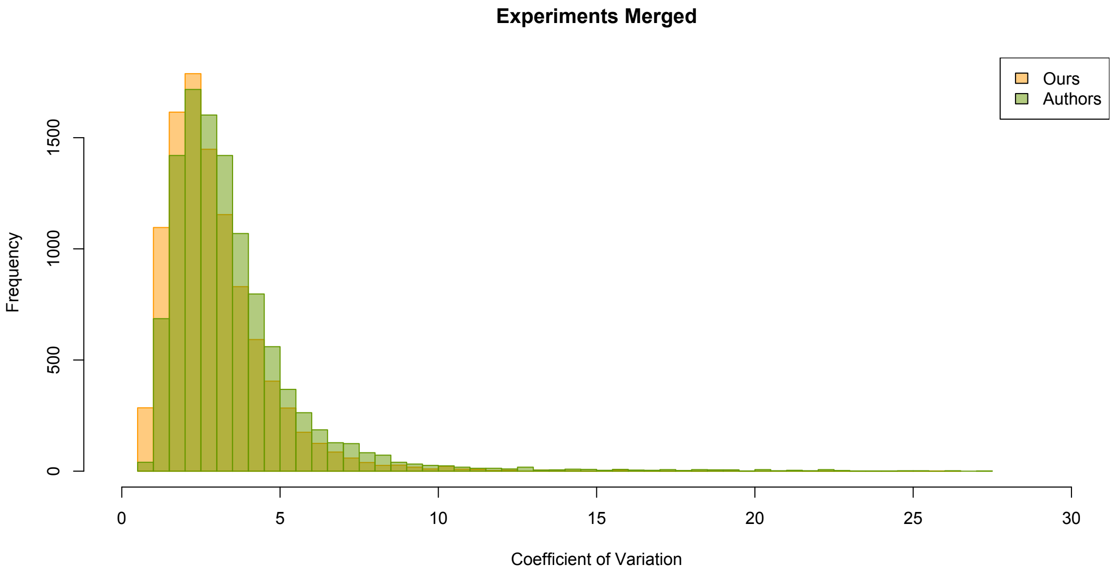


Figure 8 : Coefficient of Variation of Genes from All Experiments Merged

We looked at the CVs of few housekeeping genes [48] and found that we had a similar result as the authors. Figure 9 shows the side-by-side comparison of CVs of the housekeeping genes in each experiment. This gives us a platform for the future to study in detailed about the genes that got expressed higher or lower than the actual experiments. When this pattern is observed among most of the experiments analyzed it intrigue us even further. When comparing the expression data from experiments that used different reference genome version, ours pipeline was able to capture the difference.

