

A DISSERTATION ON
Analysis of RNA Seq Data for Psoriasis Using R

SUBMITTED TO THE
DEPARTMENT OF BIOENGINEERING
FACULTY OF ENGINEERING
INTEGRAL UNIVERSITY, LUCKNOW



IN PARTIAL FULFILMENT
FOR THE
DEGREE OF MASTER OF TECHNOLOGY
IN BIOINFORMATICS

BY
Siddharth Gupta
M. Tech Bioinformatics (IV Semester)
Roll No: 2001081006

UNDER THE SUPERVISION OF
Dr. Rakesh Pandey
Assistant Professor and Coordinator
Bioinformatics
Mahila Mahavidyalaya, Banaras Hindu University, Varanasi



INTEGRAL UNIVERSITY, DASAULI, KURSI ROAD
LUCKNOW- 226026

DECLARATION FORM

I, **Siddharth Gupta**, a student of **M.Tech Bioinformatics** (2nd Year/ 4th Semester), Integral University have completed my six months dissertation work entitled “**Analysis of RNA Seq Data for Psoriasis Using R**” successfully from Mahila Mahavidyalaya, Banaras Hindu University, Varanasi, under the able guidance of **Dr. Rakesh Pandey**.

I, hereby, affirm that the work has been done by me in all aspects. I have sincerely prepared this project report and the results reported in this study are genuine and authentic.

Siddharth Gupta

Dr. Mohammed Kalim Ahmad Khan
Course Coordinator
Department of Bioengineering

काशी हिन्दू
विश्वविद्यालय



BANARAS HINDU
UNIVERSITY

AN INSTITUTION OF NATIONAL IMPORTANCE ESTABLISHED BY AN ACT OF PARLIAMENT

Dr. Rakesh Pandey
Assistant Professor and Coordinator
Bioinformatics
Mahila Mahavidyalaya
Banaras Hindu University
Varanasi, Uttar Pradesh, India

16th July 2022

Certificate

To Whom It May Concern

It is to certify that **Mr Siddharth Gupta** has completed his M. Tech. dissertation work under my supervision. He is doing M.Tech. in Bioinformatics from **Integral University, Lucknow**. His training period is from 17th January to 16th July 2022. The title of his dissertation is “**Analysis of RNA Seq Data for Psoriasis Using R**”. During the training, he has been very sincere about the tasks assigned to him and shown a willingness to learn new things.

I wish him all the very best for his future.

Yours Faithfully,

Rakesh Pandey



Varanasi- 221005
Mobile: 7678344922
Email: rakeshpandey.bhu.ac.in:
drakeshpandey@gmail.com



INTEGRAL UNIVERSITY

Established Under the Integral University Act 2004 (U.P. Act No.9 of 2004)

Approved by University Grant Commission

Phone No.: +91(0522) 2890812, 2890730, 3296117, 6451039, Fax No.: 0522-2890809

Kursi Road, Lucknow-226026 Uttar Pradesh (INDIA)

CERTIFICATE BY INTERNAL ADVISOR

This is to certify that **Siddharth Gupta**, a student of **M.Tech Bioinformatics** 2nd year/ 4th semester, Integral University has completed his six months dissertation work entitled “**Analysis of RNA Seq Data for Psoriasis Using R**” successfully. He has completed this work from Mahila Mahavidyalaya, Banaras Hindu University under the guidance of Dr. Rakesh Pandey, Assistant Professor and Coordinator, Bioinformatics. The dissertation was a compulsory part of his **M.Tech Bioinformatics**.

I wish him good luck and bright future.

Dr. Ashish

Assistant Professor

Department of Bioengineering

Faculty of Engineering



INTEGRAL UNIVERSITY

Established Under the Integral University Act 2004 (U.P. Act No.9 of 2004)

Approved by University Grant Commission

Phone No.: +91(0522) 2890812, 2890730, 3296117, 6451039, Fax No.: 0522-2890809

Kursi Road, Lucknow-226026 Uttar Pradesh (INDIA)

TO WHOM IT MAY CONCERN

This is to certify that **Siddharth Gupta**, a student of **M.Tech Bioinformatics** 2nd Year 4th semester, Integral University has completed her six months dissertation work entitled “**Analysis of RNA Seq Data for Psoriasis Using R**” successfully. He has completed this work from Mahila Mahavidyalaya, Banaras Hindu University, Varanasi, under the guidance of **Dr. Rakesh Pandey, Assistant Professor and Coordinator, Bioinformatics**. The dissertation was a compulsory part of his **M.Tech Bioinformatics**.

I wish him good luck and bright future.

Dr. Alvina Farooqui

Head

Department of Bioengineering

Faculty of Engineering

ACKNOWLEDGEMENT

Firstly, I would also like to take this opportunity to extend my sincere thanks to the authority of the University i.e., Chancellor Sir **Prof. S.W. Akthar**, Pro-Chancellor Sir **Dr. Syed Nadeem Akthar**, Vice-Chancellor Sir **Prof. Javed Musarrat**, Pro Vice-Chancellor Sir **Prof. Aqil Ahmad** for their valuable inputs for my research work and administrative support and guidance throughout my journey of this research work in the Integral University, Lucknow.

My sincere thanks also goes to Dean Engineering **Prof. T. Usmani**, Head **Dr. Alvina Farooqui**, PG Coordinator **Dr. Roohi**, Course Coordinator **Dr. Mohammed Kalim Ahmad Khan** who provided me an opportunity to join their team as an intern, and who gave me access to the laboratory and research facilities. Without their precious support, it would not be possible to conduct this research.

I would like to express my sincere gratitude to my Supervisor **Dr. Rakesh Pandey** and Internal Advisor **Dr. Ashish** for the continuous support of my thesis study and related research, for their patience, motivation, and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my thesis study.

At last, I would like to extend my heartfelt thanks to my parent because without their help this project would not have been successful. Finally, I would like to thank my dear friends, all lab assistants, faculty members for supporting me in this Dissertation and Thesis work.

Date:

Siddharth Gupta

CONTENTS

S.NO	List of Particulates	Page No.
1	Introduction	1
2	Review of literature	6
3	Material and methodology	13
	3.1 Quality control by Fast QC	14
	3.2 Building the reference index by Rstudio	18
	3.3 Alignment using Rsubread	18
	3.4 Feature Count using Rsubread in terminal	19
	3.5 Differentially Expressed Gene Analysis	21
4	Result and Discussion	22
	4.1 Volcano Plot	22
	4.2 MA Plot	23
	4.3 Heatmap	24
	4.4 Dispersion Plot	25
	4.5 PCA Biplot	26
5	Conclusion	28
6	Bibliography	29

LIST OF FIGURES

Figure No.	List of Particulates	Page No.
Fig 1	Workflow for RNA-Seq Analysis	13
Fig 2	Metadata of the sample on NCBI	14
Fig 3	Shows the quality control by using Fast QC of the following samples	14
	Fig 3a: SRR14108901_1	14
	Fig 3b: SRR14108901_2	14
	Fig 3c: SRR14108902_1	15
	Fig 3d: SRR14108902_2	15
	Fig 3e: SRR14108903_1	15
	Fig 3f: SRR14108903_2	15
	Fig 3g: SRR14108904_1	15
	Fig 3h: SRR14108904_2	15
	Fig 3i: SRR14108905_1	16
	Fig 3j: SRR14108905_2	16
	Fig 3k: SRR14108906_1	16
	Fig 3l: SRR14108906_2	16
	Fig 3m: SRR14108907_1	16
	Fig 3n: SRR14108907_2	16
	Fig 3o: SRR14108908_1	17
	Fig 3p: SRR14108908_2	17
	Fig 3q: SRR14108909_1	17
	Fig 3r: SRR14108909_2	17
	Fig 3s: SRR14108910_1	17
	Fig 3t: SRR14108910_2	17
Fig 4	Reference index was built using Rsubread in RStudio	18
Fig 5	The list of BAM files after alignment	19
Fig 6	Feature Count using Rsubread	19
Fig 7	Volcano Plot generated from DESeq2 Dataset	22

Fig 8	MA Plot generated from DESeq2 dataset	23
Fig 9	Heatmap generated from DESeq2 dataset	24
Fig 10	Dispersion Plot generated from DESeq2 dataset	25
Fig 11	PCA Biplot generated from DESeq2 dataset	26

LIST OF TABLES

Table No.	List of Particulates	Page No.
Table 1	Table of first 20 output after feature count of the sequence data	20
Table 2	First 20 outputs of the fold change and p value of the samples	27

ABBREVIATION

Abbreviation	Full form
RNA	Ribonucleic Acid
DNA	Deoxyribonucleic Acid
NGS	Next Generation Sequencing
HTS	High Throughput Sequencing
mRNA	Messenger Ribonucleic Acid
tRNA	Transfer Ribonucleic Acid
snRNA	Small nucleus Ribonucleic Acid
siRNA	Small intrinsic disturbances Ribonucleic Acid
lincRNA	Long-scatted non coding Ribonucleic Acid
tiRNA	Transcription initiation Ribonucleic Acid
miRNA	Micro- Ribonucleic Acid
TSSa RNA	Transcription initiation site-related Ribonucleic Acid
IL	Interleukin
TNF	Tumour necrosis factor
TWEAK	Tumour necrosis factor-like weak inducer of apoptosis
FPKM	Fragments per kilo base of transcript per million mapped fragments
GEO	Gene expression omnibus
QC	Quality control
DEG	Differentially expressed genes
scRNA	Small conditional Ribonucleic Acid
NCBI	National centre for biotechnology information
PCA	Transcription initiation site-related Principal components analysis

FASTA	Fast alignment sequence test for application
BAM	Binary alignment map
SAM	Sequence alignment map

INTRODUCTION

The Central Dogma of Molecular Biology outlines the flow of information that is stored in genes as DNA, transcribed into RNA, and finally translated into proteins (Crick, 1958; Crick, 1970). Early gene expression studies relied on low-throughput methods such as Northern blots and quantitative polymerase chain reaction (qPCR), but these were limited to single transcript measurements.

The development of next-generation high-throughput sequencing (NGS) has revolutionized transcriptomics by enabling RNA analysis with complementary DNA (cDNA) sequencing (Wang *et al.*, 2009). This method, called RNA-Sequencing, has clear advantages over previous approaches and has revolutionized the understanding of the complex and dynamic nature of the transcriptome. RNA-Sequencing provides a more detailed and quantitative view of gene expression, alternative splicing, and allele-specific expression. Recent advances in RNA-Sequencing workflows, from sample preparation to sequencing platforms to bioinformatics data analysis, have enabled detailed transcriptome profiling and the ability to elucidate a variety of physiological and pathological conditions. rice field. The advent of high-throughput next-generation sequencing (NGS) technology has revolutionized transcriptomics. This technological development solves many of the challenges posed by the hybridization-based microarray and Sanger sequencing-based approaches previously used to measure gene expression.

High-throughput sequence (HTS) data analysis is a complex multi-step process. Many bioinformatics tools are available at most steps, and most tools require different parameters to be set. Due to this complexity, HTS data analysis is particularly prone to reproducibility and consistency issues. The high-throughput sequencer enables transcriptome inspection. The transcriptome is a set of intracellular ribonucleic acids, including messenger ribonucleic acid (mRNA), transfer ribonucleic acid (tRNA), ribosomal ribonucleic acid (rRNA), small nucleus ribonucleic acid (snRNA), and non-coding ribonucleic acid (ncRNA), others. These RNAs are expressed differentially depending on the tissue, physiological state, or developmental stage (Gupta *et al.*, 2021). Interpreting the complexity of the transcriptome is an important goal for understanding the functional elements of the genome, and therefore for understanding how the disease functions and signs of progress. In this sense, the amount of non-coding DNA has recently been shown to increase with biological complexity, increasing by 0.25% in the prokaryotic genome and 98.8% in the

human genome. Existing complexity associated with the discovery of small intrinsic disturbances RNA (siRNA), long-scattered non-coding RNA (lincRNA), transcription initiation RNA (tiRNA), microRNA (miRNA), transcription initiation site-related RNA (TSSa-RNA), etc. is the transcription puzzles we need. Represents a piece of. Elucidate to understand how the genome works.

Psoriasis is one of the most common immune inflammatory skin diseases, affecting approximately 125 million people worldwide and more than 8 million in the United States (Rachakonda *et al.*, 2014). Psoriasis lesions can exhibit a variety of clinical manifestations, including acanthosis (increased epidermal thickness), keratin proliferation, parakeratosis, hypervascularization, and dense skin infiltration of immune cells (Gran *et al.*, 2020). Keratinocytes have central importance for inducing early pathogenic events and for increasing psoriatic inflammation during the course of the disease (Albanesi *et al.*, 2018, Benhadou *et al.*, 2019). In response to external and internal threat stimuli, keratinocytes can be a source of innate immune mediators. These include various pro-inflammatory cytokines and chemokines that mobilize cells important for innate and adaptive immune responses (Li *et al.*, 2014, Takagi *et al.*, 2016). The IL-23 / IL-17 axis and TNF were first identified in animal studies as the centre of pathogenesis for skin inflammation such as psoriasis, and their role is now being demonstrated in humans. IL-36 γ is also strongly associated with human psoriasis. IL-36 γ is produced by keratinocytes and can induce the expression of the IL-23 gene in keratinocytes (Goldstein *et al.*, 2020). Therefore, it is possible to drive a strengthening loop from IL-23 back to IL-17, IL-36 γ , and IL-23, thereby maintaining the condition. All of these cytokines are elevated in psoriatic skin lesions, and proper neutralization of TNF, IL-23 p19, or IL-17A has shown potential therapeutic effects in psoriatic patients (Gran *et al.*, 2020, Schon, 2019, Yamanaka *et al.*, 2021). Although these current treatments have proven to be effective, some patients do not respond or become refractory over time, or the disease relapses when treatment is stopped. Therefore, understanding the pathological mechanisms that can occur in psoriasis requires further efforts, such as identifying new molecules that can be targeted alone or in combination with existing therapies.

TNF and IL-17 are two cytokines that promote dysregulated keratinocyte activity, and their targeting is very effective in psoriasis patients, but whether these molecules interact with other inflammatory factors. Is not clear. Here, mice with a keratinocyte-specific deletion of Fn14 (Tnfrsf12a), a receptor for the TNF superfamily cytokine TWEAK (Tnfsf12), have

imiquimod-induced skin inflammation such as decreased epidermal hyperplasia and decreased expression of the psoriasis signature gene. Indicates a decrease in. This corresponded to the expression of Fn14 in the keratinocytes of human psoriasis lesions and TWEAK being found in several sub-sets of skin cells. Transcriptomic studies in human keratinocytes revealed that TWEAK strongly overlaps with IL-17A and TNF in upregulating the expression of CXC chemokines, along with cytokines such as IL-23, inflammation-associated proteins like S100A8/9 and SERPINB1/B9, all previously found to be highly expressed in the lesional skin of psoriasis patients (Gupta *et al.*, 2021)

Although these current treatments have proven efficacy, some patients fail to respond or become resistant to therapy over time, or their disease comes back when treatment is stopped. Therefore, continuing efforts to understand the pathological mechanisms that might occur in psoriasis are needed, including identifying novel molecules that can be targeted alone or combined with existing therapies. TNF-like weak inducer of apoptosis (TWEAK, TNFSF12) can be expressed similar to TNF (TNFSF2) is a membrane-bound molecule or soluble cytokine by a variety of cell types including structural and immune cells (Chicheportiche *et al.*, 1997, Bird *et al.*, 2013). TWEAK binds to Fn14 (fibroblast growth factor inducible 14, TNFRSF12A) and regulates many cellular activities such as proliferation, migration, differentiation, apoptosis, and angiogenesis (Leng *et al.*, 2011). TWEAK is involved in the pathogenesis of several inflammatory and autoimmune diseases (Burkly, 2014, Doerner *et al.*, 2016). Recently, we have discovered that TWEAK-deficient mice are protected from exhibiting severe imiquimod-induced skin inflammation with some characteristics of psoriasis. Gene set enrichment analysis suggests an association between Fn14 transcripts and their signaling mediators in human psoriasis lesions (Leng *et al.*, 2011). The pathogenic activity of TWEAK was subsequently validated by another group using Fn14-deficient mice in the same experimental model (Doerner *et al.*, 2015). Other literature has found that soluble TWEAK is upregulated in the sera of psoriasis patients and that expression of both TWEAK and Fn14 is detected at high levels in tissue sections of psoriasis-damaged skin (Sidler *et al.*, 2017, Peng *et al.*, 2018). A new therapeutic approach to reduce skin lesions in psoriasis. The TWEAK primary cell target in the skin is unclear. Subcutaneous injection of recombinant TWEAK bolus into mice was found to result in skin inflammation and some histological features reminiscent of human psoriasis. It was associated with the production of a series of chemokines that attract the innate and adaptive immune cells characteristic of psoriasis (Sidler *et al.*, 2017). Many of these chemokines are products of keratinocytes, and Fn14 is expressed in keratinocytes

(Sidler *et al.*, 2017), suggesting that this cell type may be central to the action of TWEAK. Before considering clinical treatment for this pathway, how TWEAK in the skin, especially on keratinocytes, and its relationship to other pathogenic molecules such as IL-17 and TNF that also have receptors on keratinocytes

In this study, we investigated if TWEAK signalling specifically in keratinocytes is required to develop psoriasis-like skin lesions after imiquimod treatment using Fn14-conditional knockout mice, and also performed RNA-sequencing analysis in human epidermal keratinocytes to determine how TWEAK alone or in combination with IL-17 and TNF controls expression of a variety of gene sets found to be upregulated in human psoriasis. Our data demonstrate that Fn14 signalling in keratinocytes is crucial for the development of imiquimod-induced skin inflammation. Furthermore, transcriptomic data establish substantial similarities in the genes induced in keratinocytes by TWEAK, IL-17, and TNF, and notably, we found strong synergistic activities of these cytokines acting together on a number of genes associated with psoriasis. Correspondingly, a similar effect of blocking TWEAK therapeutically was observed in reducing skin lesions in mice compared to blocking either TNF or IL-17A, and no greater effect was seen with combination treatments. These results suggest that TWEAK might be as good a target to counter the keratinocyte hyperresponsiveness and dysregulated immune system seen in psoriasis as observed when IL-17 and TNF are neutralized (Wang *et al.*, 2021, Bilgic *et al.*, 2016)

The main goal of many gene expression experiments is to detect transcripts that exhibit differential expression under a variety of conditions. Extensive statistical approaches have been developed to test differential expression using microarray data, and the continuous probe intensity of the entire replication can be approximated by a normal distribution (Chandran and Raychaudhuri, 2010, Cui and Churchill, 2003, Smyth, 2004). While these approaches can, in principle, be applied to RNA-Sequencing data, other statistical models of discrete read counts that do not fit the normal distribution should be considered. Early RNA-Sequencing studies showed that the distribution of read counts throughout replication follows a Poisson distribution. This formed the basis for modelling RNA-Sequencing count data (Grant *et al.*, 2005). However, further studies have shown that biological variability is not captured by Poisson's assumptions and leads to high false positive rates due to underestimation of sampling errors (Marioni *et al.*, 2008, Anders and Huber, 2010, Lanham et al., 2010). Therefore, a negative binomial distribution model that describes

overdispersion or extra-Poisson variability has been shown to best fit the distribution of read counts across biological replication.

