Differentially Expressed Gene Analysis on Macrophages after in Vitro Exposure to Flat and Textured Breast Silicone Mammary Implants

Master thesis of	:	Anqi Sun, 1340093
Departement	:	Biomedical Engineering – Computational Biology
Course code	:	8ZM104
Organisation	:	Eindhoven University of Technology (TU/e)
Graduation committee	:	Prof. dr. P.A.J. Hilbers (Supervising professor, TU/e) Prof. dr. Boer, J. de (Supervisor, TU/e) Prof. dr. ir. N.A.W. van Riel (Supervisor, TU/e) Cabbia, A., MSc (Supervisor, TU/e) Sudarsanam, P.K., MSc (Supervisor, TU/e)
Period	:	22/10/2019 – 15/10/2020

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

Abstract

Silicone mammary implants (SMIs) are indispensable in the aesthetic breast augmentation and the postmastectomy breast reconstruction. However, about 20% of customers or patients suffered from its common systemic and local complications of SMIs, capsular contracture (CC). The formation of CC could be thought of as an issue related to implant biocompatibility. The research on the interactions between human peripheral blood mononuclear cells (PBMCs) and SMIs in vitro indicated that pro-inflammatory cytokines $IL-1\beta$, $IL-6$, and $TNF-\alpha$ that are related to the activation of macrophages could be found on two textured SMI surfaces compared to smooth SMI surfaces. Moreover, a moderate upregulation in an anti-inflammatory and profibrotic cytokine $TGF-\beta 1$ on them. Therefore, the SMI surfaces' roughness may play an essential role in regulating inflammation-related cytokines in macrophages, and the differences in gene expression patterns induced by various SMIs may be the key to the formation of fibrosis.

In our study, the RNA-seq data of macrophages derived from PBMCs are analyzed to investigate the differences between the gene expression pattern between macrophages cultured on various SMIs. Macrophages were derived from the peripheral blood mononuclear cells (PBMCs) of six healthy females donors, and they were cultured on the flat and the textured breast SMIs for 24 hours and 96 hours, respectively. We used differentially expressed genes analysis (DEGA) to compare the gene expression profile of macrophages of the flat and the textured phenotypes to investigate if the flat and the textured SMIs can induce different gene expression profiles in macrophages.

Our study presents a complete DEGA pipeline on macrophages cultured on the breast SIMs with various surface structures. The DEGA pipeline includes quality control, read alignment, differential expression analysis, geneset enrichment analysis, connectivity map query, and gene network construction. Based on the analysis of two comparison groups Text24h VS Flat24h and Text96h VS Flat96h, we could conclude that DEGs can distinguish macrophages cultured on the flat-surface implant from those cultured on the textured-surface implant. Compared to the textured surface, the higher expression level of pro-inflammatory factors, including IFN-α, and -γ, and TNF-α that can lead to the formation of CC has been found in macrophages on the flat surface. Besides, macrophages cultured on flat and textured presented different M1/M2 macrophages polarization patterns. Connectivity map query results provided possible clinical access to preventing complications like CC and breast cancer.

Keywords: Macrophages; Flat and Textured Breast Implants; Differentially Expressed Genes Analysis Pipeline; M1/M2 Polarization.

Contents

Chapter 1

Introduction

1.1 Background and Related Work

Silicone mammary implants (SMI) breast augmentation is one of the most common aesthetic surgery procedures [1], and it is also the primary way of the postmastectomy breast reconstruction. Even though the application of SMIs can meet the needs of beauty, it will also challenge customers or patients' health. The most common systemic and local complication of SMI is capsular contracture (CC), with an incidence larger than 20% [2, 3].

The formation mechanism of the CC induced by breast augmentation is yet evident. Nevertheless, findings have confirmed that lymphocytes and T cells may be vital in the breast tissue's immunological responses to breast SMIs and the existence of the fibrous capsule. In 2012, Maria et al. found lymphocytes, primarily T cells, at the implant site in the initial stage of fibrous capsule formation [4]. Activated T cell pool clonally expansed by the simulation on lymphocyte could lead to T cells' malignant transformation into breast implant-associated anaplastic large cell lymphoma (BIA-ALCL) [5, 6]. Although peripheral blood mononuclear cells (PBMC) are mainly composed of lymphocytes (T, B, and NK cells), monocytes, and dendritic cells (DC), there is little research focused on in vitro responses of PBMC to the SMI surface.

In 2009, S.barret et al. proposed that CC could be thought of as an issue related to implant biocompatibility since CC's development is due to the body's reaction to the implant [7]. The biocompatibility was first defined by Willimas et al. as *"The ability of a material perform with an appropriate host response in a specific application"* [8]. To validate this assumption, S.barr et al. conducted studies on the structure and the biocompatibility of 13 commercially available SMIs in 2009 and 2017 [7, 9].

In 2018, based on the postulation that fibrosis is always a sequela of inflammatory processes [10], Dolores's group conducted a study to investigate the interactions between human PBMC and 7 SMIs reported by S.barr et al. in vitro [11]. The result indicated that these SMIs would not induce either the activation or the proliferation of T cells, and they also had no effects on the distribution of T cell subsets. The researchers then assessed the cytokine profile of PBMC response to different SMI surfaces to evaluate T-cell paracrine activity. They assayed the macrophage activation cytokines' expression level, cytokines important for macrophage fusion, anti-inflammatory cytokines, and T cell-activation cytokines.

Among all the cytokines quantified, only the macrophage activation cytokines $IL-1\beta$, $IL-6$, and $TNF-alpha$ were above the lower limit quantification. In conclusion, proinflammatory cytokines $IL-1\beta$, $IL-6$ and $TNF-\alpha$ that are related to the activation of macrophages and increase in fibrosis can be found on two of the commercially available SMI surfaces, Polytech Texture ($IL-1\beta$, IL-6 and $TNF-\alpha$). Moreover, a moderate but not statistically significant upregulation in . $TGF-\beta1$ on all surfaces except on the SilkSurface. $TGF-\beta1$ has both anti-inflammatory and profibrotic properties. The maximum Peak to Valley (PV) value and the Roughness value of the Polytech Texture surface are about 220 μm and 38 μm larger than those of

the Silk surface, respectively [9]. The textured surface (Mentor Siltex surface) is the only one that showed up-regulation in the monocyte/macrophage biomarkers, including CD14, CD68. Besides, there was also up-regulation in the expression level of inflammatory cytokines IL-10 and TNF-*alpha* on this textured implant. The texture on the Mentor Siltex surface is about 100- to 200-μm deeper than that on the Allergen Smooth surface [7]. In a summary, macrophages were more likely to be activated, and more anti-inflammatory cytokines will be expressed on the textured surfaces. Therefore, the roughness of the SMI surfaces may play an important role in the expression patterns of genes in macrophages and cytokines that can lead to fibrosis.

Then, in 2019, Daneshgaran and Wong et al. investigated how the CC interacted with the Allergan Smooth surface and the Mentor Siltex surface by analysing differential gene expression [12]. As results indicated, CCs around the smooth and the textured implants presented different patterns of gene expression. Also, the CC had expression patterns for MMP-3, TNNT-3, and NRG-1 that are consistent with the CC formed on the smooth implants. These results suggests new therapeutic targets for the CC.