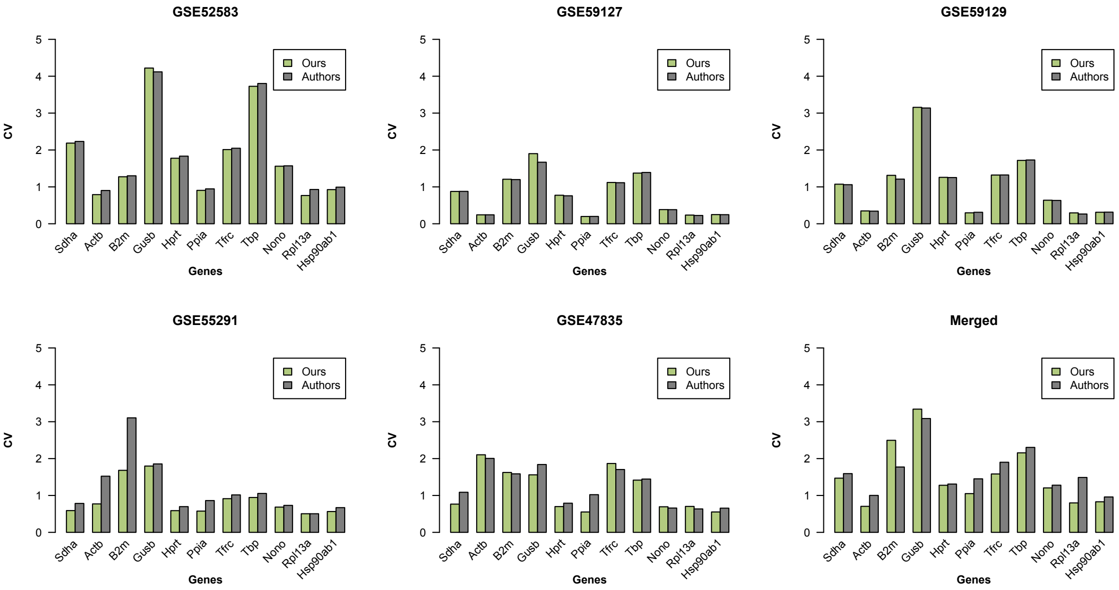


Figure 9: Comparison of Coefficient of Variations among Housekeeping Genes

To further strengthen our validation we generated the scatterplots for all 5 experiments. Figure 10 shows the CV of each gene calculated with our pipeline when compared with the original authors pipelines on their data sets. It can be seen (panel A, B, C, D and E) that in general our pipeline produces comparable CV as in the original paper, validating our modified protocol.

Scatterplots in Figure 10 shows the ratio of CV's, with the x=y line drawn. Points on the right side (x>y) are drawn in red, and opposite in black. For the experiments in (A-E), raw scRNA-seq reads are provided by NCBI's GEO service and analyzed using our customized pipeline (Figure 6). In addition to raw reads, each author provides estimates of gene activity levels through FPKM or TPM. In subplots (A-E) we plot the coefficient of variation (CV) of our pipeline results against those of the original authors. In (F) we group the experiments from (A-E). Additionally, when all experiments are pooled we found that our pipeline leads to a 4% reduction in mean CV and a reduction in CV for over half of the genes (F). One probable hypothesis is that the variation removed is caused by differing biases imposed by distinct pipelines, and such a reduction in variance could provide a significant benefit to researchers hoping to focus on inherent gene expression noise.

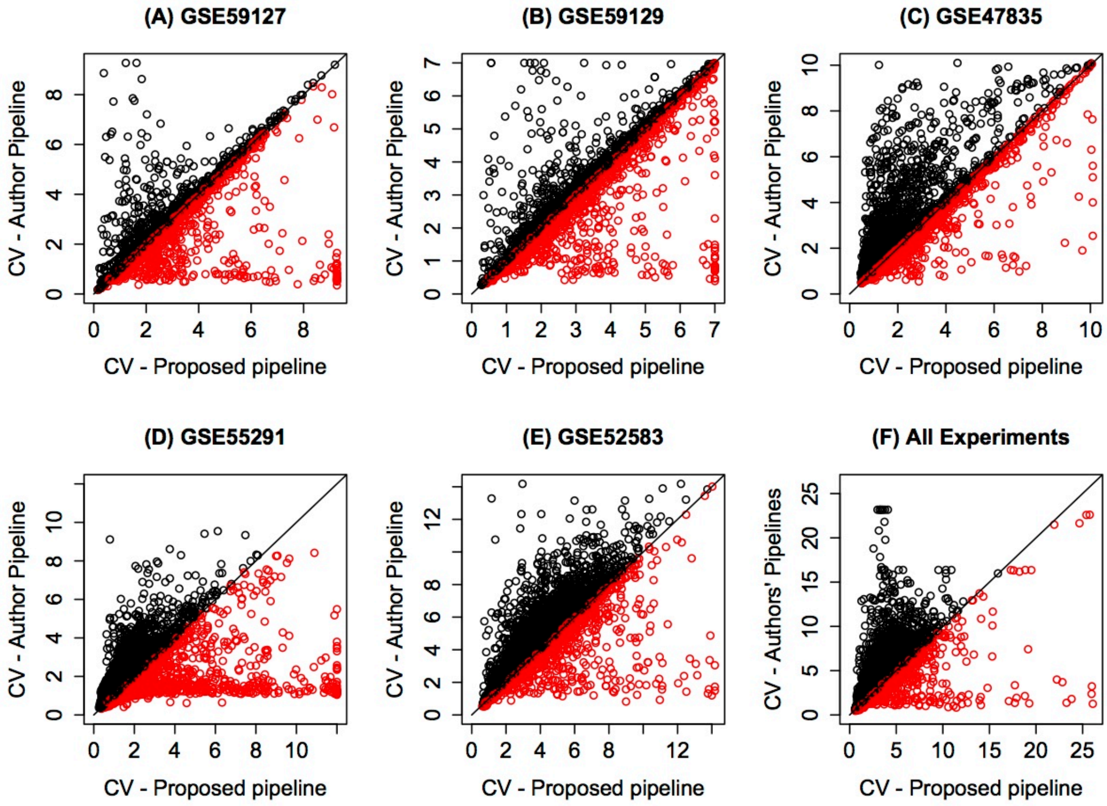


Figure 10: Ratio of Coefficient of Variation Visualized Using Scatterplot

To decipher how a transcriptional landscape regulates cell fate determination via integration of stochastic signals from various inputs, understanding the network architecture that makes up the transcriptional landscape and dissecting core regulatory networks within it is essential. Using high quality scRNA-seq data produced from our customized protocols, it is possible to employ newly developed noise-based methods to decipher co-regulated gene modules where candidate genes show unique expression stochasticity during cell reprogramming will be dissected for further modeling analyses.