REVIEW OF LITERATURE

Psoriasis Vulgaris is a chronic disease that affects 1–3% of the population (Robinson and Oshlack, 2010). In addition to the possible involvement of skin and joints, recent evidence suggests a link between psoriasis and other systemic disorders (Gelfand *et al.*, 2006). The molecular properties of psoriasis skin samples have led to a better understanding of the etiology of the disease and helped identify therapeutic targets (Lebwohi, 2003). Psoriasis is one of the most common chronic inflammatory skin diseases, affecting 1-3% of the adult population worldwide (Lebwohi, 2003). It is characterized by marked overgrowth and inadequate end differentiation of keratinocytes. In addition, complex interactions between different cell types and various cytokines are known to contribute to the development of psoriasis. The etiology is also based on complex interactions between genetic predisposition, important histocompatibility alleles, and various environmental triggers (Lowe *et al.*, 2007). However, from a molecular perspective, the mechanisms responsible for the interaction of keratinocytes with the inflammatory cells that infiltrate the epidermis are not yet fully understood. Analysis of the molecular background of psoriasis describes many disease-related genes and proteins with aberrant expression patterns (Nomura *et al.*, 2003), but little is known about the regulatory pathways responsible for this aberrant expression. Recent evidence suggests that non-coding RNAs such as microRNAs (miRNAs) and long noncoding RNAs (lncRNAs) contribute to the pathogenesis of psoriasis by affecting protein expression and function in both keratinocytes and inflammatory cells. It suggests that it may be (Sonkoly *et al.*, 2007, Zibert *et al.*, 2010, Ahn *et al.*, 2016, Gupta *et al.*, 2016, Tsoi *et al.*, 2015). RNA Sequencing Fundamentals: RNA Sequencing is the use of next-generation high-throughput sequencing technology to study, characterize, and quantify genomic transcriptomes (Morin *et al.*, 2008). Unlike previous methods, RNA sequencing uses synthetic techniques to define nucleotide sequences and quantify RNA molecules in a sample (Wang *et al.*, 2009). Next-generation sequencing (NGS) can faithfully process this data in hours to days, making it an ideal method for RNA analysis among many researchers (Kolodziejczyk *et al.*, 2015). The use of this technology in research and literature has exploded in popularity. With recent discoveries in the use of RNA sequencing in many pathologies, there are many promising potential clinical applications for RNA sequencing (Beane *et al.*, 2011). Several commercially available RNA sequencing kits are available for each sample. Most follow similar processing steps

but ultimately depend on experimental considerations (Chu and Corey, 2012). Analysis of total RNA, mRNA, and small RNA can be performed with most kits. To isolate mRNA, use poly (T) primers attached to beads or magnets to bind mRNA and isolate these strands. For small or non-coding RNA, gel electrophoresis is used to separate these molecules. Complete RNA separation uses a combination of these two techniques (Tuch *et al.*, 2010). Then ligate the adapter to the 5'end, 3'end, or both. When RNA is isolated, cDNA is generated, amplified, and fragmented. Some kits provide RNA sequencing directly without creating cDNA. Although rRNA makes up a significant proportion of total RNA and can be removed, it has little research interest. These samples are then sequenced by next-generation massively parallel sequencing technology that utilizes sequencing by synthesizing short DNA strands complementary to cDNA. Once the reads are generated, the software can be used to analyse the sequence reads and match the reads to parts of the genome. You can also create a de novo transcriptome map by mapping gene fragments with sequencing analysis software. The total number of reads for each gene product can be used to quantify proportional gene expression (Han *et al.*, 2015).

The use of RNA-Sequencing has recently increased due to advances beyond previous attempts in transcriptome research. Prior to NGS RNA sequencing, two well-known techniques were available. Hybridization of cDNA probes connected to microarrays enabled transcriptome analysis but was limited by the need for extensive knowledge of genomes, transcripts, alternative splicing, and exons. The background noise produced by cross-hybridization also limited resolution during attempts to quantify gene expression. Another technique was Sanger sequencing, which used chain termination to determine nucleotide sequences. In contrast to NGS, the Sanger method was more expensive and time-consuming and could only analyze a limited portion of the transcript (Morin *et al.*, 2008, Wang *et al.*, 2009, Burroughs *et al.*, 2013). Discovery of both non-coding RNAs such as B. miRNAs (miRNAs) have required the creation of assays to test these small non-coding RNAs with variant mRNAs at high throughput and high resolution, as well as the discovery of post-transcriptional mRNA expression regulation (Klerk and Hoen, 2015). RNA-Sequencing techniques allow researchers to perform both of these tasks and quantify RNA expression, and thus gene expression, in a single assay. The high throughput of RNA sequences allows the transcriptome to be analyzed and efficiently compared across different environmental factors such as time, different tissue samples, pathological conditions, and pharmacological interventions. The potential for de novo transcriptome

synthesis allows the analysis and discovery of new products without the need for prior genomic and transcriptional knowledge of the sample. The resolution of RNA sequences also enables the identification of single nucleotide polymorphisms, novel post-transcriptional modifications, novel alternative splicing patterns, and previously unidentified non-coding RNA molecules. RNA sequencing provides accurate quantification of mRNA expression compared to real-time PCR experiments (Scapato *et al.*, 2015, de Klerk *et al.*, 2014, Derks *et al.*, 2015). RNA sequences can be used to study the molecular basis of disease susceptibility, cancer etiology/progression, and response to treatment. RNA sequences have been used to analyze the etiology of various malignancies such as psoriasis, lung cancer, and colon cancer. RNA sequencing can identify differential expression of genes (DEGs), mutant genes, fusion genes, and gene isoforms in pathological conditions. RNA sequencing also has potential for diagnostic and therapeutic applications. Current research on colorectal disease using RNA sequencing reveals new discoveries that may help clinicians in the future management of patients with colorectal disease.

Transcriptome analysis is an important tool for characterizing and understanding the molecular basis of phenotypic changes in biology, including disease. In recent decades, microarrays have been the most important and widely used approach to such analysis, but recently high-throughput cDNA sequencing (RNA-sequencing) has emerged as a powerful alternative (Mortazavi *et al.*, 2008). Many applications have already been found (Chen *et al.*, 2011). RNA-sequencing uses next-generation sequencing (NGS) methods to sequence cDNA from RNA samples, producing millions of short reads. These reads are then typically mapped to the reference genome, and the number of reads mapped within the genomic traits of interest (such as genes or exons) is used as a measure of the frequency of the traits of the analyzed sample (Oshlack *et al.*, 2010).

Perhaps the most common use of transcriptome profiling is to search for differentially expressed (DE) genes. H. Look for genes that show differences in expression levels between conditions, or genes that are associated with a particular predictor or response. RNA-sequencing offers several advantages over microarrays for differential expression analysis. B. Ability to detect and quantify previously unknown transcripts and isoforms with increased dynamic range and reduced background levels (Agrawal *et al.*, 2010, Bradford *et al.*, 2010, Bullard *et al.*, 2010). However, analysing RNA-sequencing data can be difficult. Some of these issues are unique to next-generation sequencing methods. For

example, differences in nucleotide composition between genomic regions mean that reading ranges may not be uniform throughout the genome. In addition, more reads are mapped to longer genes than shorter genes with the same expression level. In differential expression analysis, where genes are individually tested for differences in expression between conditions, biases within the sample are usually ignored as they are expected to affect all samples in a similar manner (Agrawal *et al.*, 2010).

RNA-sequencing experiments show other types of heterogeneity between samples. First, the depth of the sequence or the library size (total number of reads allocated) usually varies from sample to sample. That is, the counts observed between the samples cannot be compared directly. In fact, even in the absence of true differential expression, if one sample is sequenced twice as deep as another, then all genes in the first sample receive twice as many as the second sample. It is expected that we would like to avoid such confusion. The effect of true differential expression. The easiest way to approach different library sizes is to simply rescale or resample the read counts to get the same library size for all samples. However, such normalization is generally not sufficient. This is because RNA-Sequencing counts essentially represent the relative abundance of genes, even if the libraries are actually the same size. Some highly expressed genes can make up a very large proportion of the reads sequenced in the experiment, so few reads need to be assigned to the remaining genes (Bullard *et al.*, 2010). Therefore, the presence of a small number of highly expressed genes suppresses the count of all other genes, and the latter group of genes are mis expressed compared to samples with more evenly distributed reads. It is misunderstood that it can appear low and can lead to many genes. More complex normalization schemes have been proposed to address this difficulty and allow counts to be compared between samples (Bullard *et al.*, 2010, Anders and Huber, 2010, Robinson and Oshlack, 2010). In addition to library size, these methods also include estimating sample-specific normalization coefficients. It is used to rescale the observed count. Using these normalization methods, the sum of the normalized counts across all genes are therefore not necessarily equal between samples (as it would be if only the library sizes were used for normalization), but the goal is instead to make the normalized counts for non-differentially expressed genes similar between the samples. In this study, we use the TMM normalization (trimmed mean of M-values (Robinson and Oshlack, 2010)) and the normalization provided in the DESeq package (Anders and Huber, 2010). A comprehensive evaluation of seven different normalization methods was recently performed (Dillies *et al.*, 2012), in which these two

methods were shown to perform similarly, and they were also the only ones providing satisfactory results with respect to all metrics used in that evaluation. Still, it is important to keep in mind that even these methods are based on an assumption that most genes are equivalently expressed in the samples, and that the differentially expressed genes are divided more or less equally between up- and downregulation (Dillies *et al.*, 2012).

Microarrays have been used routinely for differential expression analysis for over a decade, and there are well-established methods available for this purpose (such as limma (Smyth, 2004)). These methods cannot be easily migrated to the analysis of RNA-sequencing data (Robinson and Smyth, 2008).

It is different from the data obtained from the microarray. Intensities recorded from microarrays are treated as continuous measurements and are generally assumed to follow a lognormal distribution, but counts from RNA-sequencing experiments are non-negative integers and therefore essentially follow a discrete distribution. Poisson distribution and negative binomial distribution (NB) are the two most commonly used models in the method explicitly developed for differential expression analysis of this type of count data (Anders and Huber, 2010, Robinson and Symth, 2008, Auer and Doerge, 2011, Hardcastle and Kelly, 2010, Di *et al.*, 2011). Other distributions such as the beta-binomial distribution (Zhou *et al.*, 2011) have also been proposed. The Poisson distribution has the advantage of simplicity, with only one parameter, but limits the variance of the modelled variables to the mean. The negative binomial distribution has two parameters that encode the mean and variance, so you can model the more general mean and variance relationship. For RNA-sequencing, the Poisson distribution has been suggested to be suitable for the analysis of engineering replication, but with high variability between biological replications, it is accompanied by overdispersion, such as a negative binomial distribution. Distribution is required (Bullard *et al.*, 2010, Marioni *et al.*, 2008). Some software packages represent RNA-sequencing data in converted quantities instead of using integers directly. Long transcripts are expected to receive more reads than short transcripts with the same expression level, so the goal of such a conversion is to normalize the count in relation to various library sizes and transcript lengths. Is to do. Other normalization strategies can be used to address other biases, such as biases due to variable GC content in reads. After such a conversion, the resulting value will no longer be an integer count. That is, you should not plug in numerical-based methods for differential expression analysis. Therefore, of the

methods evaluated in this study, only nonparametric methods are suitable for RPKM values. Other software, such as Cufflinks / Cuffdiff (Trapnell *et al.*, 2010), provides an integrated analytical pipeline from aligned reads to derivative results by inference based on FPKM values.

The field of differential expression analysis of RNA-sequencing data is still in its infancy, and new methods are constantly being introduced. To date, there has been no general consensus on which method works best in a particular situation, and few detailed comparisons between the proposed methods have been published. In a recent publication (Kyam *et al.*, 2012), four parametric methods were compared in terms of their ability to distinguish between truly differentially expressed (DE) and truly non-DE genes under different simulation conditions. The authors also compared duplications between sets of DE genes found differently in practice data set. Another recent study (Robles *et al.*, 2012) evaluated the effect of increased sequence depth on the ability to detect the DE gene and contrasted this with the benefits of increased sample size, the latter demonstrating to be significantly greater. In (Nookaew *et al.*, 2012), the authors published a case study on *Saccharomyces cerevisiae*, comparing the results of several differential expression analysis methods of RNA-sequencing with each other, comparing them with the results of microarrays, and generally between different methods.

In this study, we investigated if TWEAK signalling specifically in keratinocytes is required to develop psoriasis-like skin lesions after imiquimod treatment using Fn14-conditional knockout mice, and also performed RNA-sequencing analysis in human epidermal keratinocytes to determine how TWEAK alone or in combination with IL-17 and TNF controls expression of a variety of gene sets found to be upregulated in human psoriasis. Our data demonstrates that Fn14 signalling in keratinocytes is crucial for the development of imiquimod-induced skin inflammation. Furthermore, transcriptomic data establish substantial similarities in the genes induced in keratinocytes by TWEAK, IL-17, and TNF, and notably we found strong synergistic activities of these cytokines acting together on a number of genes associated with psoriasis. Correspondingly, a similar effect of blocking TWEAK therapeutically was observed in reducing skin lesions in mice compared to blocking either TNF or IL-17A, and no greater effect was seen with combination treatments. These results suggest that TWEAK might be as good a target to counter the

keratinocyte hyperresponsiveness and dysregulated immune system seen in psoriasis as observed when IL-17 and TNF are neutralized (Gupta *et al.*, 2021).

MATERIALS AND METHODOLOGY

Workflow is the series of activities that are necessary to complete a task. Each step in a workflow has a specific step before it and a specific step after it. Workflow for RNA Sequencing analysis is show in figure 1.

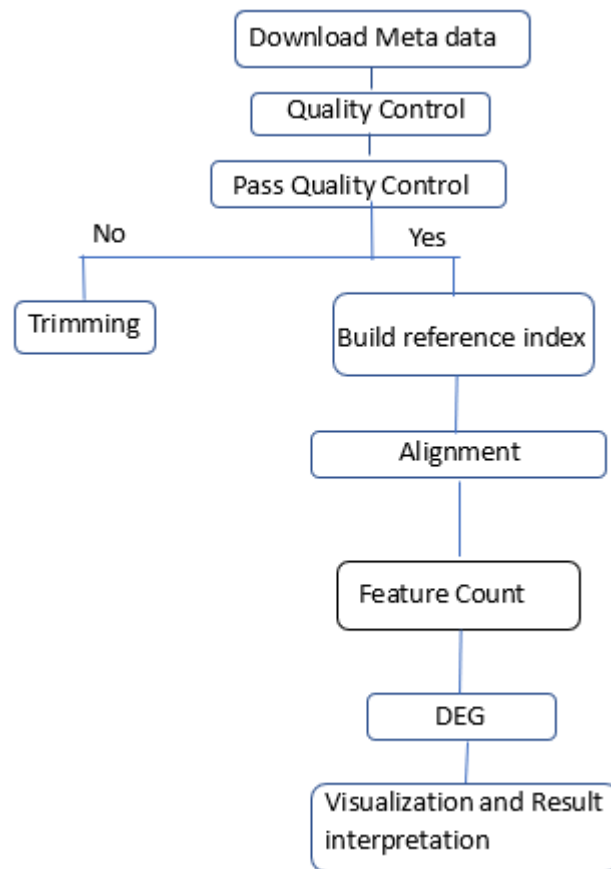


Fig 1: Workflow for RNA-Seq Analysis

The sample sequences were downloaded from the NCBI GEO Dataset (Gupta *et al.*, 2021). 10samples of paired-end sequencing were selected, out of which 6 were TWEAK stimulated and 4 were TNF stimulated, the metadata of the samples was downloaded on the workstation having an Intel Xeon 3.20GHz x20 processor and 150GB of RAM, 10 cores. The list of samples is shown in figure 2.

Accession: PRJNA718552

Common Fields:

- BioProject: PRJNA718552
- Consent: PUBLIC
- Assay Type: RNA-Seq
- AugSpotLen: 100
- Cell Line: hHEK
- Cell Type: Human epidermal keratinocytes from neonates (hHEK)
- Center Name: GEO
- DATASTORE Metatype: FASTQ_SRA
- DATASTORE Provider: GCS_MFR_13

Select:

- Total: 36 Runs, 21.95 Gb Bytes, 72.97 G Bases, Metadata, Accession List, Cloud Data Delivery, Computing
- Selected: 10 Runs, 5.84 Gb Bytes, 19.40 G Bases, Metadata, Accession List, JWT Cart, Deliver Data, Galaxy

Run	Bytes	Bases	Download	GEO Accession	Sample Name	source_name	treatment		
✓ 1	SRR14108901	SAAN18542703	1.95 G	598.86 Mb	SRX10479603	GM5220264	GM5220264	Skin Keratinocyte_TWEAK stimulated	TWEAK (100ng/ml) stimulated
✓ 2	SRR14108902	SAAN18542703	1.91 G	588.80 Mb	SRX10479603	GM5220264	GM5220264	Skin Keratinocyte_TWEAK stimulated	TWEAK (100ng/ml) stimulated
✓ 3	SRR14108903	SAAN18542702	1.76 G	540.44 Mb	SRX10479604	GM5220265	GM5220265	Skin Keratinocyte_TWEAK stimulated	TWEAK (100ng/ml) stimulated
✓ 4	SRR14108904	SAAN18542702	1.73 G	534.79 Mb	SRX10479604	GM5220265	GM5220265	Skin Keratinocyte_TWEAK stimulated	TWEAK (100ng/ml) stimulated
✓ 5	SRR14108905	SAAN18542701	2.13 G	655.60 Mb	SRX10479605	GM5220266	GM5220266	Skin Keratinocyte_TWEAK stimulated	TWEAK (100ng/ml) stimulated
✓ 6	SRR14108906	SAAN18542701	2.17 G	662.37 Mb	SRX10479605	GM5220266	GM5220266	Skin Keratinocyte_TWEAK stimulated	TWEAK (100ng/ml) stimulated
✓ 7	SRR14108907	SAAN18542700	1.84 G	574.21 Mb	SRX10479606	GM5220267	GM5220267	Skin Keratinocyte_TNF stimulated	TNF (10 ng/ml) stimulated
✓ 8	SRR14108908	SAAN18542700	1.88 G	579.79 Mb	SRX10479606	GM5220267	GM5220267	Skin Keratinocyte_TNF stimulated	TNF (10 ng/ml) stimulated
✓ 9	SRR14108909	SAAN18542699	1.99 G	617.17 Mb	SRX10479607	GM5220268	GM5220268	Skin Keratinocyte_TNF stimulated	TNF (10 ng/ml) stimulated
✓ 10	SRR14108910	SAAN18542699	2.04 G	627.51 Mb	SRX10479607	GM5220268	GM5220268	Skin Keratinocyte_TNF stimulated	TNF (10 ng/ml) stimulated

Fig 2: Metadata of the sample on NCBI

Quality Control by Fast QC

Then, the data were analyzed for quality control and trimming using Fast QC, which provides a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines, and the outcome of the Fast QC analysis shows whether the trimming is needed or not. Comparing the results from standards suggests, that trimming is not needed in the data obtained, the result of Fast QC is also shown in the figures 3. The data was good with little noise. (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>)

SRR14108901_1.Fastq.gz	SRR14108901_2.Fastq.gz	SRR14108902_1.Fastq.gz	SRR14108902_2.Fastq.gz	SRR14108901_1.Fastq.gz	SRR14108901_2.Fastq.gz	SRR14108902_1.Fastq.gz	SRR14108902_2.Fastq.gz
Basic sequence stats				Basic sequence stats			
✓ Basic Statistics	Measure	Value		✓ Basic Statistics	Measure	Value	
✓ Per base sequence quality	Filename	SRR14108901_1.Fastq.gz		✓ Per base sequence quality	Filename	SRR14108901_2.Fastq.gz	
✓ Per tile sequence quality	File type	Conventional base calls		✓ Per tile sequence quality	File type	Conventional base calls	
✓ Per sequence quality scores	Encoding	Sanger / Illumina 1.9		✓ Per sequence quality scores	Encoding	Sanger / Illumina 1.9	
✓ Per base sequence content	Total Sequences	19521208		✓ Per base sequence content	Total Sequences	19521208	
✓ Per sequence GC content	Sequences Flagged as poor quality	0		✗ Per base sequence content	Sequences Flagged as poor quality	0	
✓ Per base N content	Sequence length	50		✓ Per sequence GC content	Sequence length	50	
✓ Sequence Length Distribution	%GC	52		✓ Per base N content	%GC	53	
✗ Sequence Duplication Levels				✓ Sequence Length Distribution			
! Overrepresented sequences				✗ Sequence Duplication Levels			
✓ Adapter Content				! Overrepresented sequences			
✗ Kmer Content				✓ Adapter Content			
				✗ Kmer Content			