To conclude, based on the previous research about the formation mechanism of the CC caused by breast SMIs, macrophages from PBMC is a possible cell factor. Moreover, the roughness of SMI surfaces will affect the expression level of cytokines activating macrophages and cytokines inducing fibrosis. However, there are few studies on the immunological responses of monocytes like macrophages derived from PBMC to breast SMI surfaces with different roughness values. Further studies on this field are needed since they can indicate whether the roughness of SMI surfaces will affect the gene expression pattern of macrophages or not, if so, whether the differences can be related to the formation of the CC or other immune responses.

1.2 Research Question

This study's primary aim is to examine whether gene expression is different between macrophages grown on SMI surfaces with two kinds of structure: the flat surface and the textured surface. The differentially expressed gene analysis on the macrophages of these two phenotypes could demonstrate if macrophages cultured on the flat or the textured surfaces are more likely to induce the deregulation of biological processes, including specific genes, cytokines, or pathways. Identifying such processes could provide more clues on how macrophages react to various breast SMI environments in vitro. Eventually, the identified biological processes will be used as the pre-clinical target to avoid complications, such as the capsular contracture or the breast cancer caused by the breast SMI transplantation surgery.

1.3 Method

To achieve the objective, RNA sequencing (RNA-seq) followed by the differentially expressed gene analysis (DEGA) pipeline will be applied to compare the gene expression level of macrophages cultured on different breast SMI surfaces. Macrophages used in this study were derived from the peripheral blood donated by six healthy females, and they were cultured on flat and textured breast SMI surfaces for 24 hours and 96 hours, respectively. Due to the high-throughput of RNA-seq data and the application of adapters and the flow cell during the sequencing procedure, the RNA-seq data could contain adapter contents and the disturbed sequencing result caused by the debris on the flow cell, a completed RNA-seq quality control process is designed and applied in our pipeline. Besides, the DEGA pipeline is mainly composed by following steps, including the quality control of the RNA-seq data, alignment of the short reads within the RNA-seq data, DEGA based on the alignment, visualization of the DEGA result, gene set enrichment analysis on the DEGA result, gene network of differentially expressed genes, and connectivity map query of gene landscape of each donor.

Chapter 2

Material and Method

In this chapter, we will introduce the material, methods, and the workflow of our project. The material we used in our case is the pair-end RNA-sequencing (RNA-seq) data generated by Illumina technique, and the pipeline of the differentially expressed gene analysis (DEGA) and the visualization methods we used are presented in the Figure 2.1.

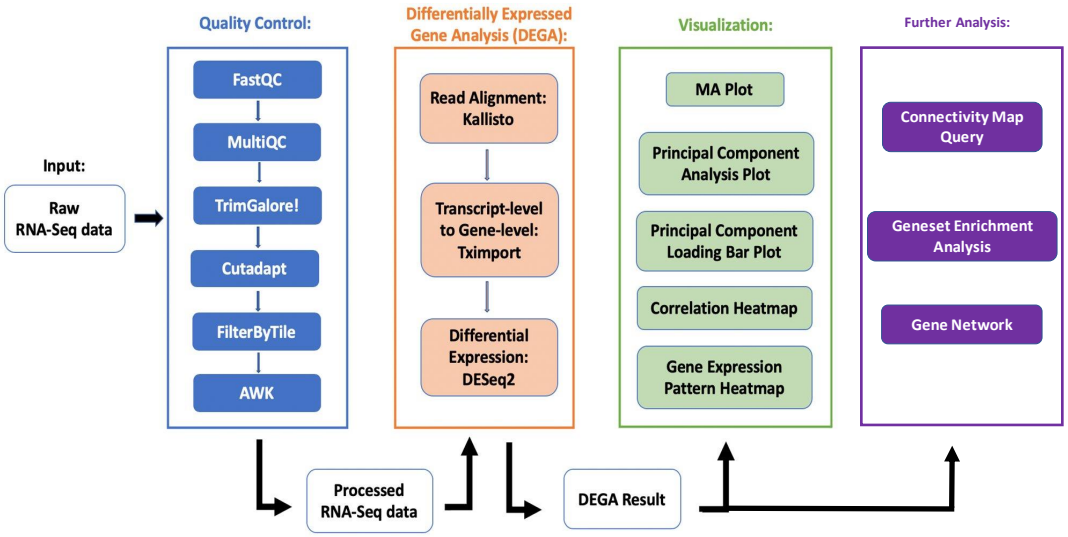


Figure 2.1: Workflow.

2.1 Human Peripheral Blood Cell Derived Macrophages

The macrophages derived from peripheral blood mononuclear cells (PBMCs) were obtained from heparinized blood by Histopaque-1077 density gradient centrifugation. Then, monocytes were isolated from PBMCs using the magnetic cell separation system (MACS). After purifying, monocytes were suspended in a complete RPMI medium containing M-CSF and seeded at ultralow attachment plates. As control, cells were cultured in full RPMI medium without M-CSF to check if differentiation would be successful. Monocyte-derived macrophages were harvested from the ultralow attachment plates after 6-day incubation. The macrophages were then washed, re-suspended, and 1.5×10^4 macrophages will be seeded in quadruplicate on different surfaces (polystyrene, flat BSMI,

and textured BSMI) inside 24-well non-tissue culture treated plates. As a control, cells will also be implanted on the three other surfaces to check the cells' attachment to the implant surfaces by staining for Phalloiddin and DAPI. After overnight incubation, the surfaces were transferred to new non-tissue culture treated plates containing fresh complete RPMI medium and M-CSF and incubated again. Macrophages will be cultured on the different surfaces for 24h and 96h. Then, surfaces will be transferred to a new 24-well plate and washed two times with cold PBS. The control wells for testing cell attachment to surfaces will be fixed with paraformaldehyde/PBS and then placed on PBS for later staining with Phalloiddin and DAPI. Trizol reagent will be used to lyse other cells in the other wells to isolate total RNA; RNA from 4 wells per surface will be combined to 1 RNA sample of each surface.

These experiments were executed to test whether gene expression is different between macrophages grown on a flat surface (Allergan Smooth surface) or a textured surface (Mentor Siltex surface). The BSMI surfaces used to culture macrophages are shown in Figure 2.2. These two breast SMI surfaces have been compared in the study on the variances in the expression of the inflammatory cytokines and T cell, monocytes, and macrophages' biomarkers [11]. The flat surface (Allergan Smooth surface) and other six textured surfaces, and the flat surface was used as the control group. Results indicated that compared to the flat surface, the textured surface (Mentor Siltex surface) is the only one that showed up-regulation in the monocyte/macrophage biomarkers, including CD14, CD68. Besides, there was also up-regulation of IL-10 and TNF-α in the Mentor Siltex surface. AK.Wong Based on these comparisons between the flat surface (Allergan Smooth surface) and the textured surface (Mentor Siltex surface), they were selected in our study to investigate if they can induce different expressions of inflammatory cytokines associated with inflammation or other complications like the breast cancer in macrophages.