For future work two components can be added to this pipeline: parameter optimization and post-processing of data. Manual effort in identifying maximum and minimum fragment length, adapter sequence, reference genome and some cases even the format of the read quality score can be automated through rule-based and distribution based approach. Post-processing can be performed on the data to remove contextual noise. This can also help in classify noise sources and validate them.

The purpose of these type of studies is to maximize the knowledge that can be extracted from vast volume of scRNA-seq data to identify patterns in gene expression noise that can then be used to uncover GRN responsible for random cell fate determination. This study can be considered as a stepping stone for developing methods in future to reconstruct gene regulatory networks by fully utilize each cell's gene expression profile and exhaustively analyze and characterize network motifs as ciphers to identify critical topologies to elucidate their integrative and coordinative regulation in transforming stochastic signals into stable decision for cell fate determination.

REFERENCES

[1] Lesk, Arthur M. Computational Molecular Biology: Sources and Methods for Sequence Analysis. Oxford: OUP, 1988. Print.

[2] “Biological Data Mining and Its Applications in Healthcare.” World Scientific Publishing Company. N.p., n.d. Web. 23 Mar. 2017.

[3] Tzanis, George, Christos Berberidis, and Ioannis Vlahavas. "Biological Data Mining." (2008). Print.

[4] “DNA sequencing uses”, http://www.dnasequencing.com/dna-sequencing-uses.html” N.p., n.d. Web. 31 Mar. 2017

[5] “DNA Sequencing | Summary.” N.p., n.d. Web. 31 Mar. 2017.

[6] Wikipedia contributors. "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 28 Mar. 2017. Web. 31 Mar. 2017.

[7] Alphey, Luke. Dna Sequencing from Experimental Methods to Bioinformatics. Oxford: Bios Scientific Publishers, 1997. Print.

[8] Holley, Robert W., James T. Madison, and Ada Zamir. “A New Method for Sequence Determination of Large Oligonucleotides.” Biochemical and Biophysical Research Communications 17.4 (1964): 389–394. ScienceDirect. Web.

[9] Heather, James M., and Benjamin Chain. “The Sequence of Sequencers: The History of Sequencing DNA.” Genomics 107.1 (2016): 1–8. ScienceDirect.

[10] Wikipedia contributors. "DNA microarray." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 3 Mar. 2017. Web. 31 Mar. 2017.

[11] “Microarray - How Does It Work? | Unsolved Mysteries of Human Health | Oregon State University.” N.p., n.d. Web. 31 Mar. 2017.

[12] Corney, David C. “RNA-Seq Using Next Generation Sequencing.” Materials and Methods (2016): n. pag. www.labome.com. Web. 31 Mar. 2017.

[13] “Thar She Blows! Ultra Long Read Method for Nanopore Sequencing · Loman Labs.” N.p., n.d. Web. 31 Mar. 2017.

[14] “Oxford Nanopore Technologies MinION.” N.p., n.d. Web. 31 Mar. 2017.

[15] “GenBank and WGS Statistics.” N.p., n.d. Web. 31 Mar. 2017.

[16] “Transcriptome Fact Sheet.” *National Human Genome Research Institute (NHGRI)*. N.p., n.d. Web. 31 Mar. 2017.

[17] “eMICE: Electronic Models Information, Communication, and Education.” N.p., n.d. Web. 31 Mar. 2017.

[18] “The Mouse Transcriptome.” N.p., n.d. Web. 31 Mar. 2017.

[19] Wang, Zhong, Mark Gerstein, and Michael Snyder. “RNA-Seq: A Revolutionary Tool for Transcriptomics.” *Nature reviews. Genetics* 10.1 (2009): 57–63. *PMC*. Web. 18 Mar. 2017.

[20] “RNA-Seqlopedia.” N.p., n.d. Web. 31 Mar. 2017.

[21] “RNA Sequencing | RNA-Seq Methods and Workflows.” N.p., n.d. Web. 31 Mar. 2017.