Fig 3a: SRR14108901_1

Fig 3b: SRR14108901_2

SRR14108901_1.fastq.gz		SRR14108902_1.fastq.gz	
Basic sequence stats			
Measure	Value		
Per base sequence quality	Filename	SRR14108902_1.fastq.gz	
Per tile sequence quality	File type	Conventional base calls	
Per sequence quality scores	Encoding	Sanger / Illumina 1.9	
Per base sequence content	Total Sequences	19070351	
Per sequence GC content	Sequences flagged as poor quality	0	
Per base N content	Sequence length	50	
Sequence Length Distribution	%GC	52	
Sequence Duplication Levels			
Overrepresented sequences			
Adapter Content			
Kmer Content			

Fig 3c: SRR14108902_1

SRR14108901_1.fastq.gz		SRR14108902_2.fastq.gz	
Basic sequence stats			
Measure	Value		
Per base sequence quality	Filename	SRR14108902_2.fastq.gz	
Per tile sequence quality	File type	Conventional base calls	
Per sequence quality scores	Encoding	Sanger / Illumina 1.9	
Per base sequence content	Total Sequences	19070351	
Per sequence GC content	Sequences flagged as poor quality	0	
Per base N content	Sequence length	50	
Sequence Length Distribution	%GC	53	
Sequence Duplication Levels			
Overrepresented sequences			
Adapter Content			
Kmer Content			

Fig 3d: SRR14108902_2

SRR14108903_1.fastq.gz		SRR14108904_1.fastq.gz	
Basic sequence stats			
Measure	Value		
Per base sequence quality	Filename	SRR14108903_1.fastq.gz	
Per tile sequence quality	File type	Conventional base calls	
Per sequence quality scores	Encoding	Sanger / Illumina 1.9	
Per base sequence content	Total Sequences	17613371	
Per sequence GC content	Sequences flagged as poor quality	0	
Per base N content	Sequence length	50	
Sequence Length Distribution	%GC	52	
Sequence Duplication Levels			
Overrepresented sequences			
Adapter Content			
Kmer Content			

Fig 3e: SRR14108903_1

SRR14108903_1.fastq.gz		SRR14108904_2.fastq.gz	
Basic sequence stats			
Measure	Value		
Per base sequence quality	Filename	SRR14108904_2.fastq.gz	
Per tile sequence quality	File type	Conventional base calls	
Per sequence quality scores	Encoding	Sanger / Illumina 1.9	
Per base sequence content	Total Sequences	17613371	
Per sequence GC content	Sequences flagged as poor quality	0	
Per base N content	Sequence length	50	
Sequence Length Distribution	%GC	53	
Sequence Duplication Levels			
Overrepresented sequences			
Adapter Content			
Kmer Content			

Fig 3f: SRR14108903_2

SRR14108903_1.fastq.gz		SRR14108904_1.fastq.gz	
Basic sequence stats			
Measure	Value		
Per base sequence quality	Filename	SRR14108904_1.fastq.gz	
Per tile sequence quality	File type	Conventional base calls	
Per sequence quality scores	Encoding	Sanger / Illumina 1.9	
Per base sequence content	Total Sequences	17309068	
Per sequence GC content	Sequences flagged as poor quality	0	
Per base N content	Sequence length	50	
Sequence Length Distribution	%GC	52	
Sequence Duplication Levels			
Overrepresented sequences			
Adapter Content			
Kmer Content			

Fig 3g: SRR14108904_1

SRR14108903_1.fastq.gz		SRR14108904_2.fastq.gz	
Basic sequence stats			
Measure	Value		
Per base sequence quality	Filename	SRR14108904_2.fastq.gz	
Per tile sequence quality	File type	Conventional base calls	
Per sequence quality scores	Encoding	Sanger / Illumina 1.9	
Per base sequence content	Total Sequences	17309068	
Per sequence GC content	Sequences flagged as poor quality	0	
Per base N content	Sequence length	50	
Sequence Length Distribution	%GC	53	
Sequence Duplication Levels			
Overrepresented sequences			
Adapter Content			
Kmer Content			

Fig 3h: SRR14108904_2

SRR14108905_1.Fastq.gz	SRR14108905_2.Fastq.gz	SRR14108906_1.Fastq.gz	SRR14108906_2.Fastq.gz	SRR14108905_1.Fastq.gz	SRR14108905_2.Fastq.gz	SRR14108906_1.Fastq.gz	SRR14108906_2.Fastq.gz																																								
<p>Basic Statistics</p> <p>Per base sequence quality</p> <p>Per tile sequence quality</p> <p>Per sequence quality scores</p> <p>Per base sequence content</p> <p>Per sequence GC content</p> <p>Per base N content</p> <p>Sequence Length Distribution</p> <p>Sequence Duplication Levels</p> <p>Overrepresented sequences</p> <p>Adapter Content</p> <p>Kmer Content</p>				<p>Basic sequence stats</p> <table border="1"> <thead> <tr> <th>Measure</th> <th>Value</th> </tr> </thead> <tbody> <tr> <td>Filename</td> <td>SRR14108905_1.Fastq.gz</td> </tr> <tr> <td>File type</td> <td>Conventional base calls</td> </tr> <tr> <td>Encoding</td> <td>Sanger / Illumina 1.9</td> </tr> <tr> <td>Total Sequences</td> <td>21321329</td> </tr> <tr> <td>Sequences flagged as poor quality</td> <td>0</td> </tr> <tr> <td>Sequence length</td> <td>50</td> </tr> <tr> <td>%GC</td> <td>52</td> </tr> </tbody> </table>				Measure	Value	Filename	SRR14108905_1.Fastq.gz	File type	Conventional base calls	Encoding	Sanger / Illumina 1.9	Total Sequences	21321329	Sequences flagged as poor quality	0	Sequence length	50	%GC	52	<p>Basic Statistics</p> <p>Per base sequence quality</p> <p>Per tile sequence quality</p> <p>Per sequence quality scores</p> <p>Per base sequence content</p> <p>Per sequence GC content</p> <p>Per base N content</p> <p>Sequence Length Distribution</p> <p>Sequence Duplication Levels</p> <p>Overrepresented sequences</p> <p>Adapter Content</p> <p>Kmer Content</p>				<p>Basic sequence stats</p> <table border="1"> <thead> <tr> <th>Measure</th> <th>Value</th> </tr> </thead> <tbody> <tr> <td>Filename</td> <td>SRR14108905_2.Fastq.gz</td> </tr> <tr> <td>File type</td> <td>Conventional base calls</td> </tr> <tr> <td>Encoding</td> <td>Sanger / Illumina 1.9</td> </tr> <tr> <td>Total Sequences</td> <td>21321329</td> </tr> <tr> <td>Sequences flagged as poor quality</td> <td>0</td> </tr> <tr> <td>Sequence length</td> <td>50</td> </tr> <tr> <td>%GC</td> <td>53</td> </tr> </tbody> </table>				Measure	Value	Filename	SRR14108905_2.Fastq.gz	File type	Conventional base calls	Encoding	Sanger / Illumina 1.9	Total Sequences	21321329	Sequences flagged as poor quality	0	Sequence length	50	%GC	53
Measure	Value																																														
Filename	SRR14108905_1.Fastq.gz																																														
File type	Conventional base calls																																														
Encoding	Sanger / Illumina 1.9																																														
Total Sequences	21321329																																														
Sequences flagged as poor quality	0																																														
Sequence length	50																																														
%GC	52																																														
Measure	Value																																														
Filename	SRR14108905_2.Fastq.gz																																														
File type	Conventional base calls																																														
Encoding	Sanger / Illumina 1.9																																														
Total Sequences	21321329																																														
Sequences flagged as poor quality	0																																														
Sequence length	50																																														
%GC	53																																														

Fig 3i: SRR14108905_1

Fig 3j: SRR14108905_2

SRR14108905_1.Fastq.gz	SRR14108905_2.Fastq.gz	SRR14108906_1.Fastq.gz	SRR14108906_2.Fastq.gz	SRR14108905_1.Fastq.gz	SRR14108905_2.Fastq.gz	SRR14108906_1.Fastq.gz	SRR14108906_2.Fastq.gz																																								
<p>Basic Statistics</p> <p>Per base sequence quality</p> <p>Per tile sequence quality</p> <p>Per sequence quality scores</p> <p>Per base sequence content</p> <p>Per sequence GC content</p> <p>Per base N content</p> <p>Sequence Length Distribution</p> <p>Sequence Duplication Levels</p> <p>Overrepresented sequences</p> <p>Adapter Content</p> <p>Kmer Content</p>				<p>Basic sequence stats</p> <table border="1"> <thead> <tr> <th>Measure</th> <th>Value</th> </tr> </thead> <tbody> <tr> <td>Filename</td> <td>SRR14108906_1.Fastq.gz</td> </tr> <tr> <td>File type</td> <td>Conventional base calls</td> </tr> <tr> <td>Encoding</td> <td>Sanger / Illumina 1.9</td> </tr> <tr> <td>Total Sequences</td> <td>21711353</td> </tr> <tr> <td>Sequences flagged as poor quality</td> <td>0</td> </tr> <tr> <td>Sequence length</td> <td>50</td> </tr> <tr> <td>%GC</td> <td>52</td> </tr> </tbody> </table>				Measure	Value	Filename	SRR14108906_1.Fastq.gz	File type	Conventional base calls	Encoding	Sanger / Illumina 1.9	Total Sequences	21711353	Sequences flagged as poor quality	0	Sequence length	50	%GC	52	<p>Basic Statistics</p> <p>Per base sequence quality</p> <p>Per tile sequence quality</p> <p>Per sequence quality scores</p> <p>Per base sequence content</p> <p>Per sequence GC content</p> <p>Per base N content</p> <p>Sequence Length Distribution</p> <p>Sequence Duplication Levels</p> <p>Overrepresented sequences</p> <p>Adapter Content</p> <p>Kmer Content</p>				<p>Basic sequence stats</p> <table border="1"> <thead> <tr> <th>Measure</th> <th>Value</th> </tr> </thead> <tbody> <tr> <td>Filename</td> <td>SRR14108906_2.Fastq.gz</td> </tr> <tr> <td>File type</td> <td>Conventional base calls</td> </tr> <tr> <td>Encoding</td> <td>Sanger / Illumina 1.9</td> </tr> <tr> <td>Total Sequences</td> <td>21711353</td> </tr> <tr> <td>Sequences flagged as poor quality</td> <td>0</td> </tr> <tr> <td>Sequence length</td> <td>50</td> </tr> <tr> <td>%GC</td> <td>53</td> </tr> </tbody> </table>				Measure	Value	Filename	SRR14108906_2.Fastq.gz	File type	Conventional base calls	Encoding	Sanger / Illumina 1.9	Total Sequences	21711353	Sequences flagged as poor quality	0	Sequence length	50	%GC	53
Measure	Value																																														
Filename	SRR14108906_1.Fastq.gz																																														
File type	Conventional base calls																																														
Encoding	Sanger / Illumina 1.9																																														
Total Sequences	21711353																																														
Sequences flagged as poor quality	0																																														
Sequence length	50																																														
%GC	52																																														
Measure	Value																																														
Filename	SRR14108906_2.Fastq.gz																																														
File type	Conventional base calls																																														
Encoding	Sanger / Illumina 1.9																																														
Total Sequences	21711353																																														
Sequences flagged as poor quality	0																																														
Sequence length	50																																														
%GC	53																																														

Fig 3k: SRR14108906_1

Fig 3l: SRR14108906_2

SRR14108907_1.Fastq.gz	SRR14108907_2.Fastq.gz	SRR14108908_1.Fastq.gz	SRR14108908_2.Fastq.gz	SRR14108907_1.Fastq.gz	SRR14108907_2.Fastq.gz	SRR14108908_1.Fastq.gz	SRR14108908_2.Fastq.gz																																								
<p>Basic Statistics</p> <p>Per base sequence quality</p> <p>Per tile sequence quality</p> <p>Per sequence quality scores</p> <p>Per base sequence content</p> <p>Per sequence GC content</p> <p>Per base N content</p> <p>Sequence Length Distribution</p> <p>Sequence Duplication Levels</p> <p>Overrepresented sequences</p> <p>Adapter Content</p> <p>Kmer Content</p>				<p>Basic sequence stats</p> <table border="1"> <thead> <tr> <th>Measure</th> <th>Value</th> </tr> </thead> <tbody> <tr> <td>Filename</td> <td>SRR14108907_1.Fastq.gz</td> </tr> <tr> <td>File type</td> <td>Conventional base calls</td> </tr> <tr> <td>Encoding</td> <td>Sanger / Illumina 1.9</td> </tr> <tr> <td>Total Sequences</td> <td>18417670</td> </tr> <tr> <td>Sequences flagged as poor quality</td> <td>0</td> </tr> <tr> <td>Sequence length</td> <td>50</td> </tr> <tr> <td>%GC</td> <td>52</td> </tr> </tbody> </table>				Measure	Value	Filename	SRR14108907_1.Fastq.gz	File type	Conventional base calls	Encoding	Sanger / Illumina 1.9	Total Sequences	18417670	Sequences flagged as poor quality	0	Sequence length	50	%GC	52	<p>Basic Statistics</p> <p>Per base sequence quality</p> <p>Per tile sequence quality</p> <p>Per sequence quality scores</p> <p>Per base sequence content</p> <p>Per sequence GC content</p> <p>Per base N content</p> <p>Sequence Length Distribution</p> <p>Sequence Duplication Levels</p> <p>Overrepresented sequences</p> <p>Adapter Content</p> <p>Kmer Content</p>				<p>Basic sequence stats</p> <table border="1"> <thead> <tr> <th>Measure</th> <th>Value</th> </tr> </thead> <tbody> <tr> <td>Filename</td> <td>SRR14108907_2.Fastq.gz</td> </tr> <tr> <td>File type</td> <td>Conventional base calls</td> </tr> <tr> <td>Encoding</td> <td>Sanger / Illumina 1.9</td> </tr> <tr> <td>Total Sequences</td> <td>18417670</td> </tr> <tr> <td>Sequences flagged as poor quality</td> <td>0</td> </tr> <tr> <td>Sequence length</td> <td>50</td> </tr> <tr> <td>%GC</td> <td>53</td> </tr> </tbody> </table>				Measure	Value	Filename	SRR14108907_2.Fastq.gz	File type	Conventional base calls	Encoding	Sanger / Illumina 1.9	Total Sequences	18417670	Sequences flagged as poor quality	0	Sequence length	50	%GC	53
Measure	Value																																														
Filename	SRR14108907_1.Fastq.gz																																														
File type	Conventional base calls																																														
Encoding	Sanger / Illumina 1.9																																														
Total Sequences	18417670																																														
Sequences flagged as poor quality	0																																														
Sequence length	50																																														
%GC	52																																														
Measure	Value																																														
Filename	SRR14108907_2.Fastq.gz																																														
File type	Conventional base calls																																														
Encoding	Sanger / Illumina 1.9																																														
Total Sequences	18417670																																														
Sequences flagged as poor quality	0																																														
Sequence length	50																																														
%GC	53																																														

Fig 3m: SRR14108907_1

Fig 3n: SRR14108907_2

SRR14108907_1.Fastq.gz	SRR14108907_2.Fastq.gz	SRR14108908_1.Fastq.gz	SRR14108908_2.Fastq.gz	SRR14108907_1.Fastq.gz	SRR14108907_2.Fastq.gz	SRR14108908_1.Fastq.gz	SRR14108908_2.Fastq.gz																																
<p>Basic Statistics</p> <p>Per base sequence quality ✔</p> <p>Per tile sequence quality ✔</p> <p>Per sequence quality scores ✔</p> <p>Per base sequence content ⚠</p> <p>Per sequence GC content ⚠</p> <p>Per base N content ✔</p> <p>Sequence Length Distribution ✔</p> <p>Sequence Duplication Levels ✖</p> <p>Overrepresented sequences ⚠</p> <p>Adapter Content ✔</p> <p>Kmer Content ✖</p>				<p>Basic Statistics</p> <p>Per base sequence quality ✔</p> <p>Per tile sequence quality ✔</p> <p>Per sequence quality scores ✔</p> <p>Per base sequence content ✖</p> <p>Per sequence GC content ✔</p> <p>Per base N content ✔</p> <p>Sequence Length Distribution ✔</p> <p>Sequence Duplication Levels ✖</p> <p>Overrepresented sequences ⚠</p> <p>Adapter Content ✔</p> <p>Kmer Content ✖</p>																																			
<table border="1"> <thead> <tr> <th>Measure</th> <th>Value</th> </tr> </thead> <tbody> <tr> <td>Filename</td> <td>SRR14108908_1.Fastq.gz</td> </tr> <tr> <td>File type</td> <td>Conventional base calls</td> </tr> <tr> <td>Encoding</td> <td>Sanger / Illumina 1.9</td> </tr> <tr> <td>Total Sequences</td> <td>18761332</td> </tr> <tr> <td>Sequences flagged as poor quality</td> <td>0</td> </tr> <tr> <td>Sequence length</td> <td>50</td> </tr> <tr> <td>%GC</td> <td>52</td> </tr> </tbody> </table>				Measure	Value	Filename	SRR14108908_1.Fastq.gz	File type	Conventional base calls	Encoding	Sanger / Illumina 1.9	Total Sequences	18761332	Sequences flagged as poor quality	0	Sequence length	50	%GC	52	<table border="1"> <thead> <tr> <th>Measure</th> <th>Value</th> </tr> </thead> <tbody> <tr> <td>Filename</td> <td>SRR14108908_2.Fastq.gz</td> </tr> <tr> <td>File type</td> <td>Conventional base calls</td> </tr> <tr> <td>Encoding</td> <td>Sanger / Illumina 1.9</td> </tr> <tr> <td>Total Sequences</td> <td>18761332</td> </tr> <tr> <td>Sequences flagged as poor quality</td> <td>0</td> </tr> <tr> <td>Sequence length</td> <td>50</td> </tr> <tr> <td>%GC</td> <td>53</td> </tr> </tbody> </table>				Measure	Value	Filename	SRR14108908_2.Fastq.gz	File type	Conventional base calls	Encoding	Sanger / Illumina 1.9	Total Sequences	18761332	Sequences flagged as poor quality	0	Sequence length	50	%GC	53
Measure	Value																																						
Filename	SRR14108908_1.Fastq.gz																																						
File type	Conventional base calls																																						
Encoding	Sanger / Illumina 1.9																																						
Total Sequences	18761332																																						
Sequences flagged as poor quality	0																																						
Sequence length	50																																						
%GC	52																																						
Measure	Value																																						
Filename	SRR14108908_2.Fastq.gz																																						
File type	Conventional base calls																																						
Encoding	Sanger / Illumina 1.9																																						
Total Sequences	18761332																																						
Sequences flagged as poor quality	0																																						
Sequence length	50																																						
%GC	53																																						