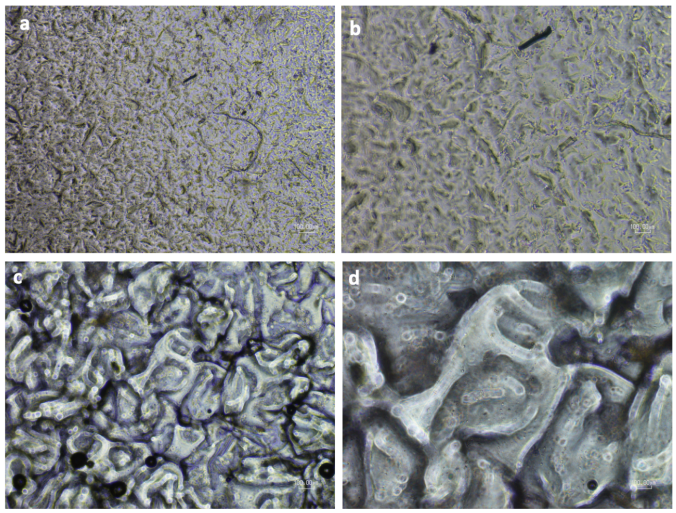


Figure 2.2: BSMI surface structures. (a)4×magnification of the surface structure of flat BSMI. (b)10×magnification of the surface structure of flat BSMI. (c)4×magnification of the surface structure of textured BSMI. (d)10×magnification of the surface structure of textured BSMI.

2.2 RNA Sequencing Data

RNA sequencing (RNA-seq) is a preferred technique for analysing transcriptome that is the complete set of transcripts in a cell, and their quantity [13]. It was first introduced in 2008 by Ugrappa et al. [14] as a quantitative sequencing-based method used for mapping transcribed regions by sequencing complementary DNA fragments belong to these regions and mapping to the genome. The high-throughput of the RNA-seq enables the generation of a high-resolution transcriptome map

of the interested cells. Over the past decade, due to the decreasing costs and the popularization of shared-resource sequencing cores at many research institutions, the RNA-seq has become more widely used [15]. RNA-seq identifies untranslated regions, introns, and coding regions less challenging, and the exploration of the transcriptional structure within a cell better. To investigate the variations in the transcriptome of macrophages cultured on different breast silicone mammary implant (BSMI) surfaces, RNA-seq data generated by Illumina flow cell platform was used to provide information in this study.

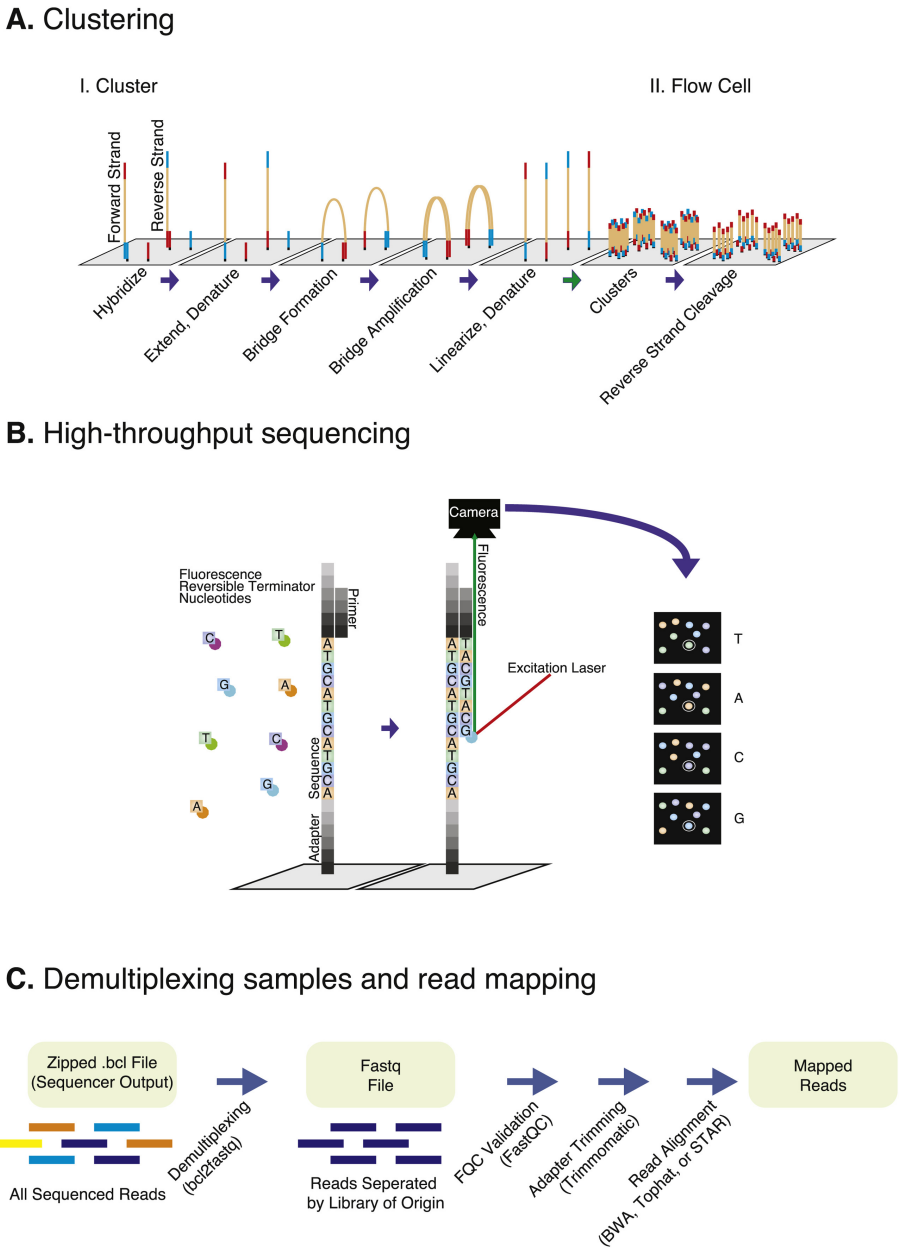


Figure 2.3: Illumina RNA-sequencing workflow.

Illumina RNA-seq platform is one of the most popular sequencing platforms since it can sequence the content of a cDNA fragment at the same time when it synthesizes the content inside the cDNA fragment. The clustering and the sequencing of Illumina RNA-seq are shown in Figure 2.3 [16]. As indicated in Figure 2.3.B, each type of dNTP (A, T, C, G) carries different colors

to identify its content. In each run, only one dNTP will be bound to a synthesized sequence. At the same time, the fluorescence it carries will be activated and recorded by the camera. This fluorescence will be used to sequence the content of the RNA sample. An example of the quality control (QC) process is described by Figure 2.3.C. The RNA-seq data is stored in the *.fastq* files, then, its quality will be checked by tools like FastQC [17]. If FastQC reports the existence of adapter contents in the RNA-seq data, tools like Trimmomatic will be used to trim adapters from the RNA-seq data. After these steps, read aligners, such as BWA, Tophat, or STAR, will match the processed RNA-seq data to the reference genome or transcriptome to investigate which genes or transcripts are contained in the *.fastq* files.

2.2.1 Sample Index

In our case, macrophages from six healthy donors were cultured on two kinds of BSMI surfaces for 24 hours and 96 hours, respectively. Thus, there were 24 samples in four groups: Flat 24h, Flat 96h, Text 24h, and Text 96h. The index was used to denote the experiment condition of the RNA-seq data to refer to the samples.