[22] Wikipedia contributors. "Single cell sequencing." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 3 Mar. 2017. Web. 31 Mar. 2017.

[23] Loman, Nicholas J. et al. “Performance Comparison of Benchtop High-Throughput Sequencing Platforms.” *Nature Biotechnology* 30.5 (2012): 434–439. *www.nature.com*.

[24] “Sequencing Platforms | Compare NGS Platforms (Benchtop, Production-Scale).” N.p., n.d. Web. 31 Mar. 2017.

[25] *Entrez Programming Utilities Help*. National Center for Biotechnology Information (US), 2010. Print.

[26] “Babraham Bioinformatics - FastQC A Quality Control Tool for High Throughput Sequence Data.” N.p., n.d. Web. 31 Mar. 2017.

[27] Schmieder, Robert, and Robert Edwards. “Quality Control and Preprocessing of Metagenomic Datasets.” Bioinformatics 27.6 (2011): 863–864. PMC. Web. 31 Mar. 2017.

[28] Kong, Yong. “Btrim: A Fast, Lightweight Adapter and Quality Trimming Program for next-Generation Sequencing Technologies.” Genomics 98.2 (2011): 152–153. ScienceDirect. Web.

[29] Dodt, Matthias et al. “FLEXBAR—Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms.” Biology 1.3 (2012): 895–905.

[30] “Najoshi/Sickle.” GitHub. N.p., n.d. Web. 31 Mar. 2017.

[31] Zhao, Shanrong, and Baohong Zhang. “A Comprehensive Evaluation of Ensembl, RefSeq, and UCSC Annotations in the Context of RNA-Seq Read Mapping and Gene Quantification.” *BMC Genomics* 16 (2015): 97. *BioMed Central*. Web.

[32] “Mouse Genome Assembly GRCm38.p5 - Genome Reference Consortium.” N.p., n.d. Web. 31 Mar. 2017.

[33] Patro, Rob, Stephen M. Mount, and Carl Kingsford. “Sailfish Enables Alignment-Free Isoform Quantification from RNA-Seq Reads Using Lightweight Algorithms.” Nature Biotechnology 32.5 (2014): 462–464. www.nature.com. Web.

[34] Langmead, Ben et al. “Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome.” Genome Biology 10 (2009): R25. BioMed Central. Web.

[35] Dobin, Alexander et al. “STAR: Ultrafast Universal RNA-Seq Aligner.” Bioinformatics 29.1 (2013): 15–21. academic.oup.com. Web.

[36] “Infphilo/Tophat.” GitHub. N.p., n.d. Web. 31 Mar. 2017.

[37] Kim, Daehwan, Ben Langmead, and Steven L. Salzberg. “HISAT: A Fast Spliced Aligner with Low Memory Requirements.” Nature Methods 12.4 (2015): 357–360. attachment. Web.

[38] Li, Heng. “Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM.” arXiv:1303.3997 [q-bio] (2013): n. pag. arXiv.org. Web. 31 Mar. 2017.

[39] “RPKM, FPKM and TPM, Clearly Explained | RNA-Seq Blog.” N.p., n.d. Web. 31 Mar. 2017.

[40] “Cole-Trapnell-Lab/Cufflinks.” GitHub. N.p., n.d. Web. 31 Mar. 2017.

[41] “Deweylab/RSEM.” GitHub. N.p., n.d. Web. 31 Mar. 2017.

[42] “Samtools/Samtools.” GitHub. N.p., n.d. Web. 31 Mar. 2017.

[43] Brunskill EW, Park JS, Chung E, Chen F et al. Single cell dissection of early kidney development: multilineage priming. *Development* 2014 Aug;141(15):3093-101.

[44] Treutlein B, Brownfield DG, Wu AR, Neff NF et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* 2014 May 15;509(7500):371-5.

[45] Streets AM, Zhang X, Cao C, Pang Y et al. Microfluidic single-cell whole-transcriptome sequencing. *Proc Natl Acad Sci U S A* 2014 May 13;111(19):7048-53.

[46] Kim DH, Marinov GK, Pepke S, Singer ZS et al. Single-cell transcriptome analysis reveals dynamic changes in lncRNA expression during reprogramming. *Cell Stem Cell* 2015 Jan 8;16(1):88-101.

[47] "If My Coefficient of Variation Is 47%, Is It Appropriate to Say..." ResearchGate. N.p., n.d. Web. 31 Mar. 2017.

[48] "Mouse Housekeeping Genes PCR Array." N.p., n.d. Web. 31 Mar. 2017.