Fig 3o: SRR14108908_1

Fig 3p: SRR14108908_2

SRR14108909_1.Fastq.gz	SRR14108909_2.Fastq.gz	SRR14108910_1.Fastq.gz	SRR14108910_2.Fastq.gz	SRR14108909_1.Fastq.gz	SRR14108909_2.Fastq.gz	SRR14108910_1.Fastq.gz	SRR14108910_2.Fastq.gz																																
<p>Basic Statistics</p> <p>Per base sequence quality ✔</p> <p>Per tile sequence quality ✖</p> <p>Per sequence quality scores ✔</p> <p>Per base sequence content ✖</p> <p>Per sequence GC content ⚠</p> <p>Per base N content ✔</p> <p>Sequence Length Distribution ✔</p> <p>Sequence Duplication Levels ✖</p> <p>Overrepresented sequences ⚠</p> <p>Adapter Content ✔</p> <p>Kmer Content ✖</p>				<p>Basic Statistics</p> <p>Per base sequence quality ✔</p> <p>Per tile sequence quality ✔</p> <p>Per sequence quality scores ✔</p> <p>Per base sequence content ✖</p> <p>Per sequence GC content ✔</p> <p>Per base N content ✔</p> <p>Sequence Length Distribution ✔</p> <p>Sequence Duplication Levels ✖</p> <p>Overrepresented sequences ⚠</p> <p>Adapter Content ✔</p> <p>Kmer Content ✖</p>																																			
<table border="1"> <thead> <tr> <th>Measure</th> <th>Value</th> </tr> </thead> <tbody> <tr> <td>Filename</td> <td>SRR14108909_1.Fastq.gz</td> </tr> <tr> <td>File type</td> <td>Conventional base calls</td> </tr> <tr> <td>Encoding</td> <td>Sanger / Illumina 1.9</td> </tr> <tr> <td>Total Sequences</td> <td>19914040</td> </tr> <tr> <td>Sequences flagged as poor quality</td> <td>0</td> </tr> <tr> <td>Sequence length</td> <td>50</td> </tr> <tr> <td>%GC</td> <td>52</td> </tr> </tbody> </table>				Measure	Value	Filename	SRR14108909_1.Fastq.gz	File type	Conventional base calls	Encoding	Sanger / Illumina 1.9	Total Sequences	19914040	Sequences flagged as poor quality	0	Sequence length	50	%GC	52	<table border="1"> <thead> <tr> <th>Measure</th> <th>Value</th> </tr> </thead> <tbody> <tr> <td>Filename</td> <td>SRR14108909_2.Fastq.gz</td> </tr> <tr> <td>File type</td> <td>Conventional base calls</td> </tr> <tr> <td>Encoding</td> <td>Sanger / Illumina 1.9</td> </tr> <tr> <td>Total Sequences</td> <td>19914040</td> </tr> <tr> <td>Sequences flagged as poor quality</td> <td>0</td> </tr> <tr> <td>Sequence length</td> <td>50</td> </tr> <tr> <td>%GC</td> <td>54</td> </tr> </tbody> </table>				Measure	Value	Filename	SRR14108909_2.Fastq.gz	File type	Conventional base calls	Encoding	Sanger / Illumina 1.9	Total Sequences	19914040	Sequences flagged as poor quality	0	Sequence length	50	%GC	54
Measure	Value																																						
Filename	SRR14108909_1.Fastq.gz																																						
File type	Conventional base calls																																						
Encoding	Sanger / Illumina 1.9																																						
Total Sequences	19914040																																						
Sequences flagged as poor quality	0																																						
Sequence length	50																																						
%GC	52																																						
Measure	Value																																						
Filename	SRR14108909_2.Fastq.gz																																						
File type	Conventional base calls																																						
Encoding	Sanger / Illumina 1.9																																						
Total Sequences	19914040																																						
Sequences flagged as poor quality	0																																						
Sequence length	50																																						
%GC	54																																						

Fig 2q: SRR14108909_1

Fig 2r: SRR14108909_2

SRR14108909_1.Fastq.gz	SRR14108909_2.Fastq.gz	SRR14108910_1.Fastq.gz	SRR14108910_2.Fastq.gz	SRR14108909_1.Fastq.gz	SRR14108909_2.Fastq.gz	SRR14108910_1.Fastq.gz	SRR14108910_2.Fastq.gz																																
<p>Basic Statistics</p> <p>Per base sequence quality ✔</p> <p>Per tile sequence quality ✔</p> <p>Per sequence quality scores ✔</p> <p>Per base sequence content ✖</p> <p>Per sequence GC content ⚠</p> <p>Per base N content ✔</p> <p>Sequence Length Distribution ✔</p> <p>Sequence Duplication Levels ✖</p> <p>Overrepresented sequences ⚠</p> <p>Adapter Content ✔</p> <p>Kmer Content ✖</p>				<p>Basic Statistics</p> <p>Per base sequence quality ✔</p> <p>Per tile sequence quality ✔</p> <p>Per sequence quality scores ✔</p> <p>Per base sequence content ✖</p> <p>Per sequence GC content ✔</p> <p>Per base N content ✔</p> <p>Sequence Length Distribution ✔</p> <p>Sequence Duplication Levels ✖</p> <p>Overrepresented sequences ⚠</p> <p>Adapter Content ✔</p> <p>Kmer Content ✖</p>																																			
<table border="1"> <thead> <tr> <th>Measure</th> <th>Value</th> </tr> </thead> <tbody> <tr> <td>Filename</td> <td>SRR14108910_1.Fastq.gz</td> </tr> <tr> <td>File type</td> <td>Conventional base calls</td> </tr> <tr> <td>Encoding</td> <td>Sanger / Illumina 1.9</td> </tr> <tr> <td>Total Sequences</td> <td>20388666</td> </tr> <tr> <td>Sequences flagged as poor quality</td> <td>0</td> </tr> <tr> <td>Sequence length</td> <td>50</td> </tr> <tr> <td>%GC</td> <td>52</td> </tr> </tbody> </table>				Measure	Value	Filename	SRR14108910_1.Fastq.gz	File type	Conventional base calls	Encoding	Sanger / Illumina 1.9	Total Sequences	20388666	Sequences flagged as poor quality	0	Sequence length	50	%GC	52	<table border="1"> <thead> <tr> <th>Measure</th> <th>Value</th> </tr> </thead> <tbody> <tr> <td>Filename</td> <td>SRR14108910_2.Fastq.gz</td> </tr> <tr> <td>File type</td> <td>Conventional base calls</td> </tr> <tr> <td>Encoding</td> <td>Sanger / Illumina 1.9</td> </tr> <tr> <td>Total Sequences</td> <td>20388666</td> </tr> <tr> <td>Sequences flagged as poor quality</td> <td>0</td> </tr> <tr> <td>Sequence length</td> <td>50</td> </tr> <tr> <td>%GC</td> <td>53</td> </tr> </tbody> </table>				Measure	Value	Filename	SRR14108910_2.Fastq.gz	File type	Conventional base calls	Encoding	Sanger / Illumina 1.9	Total Sequences	20388666	Sequences flagged as poor quality	0	Sequence length	50	%GC	53
Measure	Value																																						
Filename	SRR14108910_1.Fastq.gz																																						
File type	Conventional base calls																																						
Encoding	Sanger / Illumina 1.9																																						
Total Sequences	20388666																																						
Sequences flagged as poor quality	0																																						
Sequence length	50																																						
%GC	52																																						
Measure	Value																																						
Filename	SRR14108910_2.Fastq.gz																																						
File type	Conventional base calls																																						
Encoding	Sanger / Illumina 1.9																																						
Total Sequences	20388666																																						
Sequences flagged as poor quality	0																																						
Sequence length	50																																						
%GC	53																																						

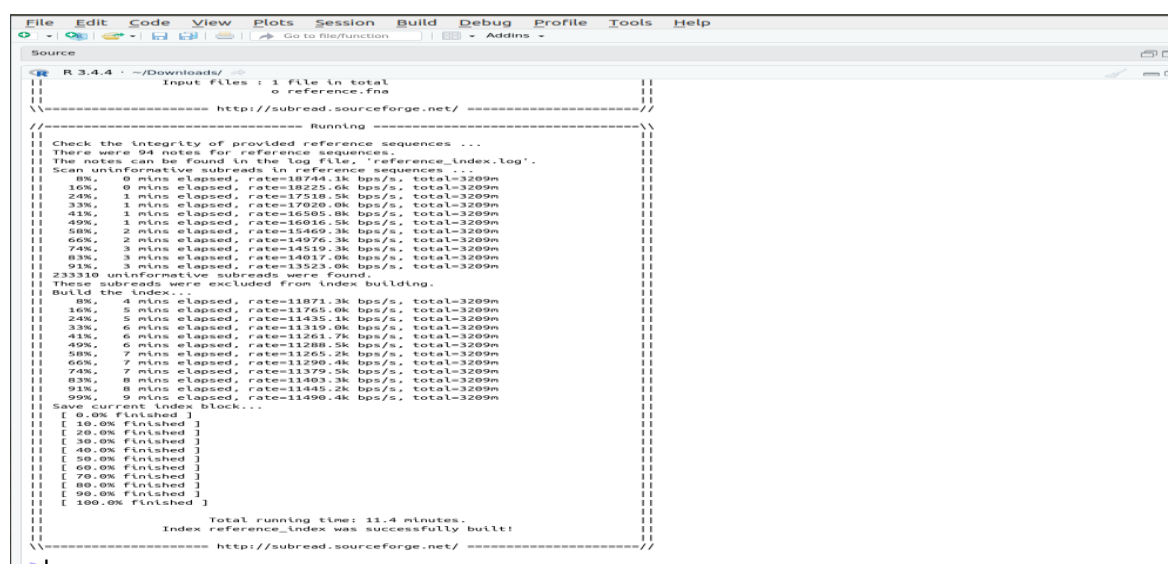
Fig 3s: SRR14108910_1

Fig 3t: SRR14108910_2

Figure 3 (Figure 3a to figure 3t) shows the quality control by using Fast QC of the following samples.

Building the reference index by RStudio

The Human reference genome of the human was downloaded for building a reference index for alignment and mapping of the sequence from NCBI (National Center for Biotechnology Information), the reference index was built using RStudio, using the Rsubread package and the base name was given as “chr1_mm10”, as shown in Figure 4. Genome indexing can be described in a similar way to book indexing. If you want to know on which page a particular word appears or where a chapter begins, it's much more efficient / faster to look it up in a ready-made index than to look it up until you find each page in the book. The same is true for linear. Indexes allow aligners to narrow down potential origins of query sequences in the genome, saving both time and memory.



```
File Edit Code View Plots Session Build Debug Profile Tools Help
Source
R 3.4.4 ~/Downloads/
Input files : 1 file in total
              o reference.fna
----- http://subread.sourceforge.net/ -----
//----- Running -----//
||
|| Check the integrity of provided reference sequences ...
|| There were 94 notes for reference sequences.
|| The notes can be found in the log file, 'reference_index.log'.
|| Scan uninformative subreads in reference sequences ...
|| 0%, 0 mins elapsed, rate=18754.1k bps/s, total=3209m
|| 16%, 0 mins elapsed, rate=18225.6k bps/s, total=3209m
|| 24%, 1 mins elapsed, rate=17518.5k bps/s, total=3209m
|| 33%, 1 mins elapsed, rate=17020.0k bps/s, total=3209m
|| 41%, 1 mins elapsed, rate=16505.8k bps/s, total=3209m
|| 49%, 1 mins elapsed, rate=16016.5k bps/s, total=3209m
|| 58%, 2 mins elapsed, rate=15469.3k bps/s, total=3209m
|| 66%, 2 mins elapsed, rate=14976.3k bps/s, total=3209m
|| 74%, 3 mins elapsed, rate=14519.3k bps/s, total=3209m
|| 83%, 3 mins elapsed, rate=14037.0k bps/s, total=3209m
|| 91%, 3 mins elapsed, rate=13523.0k bps/s, total=3209m
|| 23310 uninformative subreads were found.
|| These subreads were excluded from index building.
|| Build the index...
|| 0%, 4 mins elapsed, rate=11871.3k bps/s, total=3209m
|| 16%, 5 mins elapsed, rate=11765.0k bps/s, total=3209m
|| 24%, 5 mins elapsed, rate=11435.1k bps/s, total=3209m
|| 33%, 6 mins elapsed, rate=11319.0k bps/s, total=3209m
|| 41%, 6 mins elapsed, rate=11261.7k bps/s, total=3209m
|| 49%, 6 mins elapsed, rate=11288.5k bps/s, total=3209m
|| 58%, 7 mins elapsed, rate=11205.2k bps/s, total=3209m
|| 66%, 7 mins elapsed, rate=11296.4k bps/s, total=3209m
|| 74%, 7 mins elapsed, rate=11379.5k bps/s, total=3209m
|| 83%, 8 mins elapsed, rate=11403.3k bps/s, total=3209m
|| 91%, 8 mins elapsed, rate=11445.2k bps/s, total=3209m
|| 99%, 9 mins elapsed, rate=11496.4k bps/s, total=3209m
|| Save current index block...
|| [ 0.0% finished ]
|| [ 10.0% finished ]
|| [ 20.0% finished ]
|| [ 30.0% finished ]
|| [ 40.0% finished ]
|| [ 50.0% finished ]
|| [ 60.0% finished ]
|| [ 70.0% finished ]
|| [ 80.0% finished ]
|| [ 90.0% finished ]
|| [ 100.0% finished ]
||
|| Total running time: 11.4 minutes.
|| Index reference_index was successfully built!
||----- http://subread.sourceforge.net/ -----
=|
```

Fig 4: Reference index was built using Rsubread in RStudio

Alignment using Rsubread

Then, the alignment was done using pair-end sequencing alignment, by RStudio and by taking two FASTA files as input, the output files are in BAM format using the reference index, Rsubread can be used for many processes like- Alignment, quantification, and analysis of RNA sequencing data (including both bulk RNA-seq and scRNA-seq) and DNA sequencing data (including ATAC-seq, ChIP-seq, WGS, WES, etc). Includes functionality for reading mapping, read counting, SNP calling, structural variant detection, and gene fusion discovery. Can be applied to all major sequencing technologies and to both short

and long sequence reads (Liao *et al.*, 2019) The following results were obtained after alignment; the list of files is shown in figure 5.

		Samples	NumTotal	NumMapped	PropMapped
1	/home/rakesh/Downloads/SRR14108901.fastq.gz.subread.BAM	19521208	19371471	0.992330	
2	/home/rakesh/Downloads/SRR14108902.fastq.gz.subread.BAM	19070351	18926496	0.992457	
3	/home/rakesh/Downloads/SRR14108903.fastq.gz.subread.BAM	17613371	17495671	0.993318	
4	/home/rakesh/Downloads/SRR14108904.fastq.gz.subread.BAM	17309068	17193851	0.993344	
5	/home/rakesh/Downloads/SRR14108905.fastq.gz.subread.BAM	21321329	21188041	0.993749	
6	/home/rakesh/Downloads/SRR14108906.fastq.gz.subread.BAM	21711353	21573056	0.993630	
7	/home/rakesh/Downloads/SRR14108907.fastq.gz.subread.BAM	18417670	18222723	0.989415	
8	/home/rakesh/Downloads/SRR14108908.fastq.gz.subread.BAM	18761332	18561480	0.989348	
9	/home/rakesh/Downloads/SRR14108909.fastq.gz.subread.BAM	19914040	19699492	0.989226	
10	/home/rakesh/Downloads/SRR14108910.fastq.gz.subread.BAM	20388686	20167279	0.989141	

Fig 5: The list of BAM files after alignment

Feature Count using Rsubread in terminal

After the alignment, we got one BAM file instead of two FASTA files and then the feature count was done in order to get the count table, it was done by using Rsubread in the Ubuntu terminal and the output was in the form of the count.out file. The full analysis is shown in the figure 6.

```

=====
SUBREAD
v2.0.1

===== featureCounts setting =====

Input files : 10 BAM files
  o SRR14108901_1.fastq.gz.subread.BAM
  o SRR14108902_1.fastq.gz.subread.BAM
  o SRR14108903_1.fastq.gz.subread.BAM
  o SRR14108904_1.fastq.gz.subread.BAM
  o SRR14108905_1.fastq.gz.subread.BAM
  o SRR14108906_1.fastq.gz.subread.BAM
  o SRR14108907_1.fastq.gz.subread.BAM
  o SRR14108908_1.fastq.gz.subread.BAM
  o SRR14108909_1.fastq.gz.subread.BAM
  o SRR14108910_1.fastq.gz.subread.BAM

Output file : counts.txt
  Summary : counts.txt.summary
  Annotation : annotation.gtf (GTF)
  Dir for temp files : ./

Threads : 1
  Level : meta-feature level
  Paired-end : no
  Multimaping reads : not counted
  Multi-overlapping reads : not counted
  MtN overlapping bases : 1

===== Running =====

Load annotation file annotation.gtf ...
  Features : 2161991
  Meta-Features : 47388
  Chromosomes/contigs : 302

Process BAM file SRR14108901_1.fastq.gz.subread.BAM...
  WARNING: Paired-end reads were found.
  Total alignments : 3904246
  Successfully assigned alignments : 31460380 (80.6%)
  Running time : 0.58 minutes

Process BAM file SRR14108902_1.fastq.gz.subread.BAM...
  WARNING: Paired-end reads were found.
  Total alignments : 38140702
  Successfully assigned alignments : 30756063 (80.6%)
  Running time : 0.56 minutes

Process BAM file SRR14108903_1.fastq.gz.subread.BAM...
  WARNING: Paired-end reads were found.
  Total alignments : 35226742
  Successfully assigned alignments : 28524383 (81.0%)
  Running time : 0.51 minutes

Process BAM file SRR14108904_1.fastq.gz.subread.BAM...
  WARNING: Paired-end reads were found.
  Total alignments : 34618136
  Successfully assigned alignments : 28042929 (81.0%)
  Running time : 0.51 minutes

Process BAM file SRR14108905_1.fastq.gz.subread.BAM...
  WARNING: Paired-end reads were found.
  Total alignments : 42642658
  Successfully assigned alignments : 34522882 (81.0%)
  Running time : 0.62 minutes

Process BAM file SRR14108906_1.fastq.gz.subread.BAM...
  WARNING: Paired-end reads were found.
  Total alignments : 43422706
  Successfully assigned alignments : 35138341 (80.9%)
  Running time : 0.63 minutes

Process BAM file SRR14108907_1.fastq.gz.subread.BAM...
  WARNING: Paired-end reads were found.
  Total alignments : 43422706
  Successfully assigned alignments : 29075625 (78.9%)
  Running time : 0.55 minutes

Process BAM file SRR14108908_1.fastq.gz.subread.BAM...
  WARNING: Paired-end reads were found.
  Total alignments : 37522664
  Successfully assigned alignments : 29597023 (78.9%)
  Running time : 0.52 minutes

Process BAM file SRR14108909_1.fastq.gz.subread.BAM...
  WARNING: Paired-end reads were found.
  Total alignments : 39828080
  Successfully assigned alignments : 31795944 (79.8%)
  Running time : 0.56 minutes

Process BAM file SRR14108910_1.fastq.gz.subread.BAM...
  WARNING: Paired-end reads were found.
  Total alignments : 40777372
  Successfully assigned alignments : 32540488 (79.8%)
  Running time : 0.58 minutes

write the final count table.
write the read assignment summary.

Summary of counting results can be found in file "counts.txt.summary"
=====

```

Fig 6: Feature Count using Rsubread

The count data are structured as a table, which reports the number of sequence fragments assigned to each gene for each sample, the count data were further filtered for null, NA, and negative values in the table, as these values show errors in further steps. The count data output for 10 samples were 47895, but after filtering the negative values, NULL values, NA values and zero values, only 7322 reading were left for further analysis of Differentially Expressed Genes. Feature Count is a general-purpose read summarization function, which assigns to the genomic features (or meta-features) the mapped reads that were generated from genomic DNA and RNA sequencing. (<https://www.rdocumentation.org/packages/Rsubread/versions/1.22.2/topics/featureCounts>)

RNA-seq reads may be aligned to the transcriptome rather than the genome. In this case, there can be hundreds of thousands of transcripts, and each transcript becomes a reference sequence. featureCounts supports thread-specific read counts when thread-specific information is provided (Yang *et al.*, 2014). The output table of the first 20 output of the count table is shown in table 1.