The name and index of the samples are described in Table 2.1. E.g., the sample Donor 1 Flat 24h indicates that the macrophages are from Donor 1, and they are cultured on the Flat breast implant for 24 hours; the sample name Donor 8 Text 96h indicates that the macrophages are from Donor 8, and they are cultured on the textured breast implant for 96 hours. In the following content of this report, the sample is referred to by their index number. There are two or three replicates of each sample; RNA-seq was repeated on these samples for two or three times in different flowcell lanes, Lane 1 (001), Lane 3 (003), and Lane 4 (004). The RNA-seq data will also be referred to by their index in the following report to better indicate which replicate data is used. E.g., 005-003, in which 005 denotes that the data from Sample 005, whose macrophages are from Donor2 and cultured on the flat BSMI surface for 24 hours; 003 means that the RNA-seq data was generated in the Lane 3.

Experiment Condition	Index	Experiment Condition	Index
Donor 1 Flat 24h	001	Donor 1 Text 24h	002
Donor 1 Flat 96h	003	Donor 1 Text 96h	004
Donor 2 Flat 24h	005	Donor 2 Text 24h	006
Donor 2 Flat 96h	007	Donor 2 Text 96h	008
Donor 3 Flat 24h	009	Donor 3 Text 24h	010
Donor 3 Flat 96h	011	Donor 3 Text 96h	012
Donor 4 Flat 24h	013	Donor 4 Text 24h	014
Donor 4 Flat 96h	015	Donor 4 Text 96h	016
Donor 5 Flat 24h	017	Donor 5 Text 24h	018
Donor 5 Flat 96h	019	Donor 5 Text 96h	020
Donor 8 Flat 24h	021	Donor 8 Text 24h	022
Donor 8 Flat 96h	023	Donor 8 Text 96h	024

Table 2.1: Sample name and index.

2.3 Quality Control on RNA-seq Data

In this part, detailed information about errors of the raw RNA-seq data in the macrophage dataset will be provided, and tools used to solve these problems and their usage order will be introduced.

Quality control (QC) is the first step in the differential expressed gene analysis (DEGA) pipeline. In this step, the raw RNA-seq data quality will be checked; paired bases (bps) or short reads with poor quality will be trimmed off from the raw data. This step ensures that the quality of the RNA-seq data is eligible for the downstream analysis, generating more accurate

results. Eliminating poor-quality reads improves the quality and decreases the time needed for the analysis. In our study, QC on the macrophage RNA-seq raw data mainly completed by tools in the following order: FastQC, MultiQC, *TrimGalore!*, *Cutadapt*, *FilterByTile*, and *AWK*.

2.3.1 FastQC and MultiQC

Firstly, FastQC was used to check the quality of the raw RNA-seq data of each sample separately. It is a software developed by Simon Andrews [17], and it can examine the quality of raw sequence reads from multiple analyses, and report results in HTML format available from web browser. A FastQC report includes modules like basis statistic, per base sequence quality, per tile sequence quality, per base sequence content, per sequence GC content, overrepresented sequences, adapter content ect. Each module is annotated by a green check, red cross, or yellow exclamation mark, which denotes pass, fail, or warning, respectively. Detailed explanation of each module, criteria, and possible causes caused errors can be found in [18]. More critical, examination results of FastQC modules named by 'per base' are based on the base-pair (bp) position. This enables uers to trim bps with poor quality off from a read instead of deleting the whole read to preserve the effective information as much as possible.

According to the FastQC report, samples in the macrophage dataset contain short reads with a length of 150bp or 151bp. The RNA-seq data generated by Lane 4 is 150bp, and that generated by Lane 1 and Lane 3 is 151bp. The sequencing on the Lane 1 and 3 had one more run than that on Lane 4. The length of a short read is decided by the number of cycles executed in the RNA-seq process. The number of total short reads in each *fastq* file is various. The *fastq* file that contains the smallest number (764843) of short reads was from the Sample 003 generated by Lane 4, and the largest number (14601315) was also from the Sample 003 but generated by Lane 1.

However, the FastQC report on each sample is independent. It is hard to compare 144 reports of our macrophages data set since FastQC generates the report on a per-sample basis. It is difficult to conclude what the main problems that exist in most of the samples in a dataset are. Users need to find and compile QC results by themselves manually; however, it is time-consuming and error-prone. Thus, another tool MultiQC, was used to integrate 144 separate reports. MultiQC introduced by Philip et al. in 2016 is the first tool that can flexibly integrate FastQC reports of a sizeable RNA-seq dataset [19]. It creates a single report combining outputs from multiple FastQC reports to check global trends quickly. An advantage of MultiQC is that the data of every single sample can be compared in shared interactive plots, allowing detection of subtle differences not noticeable when switching between different files manually.

Figure 2.4 is the FastQC status checks plot given by *MultiQC*, in which the problems in each module are shown, the column is each FastQC module, and the row denotes each sample. In our macrophage dataset, RNA-seq data of a number of samples encompass poor performances in per tile sequence quality, per base sequence content, per base sequence quality, sequence duplication, overrepresented sequence, and adapter content. To gain an RNA-seq dataset with better quality, the following tools, TrimGalore!, Cutadapt, FilterByTile, and AWK, were applied and corrected the warnings and errors.

2.3.2 TrimGalore! and Cutadapt

Warnings and failures reported by adapter content and overrepresented sequences modules should be corrected first. Both adapter content and overrepresented sequences modules can give information on the type and the content of adapters. The adapter content module plots a cumulative proportion of each type of adapter at each position content in RNA-seq data, and FastQC will, in default, issue a failure if an adapter sequence is presented in more than 10% of all reads. Since no single sequence is expected to present at a high enough frequency, the overrepresented sequences module will report sequences if any sequence is found to present more than 1% of the total as failed. The sequences reported in this module are not only RNA-seq adapters but also sequences without a hit, i.e., their sources are unclear. Only one dNTP will be attached to the adapters or the formed fragments in each cycle in the Illumina platform. Since the RNA fragments are of

different lengths, there was a possibility that a certain proportion of fragments were shorter than the number of cycles. If so, adapter contents at the end of the fragment will also be synthesized in the PCR amplification process. The uncertain length of a fragment is the main reason for the existence of adapters in the RNA-seq data. The usage of adapters in RNA-seq is necessary. More critical, adapter contamination will lead to the alignment errors and an increase in unaligned reads since the adapter sequences are synthetic and do not occur in the human or homo sapiens gene library. Thus, the overrepresented sequences without hits and adapter sequences may cause errors in per base sequence quality and per base sequence content. Therefore, failures and errors should be removed first before other measures are taken to control the quality of the RNA-seq dataset. This step is also expected to improve part of errors from the per base sequence quality and per base sequence content module. To this end, *TrimGalore!* and *Cutadapt* were employed in the very first of the QC step.

In our case, the first step was trimming adapter contents off from the RNA-seq data by *TrimGalore!* [20]. It can effectively detect recognizable adapters used in the RNA-seq data and remove them away. Then, the quality of RNA-seq will be rechecked by FastQC to detect the content of adapters and the overrepresented sequences. After the RNA-seq raw data being processed by *TrimGalore!*, two problems may exist. One is that the overrepresented sequences without hit will be left; the other one is that there will be the generation of new overrepresented sequences without hit. The next step was checking the overrepresented sequences with BLAST [21] to investigate the origins of them. The result showed that there were no overrepresented sequences related to homo sapien. Thus, overrepresented sequences were also removed to decrease the bias in the read alignment. A python package called *Cutadapt* was used. *Cutadapt* is a wrapper around FastQC and *TrimGalore!* to consistently apply adapter and quality trimming to *fastq* files. *Cutadapt* can be used to trim specified sequences [22]. The other aim of the usage of *Cutadapt* is to limit the minimum length of sequences, avoiding sequences with a length of 0 in the processed RNA-seq data being recognized as overrepresented sequences. This will also lead to errors in the sequence length distribution module. In our case, the minimum length of the sequences set in *Cutadapt* is 50.

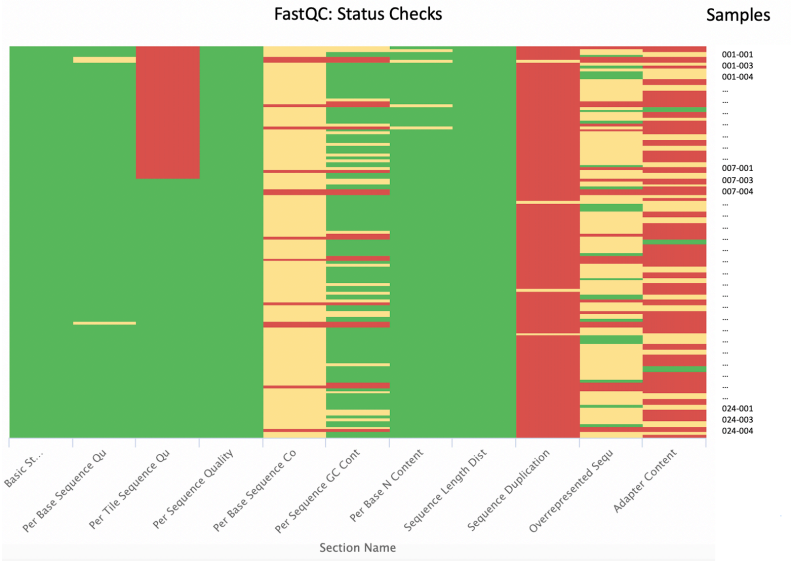


Figure 2.4: Raw macrophages RNA-seq data MultiQC report.

2.3.3 FilterByTile

Per tile sequence quality allows users to look at the quality scores from each tile across all of bps to see if there was a loss in quality associated with specific regions on a flow cell lane if the RNA-seq

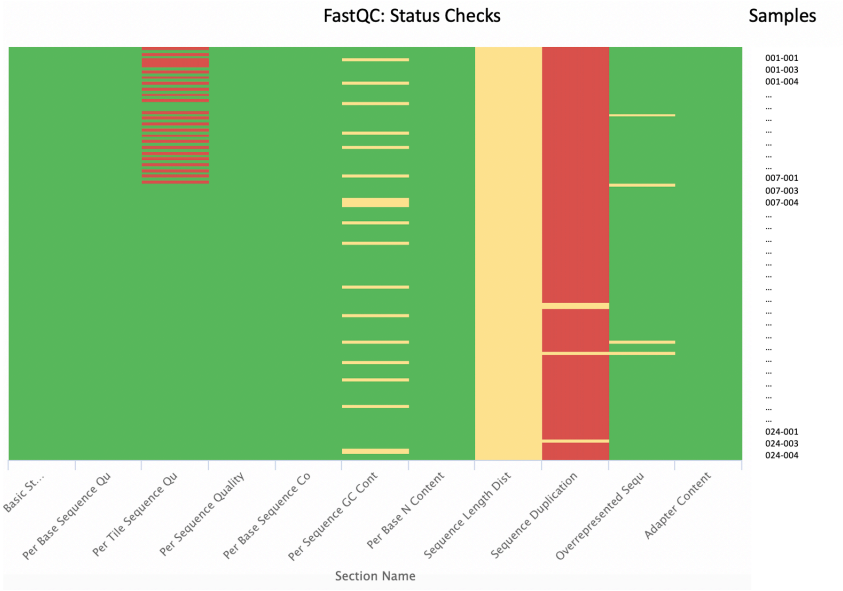


Figure 2.5: Processed macrophages RNA-seq data MultiQC report.

data is generated by Illumina platform. Figure 2.6 is an example of the plot given bey per tile sequence quality module; as indicated, red and yellow bars are the errors. On a flowcell lane, there are numerous tiles [23] that are defined as a small imaging region of view by the camera according to their coordinates on a lane, as shown in Figure 2.7. The structure inside a lane and the number of tiles are not fixed; they may be different between the different Illumina RNA-seq equipment. Inside a tile, there are massive short reads whose qualities decide the quality of the tile they are in. Errors in per tile sequence quality module indicate that the averaged *Phred* score of short reads inside a certain tile is more than five less than the mean for the same bp position across all tiles. Possible reasons for errors existing in this module suggested by *FastQC Help* are transient

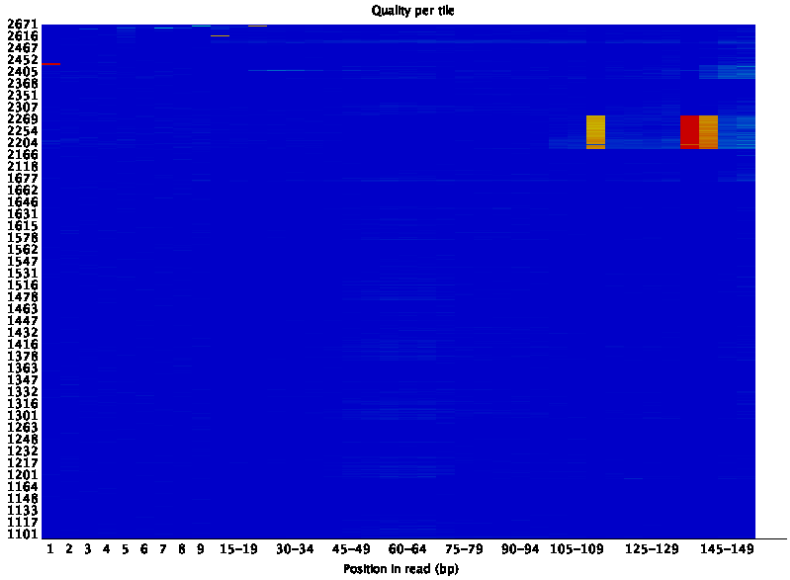


Figure 2.6: Per tile sequence quality plot.