Table 1: Table of first 20 output after feature count of the sequence data

Geneid	1	2	3	4	5	6	7	8	9	10
A4GALT	552	569	462	509	629	639	527	542	598	585
AADAT	103	99	72	102	117	102	36	43	42	41
AAMDC	116	110	120	86	115	118	107	88	113	111
AAR2	728	776	664	613	794	809	871	974	851	887
AARS2	476	531	529	491	516	563	335	321	331	332
AASDH	261	249	248	210	277	313	138	150	152	181
AATBC	39	60	51	37	69	52	65	105	63	65
AATF	668	598	602	606	744	839	539	541	595	638
ABALON	65	62	64	55	68	71	64	50	47	57
ABAT	237	304	230	262	251	343	144	179	206	185
ABCA11P	104	92	69	93	112	113	86	123	116	132
ABCA2	616	639	523	608	657	637	198	196	329	336
ABCA5	838	822	739	766	949	940	952	957	882	961
ABCB10	672	696	620	593	697	718	324	405	478	482
ABCB6	707	768	629	573	752	818	248	236	248	276
ABCB7	569	495	490	435	606	627	505	454	489	502
ABCB9	272	300	283	289	310	375	99	98	134	126
ABCC2	84	75	68	36	90	81	32	36	57	29
ABCC4	511	481	415	421	517	489	372	322	360	383
ABCD1	194	211	178	179	211	234	274	232	260	310
ABCD4	590	599	526	528	676	602	545	578	556	586

Differentially Expressed Genes Analysis

Differential expression analysis means taking normalized read count data and performing statistical analysis to discover quantitative changes in expression levels between experimental groups.

The differentially Expressed genes analysis was done in RStudio using package DESeq2, the following steps were followed, firstly the tables are converted to matrix, then the conditions are assigned to the data, the data was then loaded to DESeq pipeline and different types of plots and graphs were obtained according to the need of the analysis, like-dispersion plot, heatmap, scatter plot, histogram, MA plot, volcano plot, etc.

RESULT AND DISCUSSION

Volcano Plot

Another common and interesting comparison between the two treatment conditions is the adjusted P-value and log fold change. This figure 7 is called a volcano plot because it resembles an exploding volcano, with clusters of data points near the origin and the fanning effect moving away from its central location. The volcanic plot shows the statistical significance of the difference to the magnitude of the difference between the individual genes compared. Usually indicated by a fold change of negative base 10log or base 2log, respectively. The P-value undergoes a negative transformation, so the higher the data point along the y-axis, the smaller the P-value. Volcano graphs are generally considered to be statistically differentially expressed based on the adjusted P value of the difference between treatments, including some threshold indicators of the adjusted P value. Indicates the gene to be used. Changes in log multiples along the x-axis show a clearer difference in extrema, and data points close to 0 represent genes with similar or same mean expression levels. In the case of volcanic areas, as the name implies, it is expected to be quite widespread. The wide dispersal indicates two treatment groups with significant differences in gene expression. It is quite rare for a volcano plot to have almost or all data points gathered near the origin.

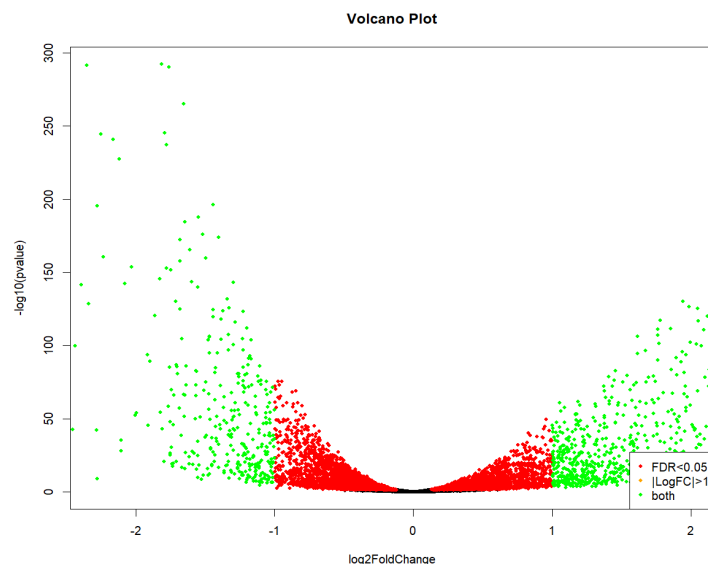


Fig 7: Volcano Plot generated from DESeq2 Dataset

MA Plot

The MA chart can only compare two treatment conditions at a time. However, all pairwise comparisons in this figure 8 can be combined in a matrix format to provide all possible combinations at once. Each cell represents a particular comparison, shown cell by cell or at the intersection of rows and columns. This visualization allows the user to view all pairwise fold change comparisons and average manifestations at once. In addition, this method allows direct comparison of pairwise treatment comparisons. It provides an approach for determining which treatment comparisons are more or less similar in terms of both fold change changes and mean expression levels. Like other matrix options, this process allows the user to visualize all treatment-based comparisons in one diagram.

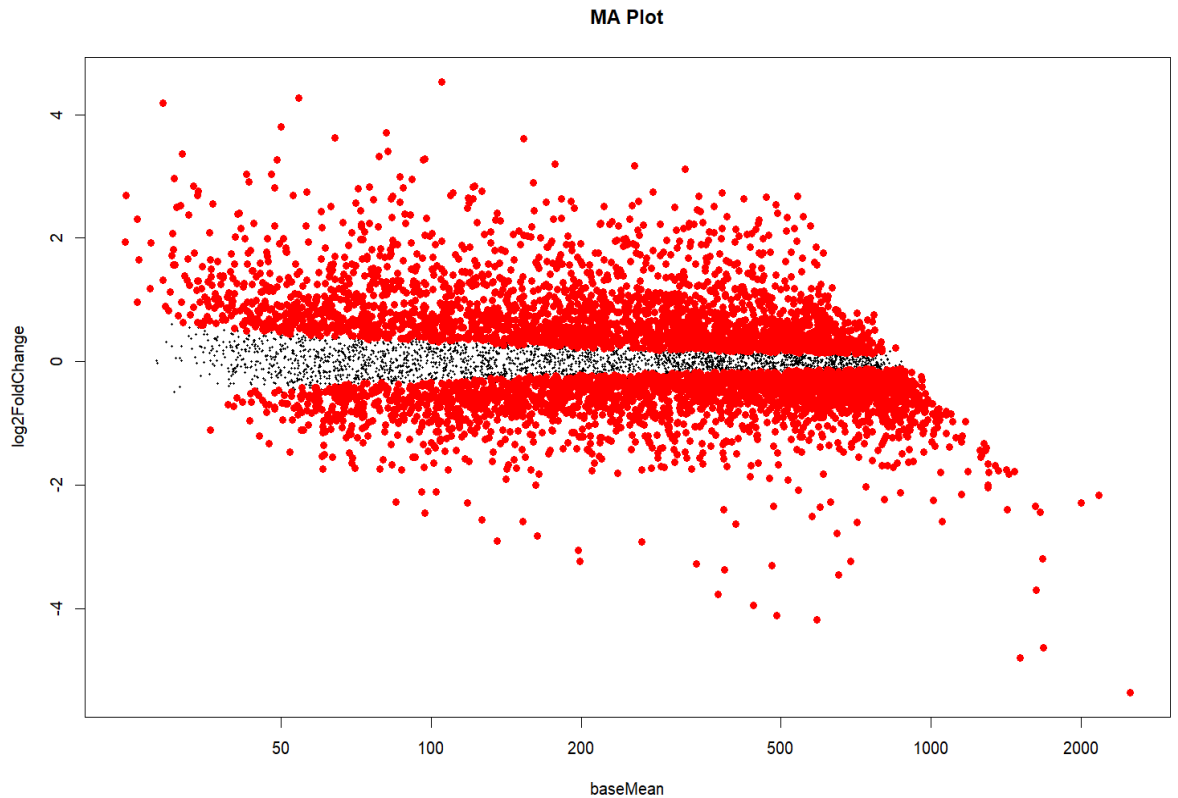


Fig 8: MA Plot generated from DESeq2 dataset

Heatmap

By comparison, you can also use a heatmap based on the number of DEG's to summarize the same information. Using a color spectrum based on the magnitude of the DEG count, the DEG heatmap can provide an easy way to read and interpret. For a DEG heatmap, each cell represents the number of DSNs in that particular intersecting row and column. Arrangements along the selected color spectrum, provide a visual indication of magnitude. Treatment group. The DEG heatmap has obvious drawbacks in terms of redundancy. For the three factor levels, this figure 9 is a good representation of the data. However, increasing the number of factor levels will generate redundant cells. Cells are usually left blank to avoid misleading the user. This method is counterproductive because it requires more effort to interpret the information efficiently. As the number of factor levels increases, the usefulness of this type of visualization diminishes and is recommended only for some factor levels. According to the heatmap, the white color shows the upregulated genes while the black color shows the downregulated genes.

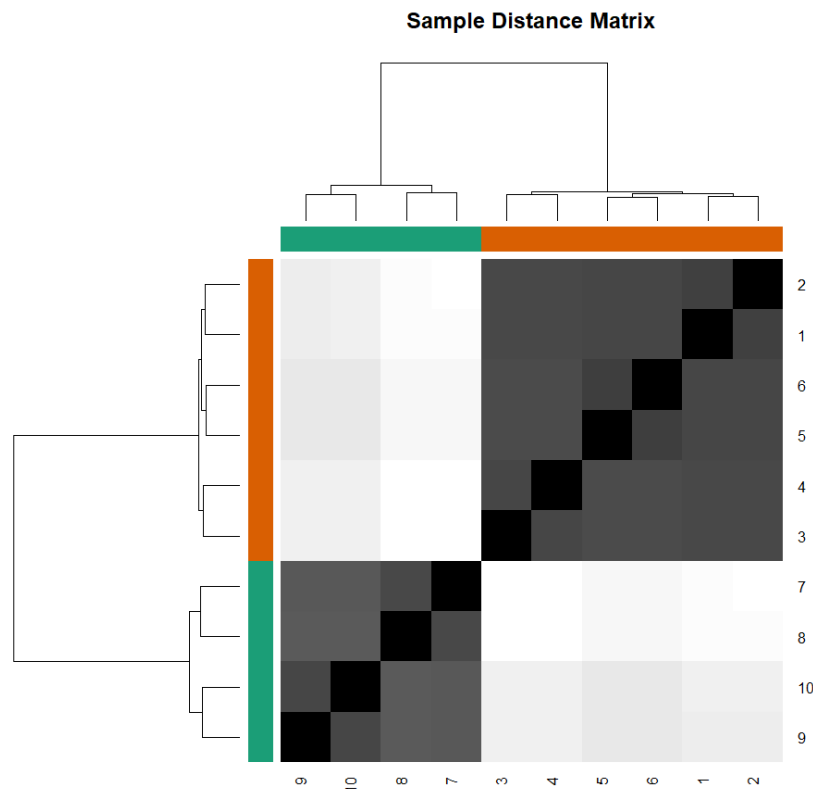


Fig 9: Heatmap generated from DESeq2 dataset

Dispersion Plot

Another relatively simple visualization method associated with Tier 1 is to compare expression levels between two samples or two treatment groups. This comparison is typically visualized using a scatter plot. Each data point represents a single gene and its placement indicates the average expression level for each of the two treatments. A scatter plot implemented in this way can be used to make a larger comparison between the two treatment groups. The axes represent the expression levels for each category, so the data points along the diagonal show similar expression levels from both groups. Data points above or below the diagonal indicate higher or lower expression levels of factor levels on the y-axis compared to factor levels on the x-axis, respectively. Considering this scatter plot as a whole, clustering of all data points along the diagonal shows two samples or treatments with very similar expression patterns across all genes, with the spread of data points from the diagonal. Larger values indicate dissimilar expression levels. Hence, the figure 10 shows that the gene is negatively regulated.

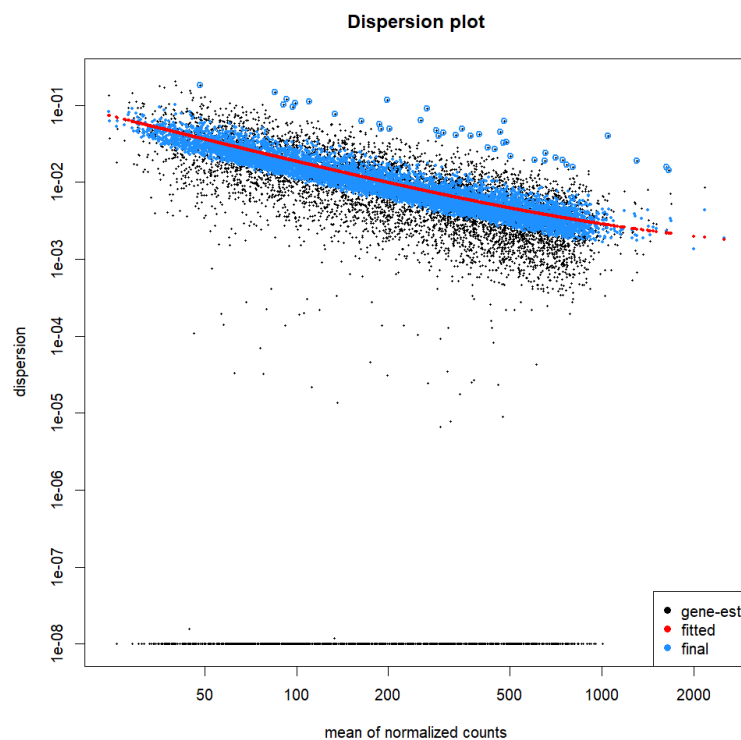


Fig 10: Dispersion Plot generated from DESeq2 dataset

PCA Biplot

The PCA Biplot also known as Principal Components Analysis Biplot is a two-dimensional chart that represents the relationship between the rows and columns. Hence, in this case the PCA Biplot is the representation of the relationship of the rows and columns of the count data in DEG as shown in figure 11.

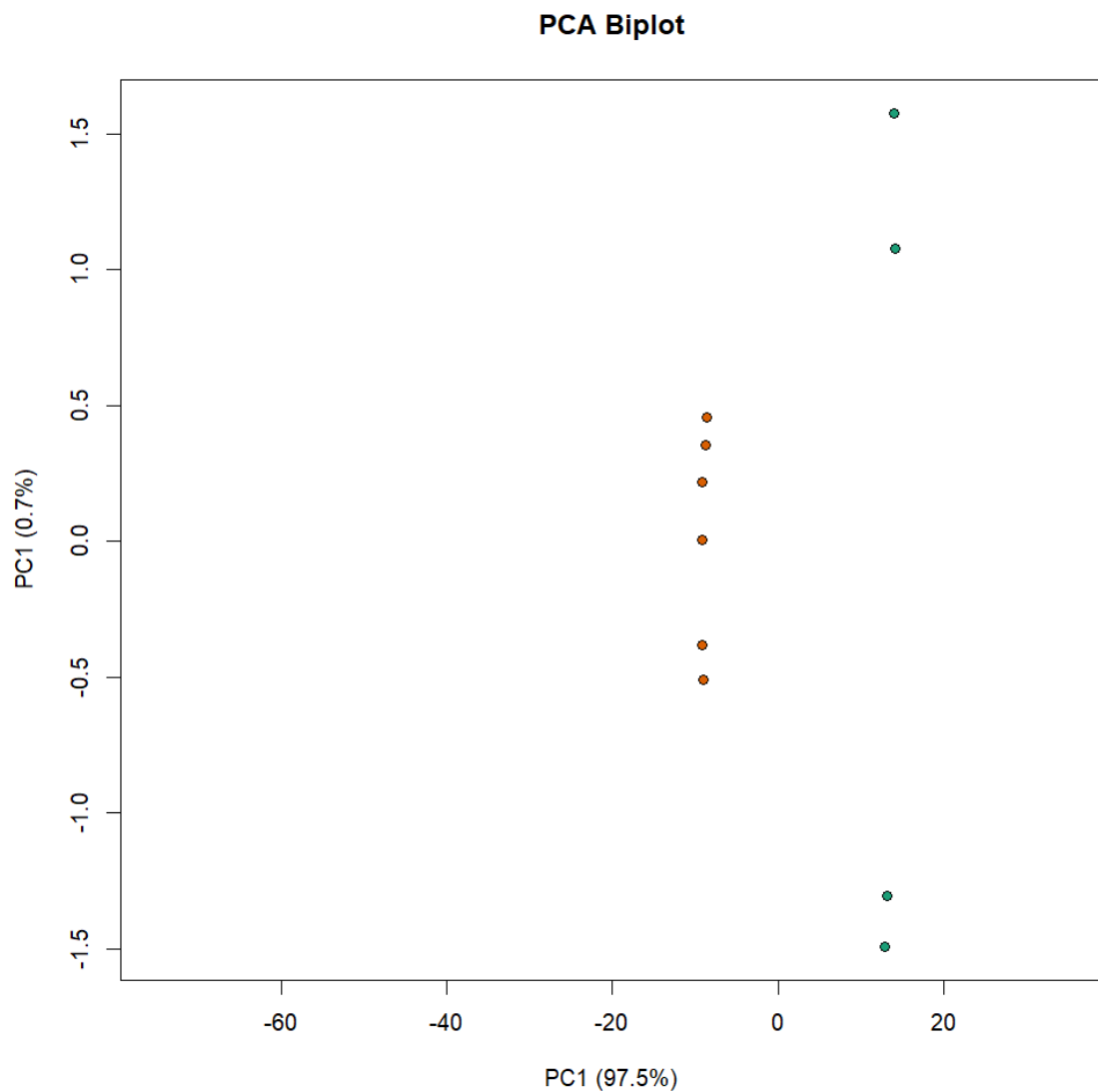


Fig 11: PCA Biplot generated from DESeq2 dataset

The table 2 shows the Gene id, base mean, log 2 fold change and p value of the samples, after differentially expressed gene analysis, it shows the first 20 output of the following table.

Table 2: First 20 output of the fold change and p value of the samples

Geneid	baseMean	log2FoldChange	stat	pvalue
A4GALT	442.1256034	-3.943573177	-40.2047	0
AADAT	591.276907	-4.187242559	-48.27	0
AAMDC	692.6433869	-3.231583464	-40.0586	0
AAR2	2512.089466	-5.363348764	-86.0756	0
AARS2	1511.415927	-4.801017016	-69.6296	0
AASDH	1683.021249	-4.626947669	-66.3029	0
AATBC	1678.557083	-3.194063629	-51.0615	0
AATF	1152.228924	-2.152591589	-39.5843	0
ABALON	1422.413078	-2.402810001	-43.0803	0
ABAT	1303.515258	-2.035337695	-40.3761	0
ABCA11P	1624.793213	-2.347929422	-45.0174	0
ABCA2	2003.794664	-2.286300956	-54.848	0
ABCA5	1437.488345	-1.815341826	-36.5894	4.21E-293
ABCB10	1053.90493	-2.588868083	-36.5871	4.59E-293
ABCB6	599.8269529	-2.352049668	-36.5403	2.55E-292
ABCB7	1365.420047	-1.760369103	-36.4537	6.01E-291
ABCB9	490.8532863	-4.110760563	-36.1429	4.82E-286
ABCC2	711.9125037	-2.60904678	-35.772	3.01E-280
ABCC4	386.9048784	-3.374336166	-35.011	1.53E-268
ABCD1	1302.18103	-1.6523181	-34.8324	7.86E-266
ABCD4	974.7166611	3.770987278	33.5146	3.96E-246

CONCLUSION

In this study, I have learned to analyse the RNA Seq data of skin disease psoriasis by using R. In this study, I have learned to check the quality of the data using Fast QC, then reference index was built for human reference genome, followed by alignment was done for pair end sequence using Rsubread package. Feature count was done to get the count data of the sample sequence. Then, differentially expressed gene analysis was done with the help of count data. Results were generated in the form of volcano plot, MA plot, heatmap, dispersion plot and PCA Biplot.

Our results suggest negative correlation through the expression levels of psoriasis. It highlights that the samples regulated by TWEAK and TNF inhibit the expression of psoriasis genes. This indicates the use of TWEAK and TNF as a possible treatment for psoriasis.