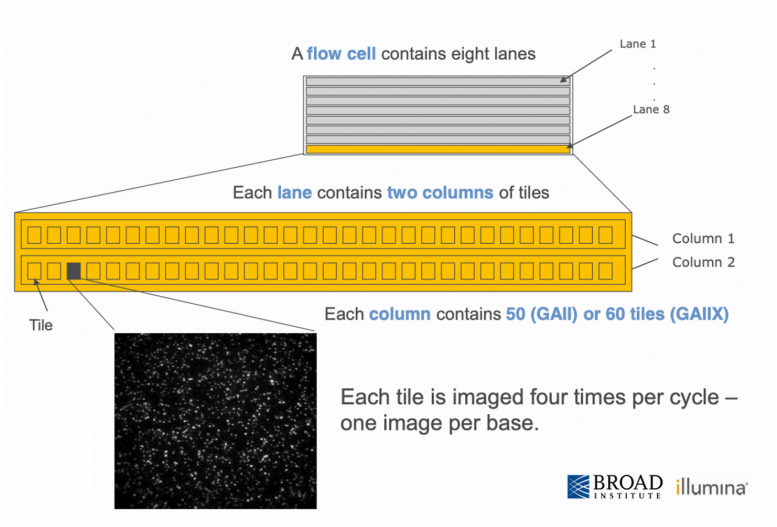


Figure 2.7: Illumina flow cell structure [23].

problems such as bubbles going through the flowcell, or permanent issues such as smudges on the flowcell or debris inside the flowcell lane [18]. Given that errors in per tile sequence quality module only in our macrophage data were generated by Lane 4. Besides, these errors occurred in the same bp positions. Thus, errors may be caused by the smudges or debris on specific regions inside the Lane 4.

To improve per tile sequence quality of our data, position-based tool *FilterByTile* from *BBMap* [24] was implemented. The reason why *FilterByTile* was applied after *TrimGalore!* and *Cutadapt* is that bases with errors of per tile sequence quality in our dataset were at the start bps (1-20bps) or the end (100-150bps) of short reads in our macrophage RNA-seq data. With the expectation that these bps were possibly included in adapters and overrepresented sequences, errors in per tile sequence quality were corrected after application of *Trim Galore!* and *Cutadapt*, this workflow design could be more time-saving and labor-saving. In our macrophage dataset, there are 1575 tiles in each lane. They are with a size of 32000×64000 micrometer^2. *FilterByTile* divides each tile into microtiles with the same size (default 500×500) to avoid removing all the sequences inside tiles with low Phred quality. Short reads within any microtiles that are unsatisfied with the thresholds of read quantity, quality, uniqueness, error-free probability thresholds will be removed. These detailed information can be checked in the report given by *FilterByTile* [24].

However, only implementing *FilterByTile* once cannot get rid of all the errors completely. Thus, we applied *FilterByTile* once again on the processed RNA-seq data; although this time errors can be removed, the loss of the data needs to consider. About 6% short reads were trimmed in the second implementation of *FilterByTile*. The trade-off between the quality of RNA-seq data and the number of short reads inside the RNA-seq data needs more investigation. To keep the information as much as possible, we used data only processed by *FilterByTile* once to do the downstream analysis. As shown in Figure C.1 and C.2, the number of the total sequences in the processed dataset was compared to that of the raw dataset. As indicated, the median of the number of reads remained in the dataset processed by FilterByTile only once was about 92.8%. However, if the RNA-seq data generated by Lane 4 processed by FilterByTile twice, the median was 85.6%. Thus, the secondary treatment on the RNA-seq data by FilterByTile caused about 7.2% loss of the data. Besides, FilterByTile twice still could not get rid of all the errors in the FastQC module per tile sequence quality, as shown in Figure B.1. Thus, in our study, we keep the remained errors in this FastQC module and used the RNA-seq data processed by FilterByTile once for the following DEGA. A better solution for tackling this kind of error should be studied to provide a more efficient way to improve the quality of the RNA-seq data. Also, before the

sequencing, the equipment like flow cell should be examined to ensure no bubbles or debris in lanes.

2.3.4 AWK

The remaining quality problems in our macrophage RNA-seq dataset were failures in per base sequence quality and the per base sequence content modules. These errors mainly exist at the beginning of the end of short reads. At the final step of QC, *AWK*, a domain-specific language designed for text processing [25] was selected to complete the workflow. It is a suitable and convenient choice for trimming bps on specific positions since the *fastq* file that stores RNA-seq data can be read as text, and certain bps in short reads can be trimmed as characters. In our study, based on the information provided by per base sequence content, bp 1-20 and bp 140-150 were trimmed from each short reads according to the FastQC report.

After applying the workflow composed by mentioned tools to process the RNA-seq data of the macrophage dataset, the MultiQC report shown in Figure 2.5 indicates that even though there were still errors in module per tile sequence quality and sequence duplication, the quality of our RNA-seq library has been improved a lot. The dynamic range of RNA-seq is more expansive than other sequencing technology, such as microarray. The range is from 0 to 1e7. One of the advantages of RNA-seq is its ability to detect RNA with relatively low expression levels; however, this merit can also cause a problem. That is the over-expression of sequence that already possesses a high expression level, and this will be reflected as an error in the sequence duplication FastQC module. In the studies on RNA-seq, this error may indicate the quality of RNA-seq is good. Thus, the errors in the sequence duplication FastQC were not removed in our study to avoid missing useful information in the RNA-seq data.

In summary, based on the comparison between the status check shown in Figure 2.5 and 2.4, most of the errors and warnings have been removed, this indicates that the quality of the raw RNA-seq data has been improved by our QC pipeline. Nevertheless, there are still errors in per tile sequence quality and sequence duplication, and warnings in per sequence GC content and sequence length distribution. As we mentioned, the errors remained in the per tile sequence quality need to be further study and the removal of them will cause more than 7% loss of the total read amount. Thus, the errors were kept in the processed RNA-seq data. Also, errors in the sequence duplication may explain that specific transcripts possessed much higher expression than others in our RNA-seq dataset, thus, the errors were also left since the correction of them may cause the loss of useful and meaningful information and bias in the alignment of the data. In a nutshell, errors of per tile sequence quality and the sequence duplication were kept to ensure enough information can be gained and used for DEGA.

As for the warnings in the sequence length distribution, they were caused by the QC process. After getting rid of sequences with low quality, each read's length is impossible to be identical (150bp or 151bp). But FatsQC will raise warnings if all sequences are not the same length. FastQC supposes that the averaged GC content of all reads should be in a normal distribution, however, there may be a wider or narrower distribution of mean GC content among all transcripts due to the high throughput of specific transcripts in RNA-seq. This can cause deviations of the observed distribution compared to an idealized normal distribution. Thus, these warnings will not cause bias in the following DEGA. To conclude, even though there are remaining errors and warnings, the processed RNA-seq data quality is still reasonable for the DGEA.

2.4 Differential Gene Expression Analysis

In recent years, researchers focus on the refinement of the transcriptome within an organism and the variations in the gene expression level between various phenotypes. They applied a method called differentially expressed gene analysis (DEGA) to investigate the changing expression levels of each transcript of cells during development and under different conditions. Thus, the DEGA can helps researchers better understand the interactions between cells and various kinds of biomaterial.

Except for the quality control (QC) on the raw data, a complete DEGA should include the following steps: the alignment of the RNA-seq data to the reference genome/transcriptome by read aligner, the quantification of the expression by expression modeler, and the differential gene identification by differential expression tools. Table 2.2 presents the tools that can be used for each step of DEGA in the current stage. DEGA can be accomplished by pipeline composed of these tools.

Read Aligner	STAR, Kallisto, Salmon, Sailfish, Tophat2, SeqMap
Expression Modeler	BitSeq, Kallisto, Salmon, Cufflinks, Stringtie
Differential Expression Tool	edgeR, DESeq2, Ballgown, SAMseq, Sleuth, NOISeq

Table 2.2: Tools for DEGA.

There are many studies on the comparison between pipelines. In Williams et al. 's research, they implemented 495 unique differential gene expression pipelines on a PBMC RNA-seq library. They compared the GSEA results with four significantly differentially expressed gene datasets to evaluate the performance of them [26]. One of the criteria evaluating these pipelines' performance is the recall, which was calculated as the number of the same significantly differentially expressed genes (SDEGs) in the test RNA-Seq dataset with the reference dataset, divided by the number of SDEGs in the reference dataset. Also, their study's result indicates the choice of differential expression tool (DET) exhibited a more substantial impact on the performance of pipelines than the choice of aligner and expression quantification method. Pipelines composed of DET DESeq2 could gain high recall values in all the four datasets. In the same year, Juliana et al. compared pipelines consist of the mentioned tools. They did not identify which specific pipeline or tool could gain the optimum results in all performance measures in their experimental conditions. Nevertheless, they demonstrated that the impact of aligners on the final DEGA was minimal. Besides, in the study of Costa-Silva et al. on the comparison of DEGA pipelines composed of different DETs, including *NOIseq*, DESeq2 and *limma+vomm* methods presented the best individual results with 95%, 95% and 93% of Specificity and 80%, 84% and 81% of True Positive Rate, respectively [27]. The Specificity denotes the percentage of those genes that were not differentially expressed and were correctly identified as not differentially expressed. The True Positive Rate denotes the percentage of DEGs that are correctly identified as DEGs. Based on the Specificity and True Positive Rate values, DESeq2 could give more accurate results. However, if pipelines consist of DESeq2 perform best in all situations is still unclear.

Thus, the choices of tools should be determined by the demands and the aim of our study. To gain more accurate results, the DEGA pipeline of our macrophage should be designed in gene-resolution. Charlotte et al. have found that both abundance estimation and statistical inference of gene-resolution analyses are often more accurate and interpretable than those of transcript-solution analyses [28]. Besides, they found that conventional gene counting approaches may cause an inflated false discovery rate (FDR) in contrast to methods aggregating transcript-level counts. The precise transcript-level estimation and inference is the key to deriving appropriate gene-level results. Moreover, transcript-level misestimation can propagate to the gene level. Based on their results, the DEGA pipeline used on our macrophage dataset was designed on gene-level derived from transcriptome-level. To this end, *Kallisto* [29] and *Tximport* [28] were selected as read aligner and transcript-level to gene-level converter, respectively. Moreover, FDR (type I errors) was expected to be as low as possible in our study. Thus, the DET chosen is DESeq2; according to the comparison result of controlling FDR between various differential expression, DESeq2 is the algorithm that often achieved the highest sensitivity in controlling the type I error [30].

2.4.1 Read Aligner: Kallisto

In this study, Kallisto was selected as aligner [29]. The working process of Kallisto is illustrated in Figure 2.8. Kallisto uses fast hashing of k-mers and a transcriptome de Bruijn graph(T-DBG)

constructed from the k-mers present in the transcriptome to gain accurate pseudo alignment result of the transcriptome.

Firstly, a T-DBG will be created from a read (in black) and its possible origins, three overlapping transcripts corresponding to different colors indicated in Figure 2.8.a. The colored paths' union covers all edges of the T-DBG, and it constructs an index (v1, v2, v3, ...) for k-mers, which are hollow circles shown in Figure 2.8.b. This united path also induces a k-compatibility class for each k-mer. For each k-mer, there is a k-compatibility class induced by the path that covers the transcriptome. E.g., the most left k-mer has a k-compatibility of all the three transcripts, the three k-mers on the most top in the T-DBG graph have a k-compatibility of only blue and pink transcripts. In Figure 2.8.c, the k-mers of the three colored transcripts are hashed to the read generating the black nodes. Then, Kallisto uses a skipping method to determine the k-compatibility of the read. Black dashed lines presented in Figure 2.8.d connects the hashed k-mers(black nodes), but skips those redundant k-mers, which are defined as the k-mers with the same k-compatibility class (i.e., V_1, V_2, and V_3). The k-mers are also classified into the same 'equivalence class.' Thus, black dashes lines will only keep the first node in the same equivalence class(i.e., black nodes V_2 and V_3 will be skipped). Finally, Kallisto takes the k-compatibility classes of its constituent k-mers to decide the read's k-compatibility class.

With this pseudo-alignment process, the speed of Kallisto is higher than the conventional aligners. Rather than aligning each read to the reference transcriptome, Kallisto gives all the transcripts that are compatible with each read. The process indicated in Figure 2.8.d is what makes Kallisto faster but makes it keep the accuracy since k-mers in the same equivalence class will not change the intersection result, and looking them up in the hash provides no new information [31].

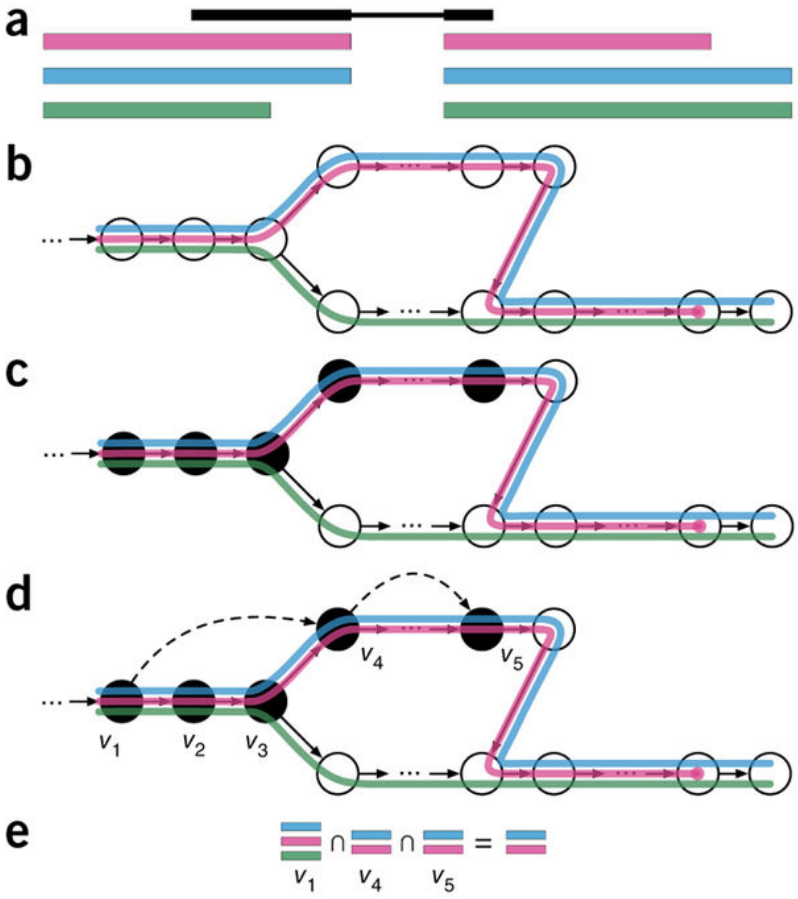


Figure 2.8: Kallisto illustration [29].