DEG is often used to determine genotype differences between two or more cellular states to support studies based on specific hypotheses. Interpretation of this information can greatly benefit from the graphic display of the result file. Tier 1 functions provide relatively basic levels of information, including read count distributions, pairwise levels, and those used to visualize DEG counts, while Tier 2 functions provide average level, use additional metrics such as multiple changes, P-values-provide more detailed and informative visualizations. Box plots, violin plots, dot plots, and read count histograms provide insight into the distribution of read counts for each sample or processing group. Scatter plots allow users to visualize the overall similarity of expression levels by showing the expression levels of each gene in the two selected treatments or samples. The DEG histogram and heatmap directly represent the number of DEGs in each comparison. MA and volcano charts are useful for showing relative expression levels, changes in log multiples, and adjusted P-values. Although not applicable to all users, 4-way plots can provide a higher level of detail by including a third treatment group or sample as a relative or control group.

BIBLIOGRAPHY

Agarwal A, Koppstein D, Rozowsky J, Sboner A, Habegger L, Hillier LW, Sasidharan R, Reinke V, Waterston RH, Gerstein M. (2010) Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. *BMC Genomics*. Ahn R, Gupta R, Lai K, Chopra N, Arron ST, Liao W. (2010) Network analysis of psoriasis reveals biological pathways and roles for coding and long non-coding RNAs. *BMC Genomics*. 2016; 17:841.

Albanesi C, Madonna S, Gisondi P, Girolomoni G, (2018) The Interplay Between Keratinocytes and Immune Cells in the Pathogenesis of Psoriasis. *Front Immunol* 9, 1549.

Anders S, Huber W. (2010) Differential expression analysis for sequence count data. *Genome Biol.* 2010;11: R106.

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Nature Precedings*, 1-1.

Auer, P. L., & Doerge, R. W. (2011). A two-stage Poisson model for testing RNA-seq data. *Statistical applications in genetics and molecular biology*, 10(1).

Auer, P. L., Srivastava, S., & Doerge, R. W. (2012). Differential expression—the next generation and beyond. *Briefings in functional genomics*, 11(1), 57-62.

Beane J, Vick J, Schembri F, Anderlind C, Gower A, Campbell J, Luo L, Zhang XH, Xiao J, Alekseyev YO, (2011) Characterizing the impact of smoking and lung cancer on the airway transcriptome using RNA-Seq. *Cancer Prev Res (Phila)* 2011;4:803–817.

Benhadou F, Mintoff D, Del Marmol V, (2019) Psoriasis: Keratinocytes or Immune Cells - Which Is the Trigger? *Dermatology* 235, 91–100.

Bilgiç Ö, Sivrikaya A, Toker A, Ünlü A, Altınyazar C, (2016) Serum levels of TWEAK in patients with psoriasis vulgaris. *Cytokine* 77, 10–13.

Bird TG, Lu WY, Boulter L, Gordon-Keylock S, Ridgway RA, Williams MJ, Taube J, Thomas JA, Wojtacha D, Gambardella A, Sansom OJ, Iredale JP, Forbes SJ, (2013) Bone marrow injection stimulates hepatic ductular reactions in the absence of injury via macrophage-mediated TWEAK signaling. *Proc Natl Acad Sci U S A* 110, 6542–6547.

Bradford, J. R., Hey, Y., Yates, T., Li, Y., Pepper, S. D., & Miller, C. J. (2010). A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling. *BMC genomics*, 11(1), 1-12.

Bradford, J. R., Hey, Y., Yates, T., Li, Y., Pepper, S. D., & Miller, C. J. (2010). A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling. *BMC genomics*, 11(1), 1-12.

Burkly LC, (2014) TWEAK/Fn14 axis: the current paradigm of tissue injury-inducible function in the midst of complexities. *Semin Immunol* 26, 229–236.

Burroughs AM, Ando Y, Aravind L. (2013) New perspectives on the diversification of the RNA interference system: insights from comparative genomics and small RNA sequencing. *Wiley Interdiscip Rev RNA*. 2013; 5:141–181.

Chandran V, Raychaudhuri SP. (2010) Geoepidemiology and environmental factors of psoriasis and psoriatic arthritis. *J Autoimmun*;34: J314–321.

Chen, G., Wang, C., & Shi, T. (2011). Overview of available methods for diverse RNA-Seq data analyses. *Science China Life Sciences*, 54(12), 1121-1128.

Chicheportiche Y, Bourdon PR, Xu H, Hsu YM, Scott H, Hession C, Garcia I, Browning JL, (1997) TWEAK, a new secreted ligand in the tumor necrosis factor family that weakly induces apoptosis. *J Biol Chem* 272, 32401–32410.

Chu Y, Corey DR. (2012) RNA sequencing: platform selection, experimental design, and data interpretation. *Nucleic Acid Ther*. 2012; 22:271–274.

Crick F. (1970) Central dogma of molecular biology. *Nature*; 227:561–563.

Crick FH. (1958) On protein synthesis. *Symp Soc Exp Biol*; 12:138–163.

Cui X, Churchill GA. (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*; 4:210.

de Klerk E, den Dunnen JT, 't Hoen PA. (2014) RNA sequencing: from tag-based profiling to resolving complete transcript structure. *Cell Mol Life Sci*. 2014; 71:3537–3551.

Derks KW, Misovic B, van den Hout MC, Kockx CE, Gomez CP, Brouwer RW, Vrieling H, Hoeijmakers JH, van IJcken WF, Pothof J. (2015) Deciphering the RNA landscape by RNAome sequencing. *RNA Biol*. 2015; 12:30–42.

- Di, Y., Schafer, D. W., Cumbie, J. S., & Chang, J. H. (2011). The NBP negative binomial model for assessing differential gene expression from RNA-Seq. *Statistical applications in genetics and molecular biology*, 10(1).
- Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., ... & Jaffrézic, F. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in bioinformatics*, 14(6), 671-683.
- Doerner J, Chalmers SA, Friedman A, Putterman C, (2016) Fn14 deficiency protects lupus-prone mice from histological lupus erythematosus-like skin inflammation induced by ultraviolet light. *Exp Dermatol* 25, 969–976.
- Doerner JL, Wen J, Xia Y, Paz KB, Schairer D, Wu L, Chalmers SA, Izmirly P, Michaelson JS, Burkly LC, Friedman AJ, Putterman C, (2015) TWEAK/Fn14 Signaling Involvement in the Pathogenesis of Cutaneous Disease in the MRL/lpr Model of Spontaneous Lupus. *J Invest Dermatol* 135, 1986–1995.
- Gelfand JM, Neimann AL, Shin DB, (2006) Risk of myocardial infarction in patients with psoriasis. *JAMA*. 2006; 296:1735–1741.
- Goldstein, J. D., Basso, E. Y., Caruso, A., Palomo, J., Rodriguez, E., Lemeille, S., & Gabay, C. (2020). IL-36 signaling in keratinocytes controls early IL-23 production in psoriasis-like dermatitis. *Life science alliance*, 3(6).
- Grän F, Kerstan A, Serfling E, Goebeler M, Muhammad K, (2020) Current Developments in the Immunology of Psoriasis. *Yale J Biol Med* 93, 97–110.
- Grant GR, Liu J, Stoeckert CJ., Jr. (2005) A practical false discovery rate approach to identifying patterns of differential expression in microarray data. *Bioinformatics*. 2005; 21:2684–2690.
- Gupta R, Ahn R, Lai K, Mullins E, Debbaneh M, Dimon M, Arron S, Liao W. (2016) Landscape of long noncoding RNAs in psoriatic and healthy skin. *J Invest Dermatol*. 2016; 136:603–9.
- Gupta, R. K., Gracias, D. T., Figueroa, D. S., Miki, H., Miller, J., Fung, K., Croft, M. (2021). TWEAK functions with TNF and IL-17 on keratinocytes and is a potential target for psoriasis therapy. *Science immunology*, 6(65), eabi8823.
- Han Y, Gao S, Muegge K, Zhang W, Zhou B. (2015) Advanced Applications of RNA Sequencing and Challenges. *Bioinform Biol Insights*. 2015; 9:29–46.

- Hardcastle, T. J., & Kelly, K. A. (2010). baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC bioinformatics*, 11(1), 1-14.
- Jabbari A, Johnson-Huang LM, Krueger JG. (2011) Role of the immune system and immunological circuits in psoriasis. *G Ital Dermatol Venereol*. 2011; 146:17–30.
- Klerk E, Hoen PA. (2015) Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. *Trends Genet*. 2015; 31:128–139.
- Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. (2015) The technology and biology of single-cell RNA sequencing. *Mol Cell*. 2015; 58:610–620.
- Kvam, V. M., Liu, P., & Si, Y. (2012). A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *American journal of botany*, 99(2), 248-256.
- Langmead B, Hansen KD, Leek JT. (2010) Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biol*. 2010;11: R83.
- Lebwohl M. (2003) Psoriasis. *Lancet*; 361:1197–204.
- Leng RX, Pan HF, Qin WZ, Wang C, Chen LL, Tao JH, Ye DQ, (2011) TWEAK as a target for therapy in systemic lupus erythematosus. *Mol Biol Rep* 38, 587–592.
- Li N, Yamasaki K, Saito R, Fukushi-Takahashi S, Shimada-Omori R, Asano M, Aiba S, (2014) Alarmin function of cathelicidin antimicrobial peptide LL37 through IL-36 γ induction in human epidermal keratinocytes. *J Immunol* 193, 5140–5148.
- Liao, Y., Smyth, G. K., & Shi, W. (2019). The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic acids research*, 47(8), e47-e47.
- Lowes MA, Bowcock AM, Krueger JG. (2007) Pathogenesis and therapy of psoriasis. *Nature*. 2007; 445:866–73.
- Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. (2008) RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*. 2008; 18:1509–1517.
- Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., & Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9), 1509-1517.

- Morin R, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh T, McDonald H, Varhol R, Jones S, Marra M. (2008) Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques*. 2008; 45:81–94.
- Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7), 621-628.
- Nalpas NC, Magee DA, Conlon KM, Browne JA, Healy C, McLoughlin KE, Rue-Albrecht K, McGettigan PA, Killick KE, Gormley E, (2015). RNA sequencing provides exquisite insight into the manipulation of the alveolar macrophage by tubercle bacilli. *Sci Rep*. 2015; 5:13629.
- Nomura I, Gao B, Boguniewicz M, Darst MA, Travers JB, Leung DY. (2007) Distinct patterns of gene expression in the skin lesions of atopic dermatitis and psoriasis: a gene microarray analysis. *J Allergy Clin Immunol*. 2003; 112:1195–202.
- Nookaew, I., Papini, M., Pornputtapong, N., Scalcinati, G., Fagerberg, L., Uhlén, M., & Nielsen, J. (2012). A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in *Saccharomyces cerevisiae*. *Nucleic acids research*, 40(20), 10084-10097.
- Oshlack, A., Robinson, M. D., & Young, M. D. (2010). From RNA-seq reads to differential expression results. *Genome biology*, 11(12), 1-10.
- Peng L, Li Q, Wang H, Wu J, Li C, Liu Y, Liu J, Xia L, Xia Y, Fn14 deficiency ameliorates psoriasis-like skin disease in a murine model. *Cell Death Dis* 9, 801 (28).
- Rachakonda TD, Schupp CW, Armstrong AW, (2014) Psoriasis prevalence among adults in the United States. *J Am Acad Dermatol* 70, 512–516.
- Robinson MD, Oshlack A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. 2010;11: R25.
- Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology*, 11(3), 1-9.
- Robinson, M. D., & Smyth, G. K. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 9(2), 321-332.

- Robles, J. A., Qureshi, S. E., Stephen, S. J., Wilson, S. R., Burden, C. J., & Taylor, J. M. (2012). Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. *BMC genomics*, 13(1), 1-14.
- Scarpato M, Federico A, Ciccodicola A, Costa V. (2015) Novel transcription factor variants through RNA-sequencing: the importance of being “alternative” *Int J Mol Sci*. 2015;16:1755–1771.
- Schön MP, (2015) Adaptive and Innate Immunity in Psoriasis and Other Inflammatory Disorders. *Front Immunol* 10, 1764.
- Sidler D, Wu P, Herro R, Claus M, Wolf D, Kawakami Y, Kawakami T, Burkly L, Croft M, (2017) TWEAK mediates inflammation in experimental atopic dermatitis and psoriasis. *Nat Commun* 8, 15395.
- Smyth GK. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statist Appl Genetics Mol Biol*. 2004;3 Article 3.
- Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1).
- Sonkoly E, Wei T, Janson PC, Saaf A, Lundeberg L, Tengvall-Linder M, Norstedt G, Alenius H, Homey B, Scheynius A, (2007) MicroRNAs: novel regulators involved in the pathogenesis of psoriasis? *PLoS One*. 2007;2: e610.
- Takagi H, Arimura K, Uto T, Fukaya T, Nakamura T, Chojjookhuu N, Hishikawa Y, Sato K, (2016) Plasmacytoid dendritic cells orchestrate TLR7-mediated innate and adaptive immunity for the initiation of autoimmune inflammation. *Sci Rep* 6, 24477.
- Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., ... & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5), 511-515.
- Tsoi LC, Iyer MK, Stuart PE, Swindell WR, Gudjonsson JE, Tejasvi T, Sarkar MK, Li B, Ding J, Voorhees JJ, (2015) Analysis of long non-coding RNAs highlights tissue-specific expression patterns and epigenetic profiles in normal and psoriatic skin. *Genome Biol*. 2015; 16:24.

Tuch BB, Laborde RR, Xu X, Gu J, Chung CB, Monighetti CK, Stanley SJ, Olsen KD, Kasperbauer JL, Moore EJ, (2010) Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations. *PLoS One*. 2010;5:e9317.

Wang H, Wang S, Li L, Wang X, Liu C, Lu M, Xia Y, Liu Y, (2021) Involvement of the cytokine TWEAK in the pathogenesis of psoriasis vulgaris, pustular psoriasis, and erythrodermic psoriasis. *Cytokine* 138, 155391.

Wang Z, Gerstein M, Snyder M. (2009) RNA-Seq: A revolutionary tool for transcriptomics. *Nat Rev Genet*; 10:57–63.

Wang Z, Gerstein M, Snyder M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009; 10:57–63.

Yamanaka K, Yamamoto O, Honda T, (2021) Pathophysiology of psoriasis: A review. *J Dermatol* 48, 722–731.

Yang Liao, Gordon K. Smyth, Wei Shi, (2013) featureCounts: an efficient general-purpose program for assigning sequence reads to genomic features, *Bioinformatics*, Volume 30, Issue 7, Pages 923–930.




Zhou, Y. H., Xia, K., & Wright, F. A. (2011). A powerful and flexible approach to the analysis of RNA sequence count data. *Bioinformatics*, 27(19), 2672-2678.

Zibert JR, Lovendorf MB, Litman T, Olsen J, Kaczkowski B, Skov L. (2010) MicroRNAs and potential target interactions in psoriasis. *J Dermatol Sci*. 2010; 58:177–85.

Document Information

Analyzed document	Siddharth_Bioinformatics.pdf (D142110981)
Submitted	7/16/2022 3:16:00 PM
Submitted by	
Submitter email	siddharth.g8564@gmail.com
Similarity	13%
Analysis address	cenlib2014.bhuni@analysis.orkund.com

Sources included in the report

W	URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4863231/ Fetched: 10/27/2019 12:42:54 PM	 7
W	URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5413777/ Fetched: 12/19/2019 5:11:14 PM	 12
SA	Park_Jiae_33862914_BIO309.pdf Document Park_Jiae_33862914_BIO309.pdf (D138334149)	 1

Entire Document

100%

MATCHING BLOCK 1/20

W

[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4 ...](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4...)

Introduction The Central Dogma of Molecular Biology outlines the flow of information that is stored in genes as DNA, transcribed into RNA, and finally translated into proteins (Crick, 1958; Crick, 1970).

Early

79%

MATCHING BLOCK 7/20

W

[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4 ...](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4...)

gene expression studies relied on low-throughput methods such as Northern blots and quantitative polymerase chain reaction (qPCR), but these were limited to single

transcript measurements.

55%

MATCHING BLOCK 2/20

W

[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4 ...](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4...)

The development of next-generation high-throughput sequencing (NGS) has revolutionized transcriptomics by enabling RNA analysis with complementary DNA (cDNA) sequencing (Wang et al., 2009). This method, called RNA-Sequencing, has clear advantages over previous approaches and has revolutionized the understanding of the complex and dynamic nature of the transcriptome. RNA-Sequencing provides a more detailed and quantitative view of gene expression, alternative splicing, and allele-specific expression. Recent advances in RNA-Sequencing workflows, from sample preparation to sequencing platforms to bioinformatics data analysis, have enabled detailed transcriptome profiling and the ability to elucidate

a variety of physiological and pathological conditions. rice field.

66%

MATCHING BLOCK 3/20

W

[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4 ...](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4...)

The advent of high-throughput next-generation sequencing (NGS) technology has revolutionized transcriptomics. This technological development solves many

of the challenges posed by the hybridization-based microarray and Sanger sequencing-based approaches previously used to measure gene expression. High-throughput sequence (HTS) data analysis is a complex multi-step process. Many bioinformatics tools are available at most steps, and most tools require different parameters to be set. Due to this complexity, HTS data analysis is particularly prone to reproducibility and consistency issues. The high-throughput sequencer enables transcriptome inspection. The transcriptome is a set of intracellular ribonucleic acids, including messenger ribonucleic acid (mRNA), transfer ribonucleic acid (tRNA), ribosomal ribonucleic acid (rRNA), small nucleus ribonucleic acid (snRNA), and non-coding ribonucleic acid (ncRNA), others. These RNAs are expressed differentially depending on the tissue, physiological state, or developmental stage (Gupta et al., 2021). Interpreting the complexity of the transcriptome is an important goal for understanding the functional elements of the genome, and therefore for understanding how the disease functions and signs of progress. In this sense, the amount of non-coding DNA has recently been shown to increase with biological complexity, increasing by 0.25% in the prokaryotic genome and 98.8% in the human genome. Existing complexity associated with the discovery of small intrinsic disturbances RNA (siRNA), long-scattered non-coding RNA (lincRNA), transcription initiation RNA (tiRNA), microRNA (miRNA), transcription initiation site-related RNA (TSSa-RNA), etc. is the transcription puzzles we need. Represents a piece of. Elucidate to understand how the genome works. Psoriasis is one of the most common immune inflammatory skin diseases, affecting approximately 125 million people worldwide and more than 8 million in the United States (Rachakonda et al., 2014). Psoriasis lesions can exhibit a variety of clinical manifestations, including acanthosis (increased epidermal thickness), keratin proliferation, parakeratosis, hypervascularization, and dense skin infiltration of immune cells (Gran et al., 2020). Keratinocytes have central importance for inducing early pathogenic events and for increasing psoriatic inflammation during the course of the disease (Albanesi et al., 2018, Benhadou et al., 2019). In response to external and internal threat stimuli, keratinocytes can be a source of innate immune mediators. These include various pro-inflammatory cytokines and chemokines that mobilize cells important for innate and adaptive immune responses (Li et al., 2014, Takagi et al., 2016). The IL-23 / IL-17 axis and TNF were first identified in animal studies as the centre of pathogenesis for skin inflammation such as psoriasis, and their role is now being demonstrated in humans. IL-36γ is also strongly associated with human psoriasis. IL-36γ is produced by keratinocytes and can induce the expression of the IL-23 gene in keratinocytes (Goldstein et al., 2020). Therefore, it is possible to drive a strengthening loop from IL-23 back to IL-17, IL-36γ, and IL-23, thereby maintaining the condition. All of these cytokines are elevated in psoriatic skin lesions, and proper neutralization of TNF, IL-23 p19, or IL-17A has shown potential therapeutic effects in psoriatic patients (Gran et al., 2020, Schon, 2019, Yamanaka et al., 2021). Although these current treatments have proven to be effective, some patients do not respond or become refractory over time, or the disease relapses when treatment is stopped. Therefore, understanding the pathological mechanisms that can occur in psoriasis requires further efforts, such as identifying new molecules that can be targeted alone or in combination with existing therapies. TNF and IL-17 are two cytokines that promote dysregulated keratinocyte activity, and their targeting is very effective in psoriasis patients, but whether these molecules interact with other inflammatory factors. Is not clear. Here, mice with a keratinocyte-specific deletion of Fn14 (Tnfrsf12a), a receptor for the TNF superfamily cytokine TWEAK (Tnfsf12), have imiquimod-induced skin inflammation such as decreased epidermal hyperplasia and

decreased expression of the psoriasis signature gene. Indicates a decrease in. This corresponded to the expression of Fn14 in the keratinocytes of human psoriasis lesions and TWEAK being found in several sub-sets of skin cells. Transcriptomic studies in human keratinocytes revealed that TWEAK strongly overlaps with IL-17A and TNF in upregulating the expression of CXC chemokines, along with cytokines such as IL-23, inflammation-associated proteins like S100A8/9 and SERPINB1/B9, all previously found to be highly expressed in the lesional skin of psoriasis patients (Gupta et al., 2021) Although these current treatments have proven efficacy, some patients fail to respond or become resistant to therapy over time, or their disease comes back when treatment is stopped. Therefore, continuing efforts to understand the pathological mechanisms that might occur in psoriasis are needed, including identifying novel molecules that can be targeted alone or combined with existing therapies. TNF-like weak inducer of apoptosis (TWEAK, TNFSF12) can be expressed similar to TNF (TNFSF2) is a membrane-bound molecule or soluble cytokine by a variety of cell types including structural and immune cells (Chicheportiche et al., 1997, Bird et al., 2013). TWEAK binds to Fn14 (fibroblast growth factor inducible 14, TNFRSF12A) and regulates many cellular activities such as proliferation, migration, differentiation, apoptosis, and angiogenesis (Leng et al., 2011). TWEAK is involved in the pathogenesis of several inflammatory and autoimmune diseases (Burkly, 2014, Doerner et al., 2016). Recently, we have discovered that TWEAK-deficient mice are protected from exhibiting severe imiquimod-induced skin inflammation with some characteristics of psoriasis. Gene set enrichment analysis suggests an association between Fn14 transcripts and their signaling mediators in human psoriasis lesions (Leng et al., 2011). The pathogenic activity of TWEAK was subsequently validated by another group using Fn14-deficient mice in the same experimental model (Doerner et al., 2015). Other literature has found that soluble TWEAK is upregulated in the sera of psoriasis patients and that expression of both TWEAK and Fn14 is detected at high levels in tissue sections of psoriasis-damaged skin (Sidler et al., 2017, Peng et al., 2018). A new therapeutic approach to reduce skin lesions in psoriasis. The TWEAK primary cell target in the skin is unclear. Subcutaneous injection of recombinant TWEAK bolus into mice was found to result in skin inflammation and some histological features reminiscent of human psoriasis. It was associated with the production of a series of chemokines that attract the innate and adaptive immune cells characteristic of psoriasis (Sidler et al., 2017). Many of these chemokines are products of keratinocytes, and Fn14 is expressed in keratinocytes (Sidler et al., 2017), suggesting that this cell type may be central to the action of TWEAK. Before considering clinical treatment for this pathway, how TWEAK in the skin, especially on keratinocytes, and its relationship to other pathogenic molecules such as IL-17 and TNF that also have receptors on keratinocytes In this study, we investigated if TWEAK signalling specifically in keratinocytes is required to develop psoriasis-like skin lesions after imiquimod treatment using Fn14-conditional knockout mice, and also performed RNA-sequencing analysis in human epidermal keratinocytes to determine how TWEAK alone or in combination with IL-17 and TNF controls expression of a variety of gene sets found to be upregulated in human psoriasis. Our data demonstrate that Fn14 signalling in keratinocytes is crucial for the development of imiquimod-induced skin inflammation. Furthermore, transcriptomic data establish substantial similarities in the genes induced in keratinocytes by TWEAK, IL-17, and TNF, and notably, we found strong synergistic activities of these cytokines acting together on a number of genes associated with psoriasis. Correspondingly, a similar effect of blocking TWEAK therapeutically was observed in reducing skin lesions in mice compared to blocking either TNF or IL-17A, and no greater effect was seen with combination treatments. These results suggest that TWEAK might be as good a target to counter the keratinocyte hyperresponsiveness and dysregulated immune system seen in psoriasis as observed when IL-17 and TNF are neutralized (Wang et al., 2021, Bilgic et al., 2016) The main goal of many gene expression experiments is to detect transcripts that exhibit differential expression under a variety of

85%

MATCHING BLOCK 4/20

W

[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4 ...](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4...)

conditions. Extensive statistical approaches have been developed to test differential expression using microarray data,

and the continuous probe intensity of the entire replication can be approximated by a normal distribution (Chandran and Raychaudhuri, 2010, Cui and Churchill, 2003, Smyth, 2004). While these approaches can, in principle, be applied to RNA-Sequencing data, other statistical models of discrete read counts that do not fit the normal distribution should be considered.

59%

MATCHING BLOCK 5/20

W

[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4 ...](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4...)

Early RNA-Sequencing studies showed that the distribution of read counts throughout replication follows a Poisson distribution. This formed the basis for modelling RNA-Sequencing count data (Grant et al., 2005). However, further studies

have shown that biological variability is not captured by Poisson's assumptions and leads to

46%

MATCHING BLOCK 6/20

W

[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4 ...](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4...)

high false positive rates due to underestimation of sampling errors (Marioni et al., 2008, Anders and Huber, 2010, Lanhmead et al., 2010). Therefore, a negative binomial distribution model that describes overdispersion or extra-Poisson variability has been shown to best fit the distribution of read counts across biological

replication.

Review of Literature Psoriasis Vulgaris is a chronic disease that affects 1–3% of the population (Robinson and Oshlack, 2010). In addition to the possible involvement of skin and joints, recent evidence suggests a link between psoriasis and other systemic disorders (Gelfand et al., 2006). The molecular properties of psoriasis skin samples have led to a better understanding of the etiology of the disease and helped identify therapeutic targets (Lebwohi, 2003). Psoriasis is one of the most common chronic inflammatory skin diseases, affecting 1–3% of the adult population worldwide (Lebwohi, 2003). It is characterized by marked overgrowth and inadequate end differentiation of keratinocytes. In addition, complex interactions between different cell types and various cytokines are known to contribute to the development of psoriasis. The etiology is also based on complex interactions between genetic predisposition, important histocompatibility alleles, and various environmental triggers (Lowes et al., 2007). However, from a molecular perspective, the mechanisms responsible for the interaction of keratinocytes with the inflammatory cells that infiltrate the epidermis are not yet fully understood. Analysis of the molecular background of psoriasis describes many disease-related genes and proteins with aberrant expression patterns (Nomura et al., 2003), but little is known about the regulatory pathways responsible for this aberrant expression. Recent evidence suggests that non-coding RNAs such as microRNAs (miRNAs) and long noncoding RNAs (lncRNAs) contribute to the pathogenesis of psoriasis by affecting protein expression and function in both keratinocytes and inflammatory cells. It suggests that it may be (Sonkoly et al., 2007, Zibert et al., 2010, Ahn et al., 2016, Gupta et al., 2016, Tsoi et al., 2015). RNA Sequencing Fundamentals:

50%

MATCHING BLOCK 8/20

W

[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5 ...](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5...)

RNA Sequencing is the use of next-generation high-throughput sequencing technology to study, characterize, and quantify genomic transcriptomes (Morin et al., 2008). Unlike previous methods, RNA sequencing uses synthetic techniques to define nucleotide sequences and quantify RNA molecules in a sample (Wang et al., 2009). Next-generation sequencing (NGS) can faithfully process this data in hours to days, making it an ideal method for RNA analysis among many researchers (Kolodziejczyk et al., 2015). The use of this technology in research and literature has exploded in popularity.

With recent discoveries in the use of RNA sequencing in many pathologies,

60%

MATCHING BLOCK 9/20

W

[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5 ...](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5...)

there are many promising potential clinical applications for RNA sequencing (Beane et al., 2011). Several commercially available RNA sequencing kits are available for each sample. Most follow similar processing steps but ultimately depend on experimental considerations (

Chu and Corey, 2012). Analysis of

83%

MATCHING BLOCK 19/20

W

[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5 ...](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5...)

total RNA, mRNA, and small RNA can be performed with most kits.

To isolate

76%

MATCHING BLOCK 10/20

W

[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5 ...](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5...)

mRNA, use poly (T) primers attached to beads or magnets to bind mRNA and isolate these strands. For small or non-coding RNA, gel electrophoresis is used to separate these molecules. Complete RNA separation uses a combination of these two techniques (

Tuch et al., 2010). Then ligate the adapter

64%

MATCHING BLOCK 11/20

W

[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5 ...](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5...)

to the 5'end, 3'end, or both. When RNA is isolated, cDNA is generated, amplified, and fragmented. Some kits provide RNA sequencing directly without creating cDNA. Although rRNA makes up a significant proportion of total RNA

and can be removed, it has

48%

MATCHING BLOCK 12/20

W

[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5 ...](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5...)

little research interest. These samples are then sequenced by next- generation massively parallel sequencing technology that utilizes sequencing by synthesizing short DNA strands complementary to cDNA. Once the reads are generated, the software can be used to analyse the sequence reads and match the reads to parts of the genome.

You can also create a de novo transcriptome map by mapping gene fragments with sequencing analysis software. The total number of reads for each gene product can be used to quantify proportional gene expression (Han et al., 2015). The use of RNA-Sequencing has recently increased due to advances beyond previous attempts in transcriptome research. Prior to NGS RNA sequencing, two well-known

62%

MATCHING BLOCK 13/20

W

[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5 ...](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5...)

techniques were available. Hybridization of cDNA probes connected to microarrays enabled transcriptome analysis but was limited by the need for extensive knowledge of genomes, transcripts, alternative splicing, and exons.

The background noise produced by cross-hybridization also limited resolution during attempts to quantify gene expression. Another technique

86%

MATCHING BLOCK 14/20

W

[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5 ...](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5...)

was Sanger sequencing, which used chain termination to determine nucleotide sequences. In contrast to NGS,

the Sanger method was more expensive and time- consuming and could only analyze a limited portion of the transcript (Morin et al., 2008, Wang et al., 2009, Burroughs et al., 2013). Discovery of both non-coding RNAs such as. B. miRNAs (miRNAs) have required

61%

MATCHING BLOCK 15/20

W

[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5 ...](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5...)

the creation of assays to test these small non-coding RNAs with variant mRNAs at high throughput and

high resolution, as well as the discovery of post-transcriptional mRNA expression regulation (Klerk and Hoen, 2015). RNA- Sequencing techniques allow researchers to perform both of these tasks and quantify

80%

MATCHING BLOCK 16/20

W

[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5 ...](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5...)

RNA expression, and thus gene expression, in a single assay. The high throughput of RNA

sequences allows the transcriptome to be analyzed and efficiently compared across different environmental factors such as time, different tissue samples, pathological conditions, and pharmacological interventions. The potential for de novo transcriptome synthesis allows the analysis and discovery of new products without the need for

47%

MATCHING BLOCK 17/20

W

[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5 ...](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5...)

prior genomic and transcriptional knowledge of the sample. The resolution of RNA sequences also enables the identification of single nucleotide polymorphisms, novel post-transcriptional modifications, novel alternative splicing patterns, and previously unidentified non-coding RNA molecules. RNA sequencing provides accurate quantification of mRNA expression compared

to real-time PCR experiments (Scapato et al., 2015, de Klerk et al., 2014, Derks et al., 2015). RNA sequences can be used to study the molecular basis of

60%

MATCHING BLOCK 18/20

W

[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5 ...](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5...)

disease susceptibility, cancer etiology/progression, and response to treatment. RNA sequences have been used to analyze the etiology of various malignancies such as psoriasis, lung cancer, and colon cancer. RNA sequencing can identify differential expression of genes (DEGs), mutant genes, fusion genes, and gene isoforms in pathological conditions. RNA sequencing also has potential for diagnostic and therapeutic applications. Current research on colorectal disease using RNA sequencing reveals new discoveries that may help clinicians in the future

management of patients with colorectal disease. Transcriptome analysis is an important tool for characterizing and understanding the molecular basis of phenotypic changes in biology, including disease. In recent decades, microarrays have been the most important and widely used approach to such analysis, but recently high-throughput cDNA sequencing (RNA-sequencing) has emerged as a powerful alternative (Mortazavi et al., 2008). Many applications have already been found (Chen et al., 2011). RNA-sequencing uses next-generation sequencing (NGS) methods to sequence cDNA from RNA samples, producing millions of short reads. These reads are then typically mapped to the reference genome, and the number of reads mapped within the genomic traits of interest (such as genes or exons) is used as a measure of the frequency of the traits of the analyzed sample (Oshlack et al., 2010). Perhaps the most common use of transcriptome profiling is to search for differentially expressed (DE) genes. H. Look for genes that show differences in expression levels between conditions, or genes that are associated with a particular predictor or response. RNA-sequencing offers several advantages over microarrays for differential expression analysis. B. Ability to detect and quantify previously unknown transcripts and isoforms with increased dynamic range and reduced background levels (Agrawal et al., 2010, Bradford et al., 2010, Bullard et al., 2010). However, analysing RNA-sequencing data can be difficult. Some of these issues are unique to next-generation sequencing methods. For example, differences in nucleotide composition between genomic regions mean that

reading ranges may not be uniform throughout the genome. In addition, more reads are mapped to longer genes than shorter genes with the same expression level. In differential expression analysis, where genes are individually tested for differences in expression between conditions, biases within the sample are usually ignored as they are expected to affect all samples in a similar manner (Agrawal et al., 2010). RNA-sequencing experiments show other types of heterogeneity between samples. First, the depth of the sequence or the library size (total number of reads allocated) usually varies from sample to sample. That is, the counts observed between the samples cannot be compared directly. In fact, even in the absence of true differential expression, if one sample is sequenced twice as deep as another, then all genes in the first sample receive twice as many as the second sample. It is expected that we would like to avoid such confusion. The effect of true differential expression. The easiest way to approach different library sizes is to simply rescale or resample the read counts to get the same library size for all samples. However, such normalization is generally not sufficient. This is because RNA-Sequencing counts essentially represent the relative abundance of genes, even if the libraries are actually the same size. Some highly expressed genes can make up a very large proportion of the reads sequenced in the experiment, so few reads need to be assigned to the remaining genes (Bullard et al., 2010). Therefore, the presence of a small number of highly expressed genes suppresses the count of all other genes, and the latter group of genes are mis expressed compared to samples with more evenly distributed reads. It is misunderstood that it can appear low and can lead to many genes. More complex normalization schemes have been proposed to address this difficulty and allow counts to be compared between samples (Bullard et al., 2010, Anders and Huber, 2010, Robinson and Oshlack, 2010). In addition to library size, these methods also include estimating sample-specific normalization coefficients. It is used to rescale the observed count. Using these normalization methods, the sum of the normalized counts across all genes are therefore not necessarily equal between samples (as it would be if only the library sizes were used for normalization), but the goal is instead to make the normalized counts for non-differentially expressed genes similar between the samples. In this study, we use the TMM normalization (trimmed mean of M-values (Robinson and Oshlack, 2010)) and the normalization provided in the DESeq package (Anders and Huber, 2010). A comprehensive evaluation of seven different normalization methods was recently performed (Dillies et al., 2012), in which these two methods were shown to perform similarly, and they were also the only ones providing satisfactory results with respect to all metrics used in that evaluation. Still, it is important to keep in mind that even these methods are based on an assumption that most genes are equivalently expressed in the samples, and that the differentially expressed genes are divided more or less equally between up- and downregulation (Dillies et al., 2012). Microarrays have been used routinely for differential expression analysis for over a decade, and there are well-established methods available for this purpose (such as limma (Smyth, 2004)). These methods cannot be easily migrated to the analysis of RNA- sequencing data (Robinson and Smyth, 2008). It is different from the data obtained from the microarray. Intensities recorded from microarrays are treated as continuous measurements and are generally assumed to follow a lognormal distribution, but counts from RNA-sequencing experiments are non-negative integers and therefore essentially follow a discrete distribution. Poisson distribution and negative binomial distribution (NB) are the two most commonly used models in the method explicitly developed for differential expression analysis of this type of count data (Anders and Huber, 2010, Robinson and Symth, 2008, Auer and Doerge, 2011, Hardcastle and Kelly, 2010, Di et al., 2011). Other distributions such as the beta-binomial distribution (Zhou et al., 2011) have also been proposed. The Poisson distribution has the advantage of simplicity, with only one parameter, but limits the variance of the modelled variables to the mean. The negative binomial distribution has two parameters that encode the mean and variance, so you can model the more general mean and variance relationship. For RNA- sequencing, the Poisson distribution has been suggested to be suitable for the analysis of engineering replication, but with high variability between biological replications, it is accompanied by overdispersion, such as a negative binomial distribution. Distribution is required (Bullard et al., 2010, Marioni et al., 2008). Some software packages represent RNA-sequencing data in converted quantities instead of using integers directly. Long transcripts are expected to receive more reads than short transcripts with the same expression level, so the goal of such a conversion is to normalize the count in relation to various library sizes and transcript lengths. Is to do. Other normalization strategies can be used to address other biases, such as biases due to variable GC content in reads. After such a conversion, the resulting value will no longer be an integer count. That is, you should not plug in numerical-based methods for differential expression analysis. Therefore, of the methods evaluated in this study, only nonparametric methods are suitable for RPKM

values. Other software, such as Cufflinks / Cuffdiff (Trapnell et al., 2010), provides an integrated analytical pipeline from aligned reads to derivative results by inference based on FPKM values. The field of differential expression analysis of RNA-sequencing data is still in its infancy, and new methods are constantly being introduced. To date, there has been no general consensus on which method works best in a particular situation, and few detailed comparisons between the proposed methods have been published. In a recent publication (Kyam et al., 2012), four parametric methods were compared in terms of their ability to distinguish between truly differentially expressed (DE) and truly non-DE genes under different simulation conditions. The authors also compared duplications between sets of DE genes found differently in practice data set. Another recent study (Robles et al., 2012) evaluated the effect of increased sequence depth on the ability to detect the DE gene and contrasted this with the benefits of increased sample size, the latter demonstrating to be significantly greater. In (Nookaew et al., 2012), the authors published a case study on *Saccharomyces cerevisiae*, comparing the results of several differential expression analysis methods of RNA-sequencing with each other, comparing them with the results of microarrays, and generally between different methods. In this study, we investigated if TWEAK signalling specifically in keratinocytes is required to develop psoriasis-like skin lesions after imiquimod treatment using Fn14-conditional knockout mice, and also performed RNA-sequencing analysis in human epidermal keratinocytes to determine how TWEAK alone or in combination with IL-17 and TNF controls expression of a variety of gene sets found to be upregulated in human psoriasis. Our data demonstrates that Fn14 signalling in keratinocytes is crucial for the development of imiquimod-induced skin inflammation. Furthermore, transcriptomic data establish substantial similarities in the genes induced in keratinocytes by TWEAK, IL-17, and TNF, and notably we found strong synergistic activities of these cytokines acting together on a number of genes associated with psoriasis. Correspondingly, a similar effect of blocking TWEAK therapeutically was observed in reducing skin lesions in mice compared to blocking either TNF or IL-17A, and no greater effect was seen with combination treatments. These results suggest that TWEAK might be as good a target to counter the keratinocyte hyperresponsiveness and dysregulated immune system seen in psoriasis as observed when IL-17 and TNF are neutralized (Gupta et al., 2021).

Materials and Method The sample sequences were downloaded from the NCBI GEO Dataset (Gupta et al., 2021). 10 samples of paired-end sequencing were selected, out of which 6 were TWEAK stimulated and 4 were TNF stimulated, the metadata of the samples was downloaded on the workstation having an Intel Xeon 3.20GHz x20 processor and 132GB of RAM. Workflow is the series of activities that are necessary to complete a task. Each step in a workflow has a specific step before it and a specific step after it. Quality Control by Fast QC Then, the data were analyzed for quality control and trimming using Fast QC, which provides a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines, and the outcome of the Fast QC analysis shows whether the trimming is needed or not. Comparing the results from standards suggests, that trimming is not needed in the data obtained, the result of Fast QC is also shown in the below figures. The data was good with little noise. Building the reference index by RStudio The Human reference genome of the human was downloaded for building a reference index for alignment and mapping of the sequence from NCBI (National Center for Biotechnology Information), the reference index was built using RStudio, using the Rsubread package and the base name was given as "chr1_mm10", the figure is attached below. Genome indexing can be described in a similar way to book indexing. If you want to know on which page a particular word appears or where a chapter begins, it's much more efficient / faster to look it up in a ready-made index than to look it up until you find each page in the book. The same is true for linear. Indexes allow aligners to narrow down potential origins of query sequences in the genome, saving both time and memory. Alignment using Rsubread Then, the alignment was done using pair-end sequencing alignment, by RStudio and by taking two FASTA files as input, the output files are in BAM format using the reference index, Rsubread can be used for many processes like- Alignment, quantification, and analysis of RNA sequencing data (including both bulk RNA-seq and scRNA-seq) and DNA sequencing data (including ATAC-seq, ChIP-seq, WGS, WES, etc). Includes functionality for read mapping, read counting, SNP calling, structural variant detection, and gene fusion discovery. Can be applied to all major sequencing technologies and to both short and long sequence reads (Liao et al., 2019) The following results were obtained after alignment; the list of files is shown in below figure. Feature Count using Rsubread in terminal After the alignment, we got one BAM file instead of two FASTA files and then the feature count was done in order to get the count table, it was done by using Rsubread in the Ubuntu terminal and the out was in the form of the count.out file. The count data are structured as a table, which reports the number of sequence fragments assigned to each gene for each sample, the count data were further filtered for null, NA, and negative values in the table, as these values show errors in further steps. The count data output for 10 samples were 47895, but after filtering the negative values, NULL values, NA values and zero values, only 7322 reading were left for further analysis of Differentially Expressed Genes. Feature Count is a general-purpose read summarization function, which assigns to the genomic features (or meta-features) the mapped reads that were generated from genomic DNA and RNA sequencing. RNA-seq reads may be aligned to the transcriptome rather than the genome. In this case, there can be hundreds of thousands of transcripts, and each transcript becomes a reference sequence. featureCounts supports thread-specific read counts when thread-specific information is provided (Yang et al., 2014). Differentially Expressed Genes

normalized read count data and performing statistical analysis to discover quantitative changes in expression levels

between experimental groups. The differentially Expressed genes was done in RStudio using package DESeq2, the following steps were followed, firstly the tables are converted to matrix, then the conditions are assigned to the data, the data was then loaded to DESeq pipeline and different types of plots and graphs were obtained according to the need of the analysis, like- dispersion plot, heatmap, scatter plot, histogram, MA plot, volcano plot, etc. Result and Discussion

Volcano Plot Another common and interesting comparison between the two treatment conditions is the adjusted P-value and log fold change. This figure is called a volcano plot because it resembles an exploding volcano, with clusters of data points near the origin and the fanning effect moving away from its central location. The volcanic plot shows the statistical significance of the difference to the magnitude of the difference between the individual genes compared. Usually indicated by a fold change of negative base 10log or base 2log, respectively. The P-value undergoes a negative transformation, so the higher the data point along the y-axis, the smaller the P-value. Volcano graphs are generally considered to be statistically differentially expressed based on the adjusted P value of the difference between treatments, including some threshold indicators of the adjusted P value. Indicates the gene to be used. Changes in log multiples along the x-axis show a clearer difference in extrema, and data points close to 0 represent genes with similar or same mean expression levels. In the case of volcanic areas, as the name implies, it is expected to be quite widespread. The wide dispersal indicates two treatment groups with significant differences in gene expression. It is quite rare for a volcano plot to have almost or all data points gathered near the origin.

MA Plot The MA chart can only compare two treatment conditions at a time. However, all pairwise comparisons in this figure can be combined in a matrix format to provide all possible combinations at once. In this figure, each cell represents a particular comparison, shown cell by cell or at the intersection of rows and columns. This visualization allows the user to view all pairwise fold change comparisons and average manifestations at once. In addition, this method allows direct comparison of pairwise treatment comparisons. It provides an approach for determining which treatment comparisons are more or less similar in terms of both fold change changes and mean expression levels. Like other matrix options, this process allows the user to visualize all treatment-based comparisons in one diagram.

Heatmap By comparison, you can also use a heatmap based on the number of DEG's to summarize the same information. Using a color spectrum based on the magnitude of the DEG count, the DEG heatmap can provide an easy way to read and interpret. For a DEG heatmap, each cell represents the number of DSNs in that particular intersecting row and column. Arrangements along the selected color spectrum, provide a visual indication of magnitude.

Treatment group. The DEG heatmap has obvious drawbacks in terms of redundancy. For the three factor levels, this figure is a good representation of the data. However, increasing the number of factor levels will generate redundant cells. Cells are usually left blank to avoid misleading the user. This method is counterproductive because it requires more effort to interpret the information efficiently. As the number of factor levels increases, the usefulness of this type of visualization diminishes and is recommended only for some factor levels.

According to the heatmap the white colour shows the upregulated genes while the black colour shows the down regulated genes.

Dispersion Plot Another relatively simple visualization method associated with Tier 1 is to compare expression levels between two samples or two treatment groups. This comparison is typically visualized using a scatter plot. Each data point represents a single gene and its placement indicates the average expression level for each of the two treatments. A scatter plot implemented in this way can be used to make a larger comparison between the two treatment groups. The axes represent the expression levels for each category, so the data

points along the diagonal show similar expression levels from both groups. Data points above or below the diagonal indicate higher or lower expression levels of factor levels on the y-axis compared to factor levels on the x-axis, respectively. Considering this scatter plot as a whole, clustering of all data points along the diagonal shows two samples or treatments with very similar expression patterns across all genes, with the spread of data points from the diagonal. Larger values indicate dissimilar expression levels. Hence, the below graph shows that the gene is negatively regulated.

PCA Biplot The PCA Biplot also known as Principal Components Analysis Biplot is a two-dimensional chart that represents the relationship between the rows and columns. Hence, in this case the PCA Biplot is the representation of the relationship of the rows and columns of the count data in DEG.

Conclusion In this study, I have learned to analysed the RNA Seq data of skin disease psoriasis by using R. In this study, I have learned to check the quality of the data using Fast QC, then reference index was build for human reference genome, followed by alignment was done for pair end sequence using Rsubread package. Feature count was done to get the count data of the sample sequence. Then, differentially expressed gene analysis was done with the help of count data. Results were generated in the form of volcano plot, MA plot, heatmap, dispersion plot and PCA Biplot. Our results suggest negative correlation through the expression levels of psoriasis. It highlights that the samples regulated by TWEAK and TNF inhibit the expression of psoriasis genes. This indicates the use of TWEAK and TNF as a possible treatment for psoriasis.

DEG is often used to determine genotype differences between two or more cellular states to support studies based on specific hypotheses. Interpretation of this information can greatly benefit from the graphic display of the result file. Tier 1 functions provide relatively basic levels of information, including read count distributions, pairwise levels, and those used to visualize DEG counts, while Tier 2 functions provide average level, use additional metrics such as multiple changes, P-values-provide more detailed and informative visualizations. Box plots, violin plots, dot plots, and read count histograms provide insight into the distribution of read counts for each sample or processing group. Scatter plots allow users to visualize the overall similarity of expression levels by showing the expression levels of each gene in the two selected treatments or samples. The DEG histogram and heatmap directly represent the number of DEGs in each comparison. MA and volcano charts are useful for showing relative expression levels, changes in log multiples, and adjusted P-values. Although not applicable to all users, 4-way plots can provide a higher level of detail by including a third treatment group or sample as a relative or control group.

Hit and source - focused comparison, Side by Side

Submitted text As student entered the text in the submitted document.
Matching text As the text appears in the source.

1/20	SUBMITTED TEXT	32 WORDS	100% MATCHING TEXT	32 WORDS
	Introduction The Central Dogma of Molecular Biology outlines the flow of information that is stored in genes as DNA, transcribed into RNA, and finally translated into proteins (Crick, 1958; Crick, 1970).		Introduction The central dogma of molecular biology outlines the flow of information that is stored in genes as DNA, transcribed into RNA, and finally translated into proteins (Crick 1958; Crick 1970).	
	<p>W https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4863231/</p>			

2/20	SUBMITTED TEXT	89 WORDS	55% MATCHING TEXT	89 WORDS
	<p>The development of next-generation high-throughput sequencing (NGS) has revolutionized transcriptomics by enabling RNA analysis with complementary DNA (cDNA) sequencing (Wang et al., 2009). This method, called RNA-Sequencing, has clear advantages over previous approaches and has revolutionized the understanding of the complex and dynamic nature of the transcriptome. RNA-Sequencing provides a more detailed and quantitative view of gene expression, alternative splicing, and allele-specific expression. Recent advances in RNA-Sequencing workflows, from sample preparation to sequencing platforms to bioinformatics data analysis, have enabled detailed transcriptome profiling and the ability to elucidate</p>		<p>The development of high-throughput next-generation sequencing (NGS) has revolutionized transcriptomics by enabling RNA analysis through sequencing of complementary DNA (cDNA) (Wang et al. 2009). This method, termed RNA sequencing (RNA-Seq), has distinct advantages over previous approaches and has revolutionized our understanding of the complex and dynamic nature of the transcriptome. RNA-Seq provides a more detailed and quantitative view of gene expression, alternative splicing, and allele-specific expression. Recent advances in the RNA-Seq workflow, from sample preparation to sequencing platforms to bioinformatic data analysis, has enabled deep profiling of the transcriptome and the opportunity to elucidate</p>	
	<p>W https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4863231/</p>			

3/20	SUBMITTED TEXT	17 WORDS	66% MATCHING TEXT	17 WORDS
	<p>The advent of high-throughput next-generation sequencing (NGS) technology has revolutionized transcriptomics. This technological development solves many</p>		<p>The introduction of high-throughput next-generation sequencing (NGS) technologies revolutionized transcriptomics. This technological development eliminated many</p>	
	<p>W https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4863231/</p>			

4/20	SUBMITTED TEXT	15 WORDS	85% MATCHING TEXT	15 WORDS
	<p>conditions. Extensive statistical approaches have been developed to test differential expression using microarray data,</p>		<p>conditions. Extensive statistical approaches have been developed to test for differential expression with microarray data,</p>	
	<p>W https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4863231/</p>			

5/20	SUBMITTED TEXT	33 WORDS	59% MATCHING TEXT	33 WORDS
	<p>Early RNA-Sequencing studies showed that the distribution of read counts throughout replication follows a Poisson distribution. This formed the basis for modelling RNA-Sequencing count data (Grant et al., 2005). However, further studies</p>		<p>Early RNA-Seq studies suggested that the distribution of read counts across replicates fit a Poisson distribution, which formed the basis for modeling RNA-Seq count data (Marioni et al. 2008). However, further studies</p>	
	<p>W https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4863231/</p>			

6/20	SUBMITTED TEXT	49 WORDS	46% MATCHING TEXT	49 WORDS
<p>high false positive rates due to underestimation of sampling errors (Marioni et al., 2008, Anders and Huber, 2010, Lanhmead et al., 2010). Therefore, a negative binomial distribution model that describes overdispersion or extra-Poisson variability has been shown to best fit the distribution of read counts across biological</p>		<p>high false-positive rates due to underestimation of sampling error (Anders and Huber 2010; Langmead et al. 2010; Robinson and Oshlack 2010). Hence, negative binomial distribution models that take into overdispersion or extra-Poisson variation have been shown to best fit the distribution of read counts across biological</p>		
<p>W https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4863231/</p>				

7/20	SUBMITTED TEXT	24 WORDS	79% MATCHING TEXT	24 WORDS
<p>gene expression studies relied on low-throughput methods such as Northern blots and quantitative polymerase chain reaction (qPCR), but these were limited to single</p>		<p>gene expression studies relied on low-throughput methods, such as northern blots and quantitative polymerase chain reaction (qPCR), that are limited to measuring single</p>		
<p>W https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4863231/</p>				

8/20	SUBMITTED TEXT	85 WORDS	50% MATCHING TEXT	85 WORDS
<p>RNA Sequencing is the use of next-generation high-throughput sequencing technology to study, characterize, and quantify genomic transcriptomes (Morin et al., 2008). Unlike previous methods, RNA sequencing uses synthetic techniques to define nucleotide sequences and quantify RNA molecules in a sample (Wang et al., 2009). Next-generation sequencing (NGS) can faithfully process this data in hours to days, making it an ideal method for RNA analysis among many researchers (Kolodziejczyk et al., 2015). The use of this technology in research and literature has exploded in popularity.</p>		<p>RNA sequencing is the use of high throughput next generation sequencing technology to survey, characterize, and quantify the transcriptome of a genome[1]. In contrast to previous methods, RNA sequencing utilizes sequencing by synthesis technology to define the nucleotide sequences and quantify RNA molecules in a sample[2]. Next generation sequencing (NGS) can process this data in hours to days with high fidelity, making it the preferred technique for RNA analysis amongst many researchers[3]. The utilization of this technology in research and literature has been exploding in popularity.</p>		
<p>W https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5413777/</p>				

9/20	SUBMITTED TEXT	37 WORDS	60% MATCHING TEXT	37 WORDS
<p>there are many promising potential clinical applications for RNA sequencing (Beane et al., 2011). Several commercially available RNA sequencing kits are available for each sample. Most follow similar processing steps but ultimately depend on experimental considerations (</p>		<p>There are many promising potential clinical applications of RNA sequencing with recent discoveries using RNA sequencing in many disease states[4,5]. Several commercial RNA sequencing kits are available for any sample. Most follow similar processing steps, but ultimately depend on experimental considerations[6].</p>		
<p>W https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5413777/</p>				

10/20	SUBMITTED TEXT	41 WORDS	76% MATCHING TEXT	41 WORDS
<p>mRNA, use poly (T) primers attached to beads or magnets to bind mRNA and isolate these strands. For small or non-coding RNA, gel electrophoresis is used to separate these molecules. Complete RNA separation uses a combination of these two techniques (</p>		<p>mRNA isolation, poly(T) primers attached to beads or magnets are used to bind mRNA and isolate these strands. For small RNA molecules or non-coding RNA, gel electrophoresis is used to isolate these molecules. Total RNA isolation utilizes a combination of these two techniques[7].</p>		
<p>W https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5413777/</p>				

11/20	SUBMITTED TEXT	36 WORDS	64% MATCHING TEXT	36 WORDS
<p>to the 5'end, 3'end, or both. When RNA is isolated, cDNA is generated, amplified, and fragmented. Some kits provide RNA sequencing directly without creating cDNA. Although rRNA makes up a significant proportion of total RNA</p>		<p>to the 5' end, 3' end, or both. Once RNA is isolated, cDNA is generated, amplified, and then fragmented. Some kits provide direct RNA sequencing without the need to create cDNA. rRNA can be removed since it makes up a significant proportion of the total RNA</p>		
<p>W https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5413777/</p>				

12/20	SUBMITTED TEXT	50 WORDS	48% MATCHING TEXT	50 WORDS
<p>little research interest. These samples are then sequenced by next- generation massively parallel sequencing technology that utilizes sequencing by synthesizing short DNA strands complementary to cDNA. Once the reads are generated, the software can be used to analyse the sequence reads and match the reads to parts of the genome.</p>		<p>little research interest. These samples are then sequenced through massive parallel next generation sequencing technologies that utilize sequencing by synthesis of short DNA strands complimentary to the cDNA. Once the reads are produced, software is available to analyze the sequence reads and correspond the reads to portions of the genome.</p>		
<p>W https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5413777/</p>				

13/20	SUBMITTED TEXT	30 WORDS	62% MATCHING TEXT	30 WORDS
<p>techniques were available. Hybridization of cDNA probes connected to microarrays enabled transcriptome analysis but was limited by the need for extensive knowledge of genomes, transcripts, alternative splicing, and exons.</p>		<p>techniques were available before NGS RNA sequencing. Hybridization of cDNA probes attached to microarrays allowed for transcriptome analysis but was limited by the requirement for extensive knowledge of the genome, transcription products, alternative splicing, and exons.</p>		
<p>W https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5413777/</p>				

14/20	SUBMITTED TEXT	16 WORDS	86% MATCHING TEXT	16 WORDS
<p>was Sanger sequencing, which used chain termination to determine nucleotide sequences. In contrast to NGS,</p>		<p>was Sanger sequencing, which utilized chain termination methods to determine nucleotide sequences. In contrast to NGS,</p>		
<p>W https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5413777/</p>				

15/20	SUBMITTED TEXT	18 WORDS	61% MATCHING TEXT	18 WORDS
<p>the creation of assays to test these small non-coding RNAs with variant mRNAs at high throughput and</p>		<p>the creation of an assay that survey these small non-coding RNAs along with variant mRNAs with high throughput and</p>		
<p>W https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5413777/</p>				

16/20	SUBMITTED TEXT	16 WORDS	80% MATCHING TEXT	16 WORDS
<p>RNA expression, and thus gene expression, in a single assay. The high throughput of RNA</p>		<p>RNA expression and thus gene expression with a single assay. Because of the high throughput nature of RNA</p>		
<p>W https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5413777/</p>				

17/20	SUBMITTED TEXT	45 WORDS	47% MATCHING TEXT	45 WORDS
<p>prior genomic and transcriptional knowledge of the sample. The resolution of RNA sequences also enables the identification of single nucleotide polymorphisms, novel post- transcriptional modifications, novel alternative splicing patterns, and previously unidentified non-coding RNA molecules. RNA sequencing provides accurate quantification of mRNA expression compared</p>		<p>prior genomic and transcriptional knowledge of the sample is not needed, allowing analysis and discovery of novel products. The resolution of RNA sequencing also allows for the identification of single nucleotide variants, novel post-transcriptional modification, novel alternative splicing patterns, and non-coding RNA molecules that have not been previously identified. RNA sequencing provides an accurate quantification of mRNA expression as compared</p>		
<p>W https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5413777/</p>				

18/20	SUBMITTED TEXT	76 WORDS	60% MATCHING TEXT	76 WORDS
	disease susceptibility, cancer etiology/progression, and response to treatment. RNA sequences have been used to analyze the etiology of various malignancies such as psoriasis, lung cancer, and colon cancer. RNA sequencing can identify differential expression of genes (DEGs), mutant genes, fusion genes, and gene isoforms in pathological conditions. RNA sequencing also has potential for diagnostic and therapeutic applications. Current research on colorectal disease using RNA sequencing reveals new discoveries that may help clinicians in the future		disease susceptibility, cancer pathogenesis/progression, and response to therapy. RNA Sequencing has been used to analyze the pathogenesis of several malignancies such melanoma, lung cancer, and colorectal cancer. RNA sequencing can identify differential expression of genes (DEG's), mutated genes, fusion genes, and gene isoforms in disease states. RNA sequencing has the potential for diagnostic and therapeutic applications as well. Current research in colorectal disease using RNA sequencing are unlocking new discoveries that may help clinicians treating patients with colorectal disease in the future.	
	<p>W https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5413777/</p>			

19/20	SUBMITTED TEXT	13 WORDS	83% MATCHING TEXT	13 WORDS
	total RNA, mRNA, and small RNA can be performed with most kits.		Total RNA, mRNA, and small RNA analysis can be done with most kits.	
	<p>W https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5413777/</p>			

20/20	SUBMITTED TEXT	16 WORDS	90% MATCHING TEXT	16 WORDS
	normalized read count data and performing statistical analysis to discover quantitative changes in expression levels			
	<p>SA Park_Jiae_33862914_BIO309.pdf (D138334149)</p>			