Reference Transcriptome

The reference transcriptomes used to build the indices is a data set based on the December 2013 Homo sapiens high coverage assembly GRCh38 from the Genome Reference Consortium [32], which is available under cDNA on the Ensembl website. The data set is composed of gene models built from both of the gene-level alignments of the human proteome and alignments of human cDNAs using the cDNA2genome model of exonerate [33].

Input and Output

The input of *Kallisto* is the RNA-seq data and the indices generated from the reference transcriptome. After the alignment is completed, the *Kallisto* will return *.tsv* and *.h5* files containing the following information shown in Table 2.3, in which **est_counts** denotes the number of a transcript in the RNA library.

target_id	length	eff_length	est_counts	tpm
ENST00000632684.1	12	13	0	0
ENST00000335895.12	897	742.03	577.455	438.439
...	...	...	...	...
ENST00000409020.5	1680	1525.03	38.4718	14.2127
ENST00000359683.8	1626	1471.03	10.2012	3.907
ENST00000400723.7	2051	1896.03	0	0

Table 2.3: Kallisto output matrix.

2.4.2 Transcript-level to Gene-level Converter: Tximport

Tximport developed by Charlotte et al. [28] is an R package that can transfer the transcript-level read counts to gene-level result. The accuracy of the differential gene expression analysis is expected to be improved by *Tximport*.

2.4.3 Differential Expression Tool: DESeq2

DESeq2 [30] is a successor to *DESeq* method [34], an error model that employs the negative binomial (NB) distribution. DESeq2 method does not suppose that there are any differentially expressed genes, instead, it uses the NB distribution and statistically test whether the observed difference in read counts of a gene is more significant than the natural random variation to decide if a gene is differentially expressed [30]. *DESeq* assumed that the number of reads of a sample j are assigned to gene i, K_{ij}, can be modeled by the NB distribution as Equation 2.1, where the mean μ_{ij} and the dispersion σ_{ij} can be gained by fitting the read count matrix with a generalized linear model (GLM). Based on *DESeq* method, DESeq2 uses shrinkage estimators for dispersion and fold change and improves its performance in terms of gene ranking and visualization, hypothesis tests, the regularized logarithm transformation, and clustering of overdispersed count data. One of the results returned by DESeq2 is *padj*, which is derived from Benjamini-Hochberg (BH) adjustment, and it can limit FDR, i.e. if 1% FDR is acceptable, all the genes with an adjusted p value (*padj*) <0.01 are differentially expressed genes (DEGs).

$$K_{ij} \sim NB(\mu_{ij}, \sigma_{ij}^2) \tag{2.1}$$

Input and Output

A gene-level count matrix K with one row for each gene i and one column for each sample j will be generated by *Tximport* from those *.tsv* or *.h5* files given by *Kallisto*.

DESeq2 will model the count matrix and return differential expression analysis results in *DESeqDataSet*, which are DataFrame (DF) objects. For example, by calling the function *DESeq()* and *result()*, users can gain a *res* DF as shown in Table 2.4 that contains results, including $p - value$, adjusted p-value (*padj*), and log2FoldChange (Log2FC). There were more than 13,000 genes identified and analysis in our macrophages. We aimed to analyse if the DEGs of macrophages exposed to different surfaces can be related to the complications like breast cancer or capsular contracture. Thus, in this disease/symptom association project, a stringent cutoff 0.01 was set as the threshold to decide which genes were differentially expressed. The 0.01 *padj* denotes that only 1% FDR will be tolerated when defining the differentially expressed genes.

2.5 Visualization Patterns Derived from DESeqDataSet

It is impossible to get a direct view on the result of DEGA since these results will be only presented in the numeric *DESeqDataSet*. Thus, visualization patterns including MA plot, principal component analysis (PCA) plot, bar plot of *PC* loading, correlation coefficient heatmap, and gene expression pattern heatmap derived from the *DESeqDataSet* will be applied to interpret results of DEGA. These plots will only includes genes which are recognized as DEGs ($padj < 0.01$).

	baseMean	log2FoldChange	lfcSE
	<numeric>	<numeric>	<numeric>
ENSG00000000003	2.03742659724827	0.121123138669929	0.384932562133665
ENSG00000000419	258.678607016287	0.091840712649386	0.114019237056228
...	...	...	...
ENSG00000284746	0.316075461535429	0.0906610921427894	0.51846310965637
	pvalue	padj	
	<numeric>	<numeric>	
ENSG00000000003	0.503177818376351	0.652364376714041	
ENSG00000000419	0.371886531018832	0.531940188727362	
...	...	...	
ENSG00000284746	0.579698550514588	NA	

Table 2.4: DESeq2 result matrix structure.

2.5.1 MA Plot

An MA plot puts the variable M on the y-axis and A on the x-axis, and it can give a quick overview of the distribution of the genomic data [35]. In our study, the MA plot is plotted from *res* DF shown in Table 2.4. In our study, M (on Y-axis) denotes the log2 fold change in the expression levels of the same gene under various conditions; and A (on X-axis) is the average of the normalized counts gained from the median of ratios [36]. In the MA plot, genes with $padj < 0.01$ are denoted by red dots that are classified as differentially expressed genes; others are shown by black. Genes are ranked according to their *padj* value, and genes with the five smallest *padj* will be annotated by their symbols in a red circle.

2.5.2 Principal Component Analysis Plot

The principal component analysis (PCA) plot is also a scatter plot, which can be used to indicate the clustering of the samples from different experimental conditions or phenotypes. PCA is commonly used for dimensionality reduction in exploratory data analysis and emphasizing variation within a dataset [37]. In a DEGA, there are thousands of genes in the data of each sample. Each gene can be seen as a dimension of the genomic data; thus, it is impossible to visualize the data points and find patterns between them in a 2D or 3D figure. By PCA, data points are transformed

into a new coordinate system and projected onto only a few principal components (PCs) to gain a lower-dimensional dataset. PCs are calculated from the linear combination of variables expression levels of each gene as shown in Equation 2.2, in which i denotes the i-th principal component (PC); p is the p-th variable (gene); $a_{i1}X_1$ indicates the weight or the loading of each variable; and Y_i is the i-th PC [38].

$$Y_i = a_{i1}X_1 + a_{i2}X_2 + ... + a_{ip}X_p \tag{2.2}$$

The first principal component (PC) are defined as a direction that accounts for the biggest variance of the projected dataset. Data points in the plot are shaped by their index and colored by the group they belong to. In our study, two or three dimensional PCA plots were generated from varianceStabilizingTransformation (*vst()*) result. The number of dimensions was decided by the comparison on the variation of principal components. *vst()* can provide variance-stabilizing-transformed (*vst*) values in its assay slot as indicated in Table 2.5. In the following report, it will be addressed as *vsd*.

	001-003	005-001	009-003	...	023-004
ENSG00000000003	6.738703	6.852746	6.899787	...	7.017766
ENSG00000000419	8.589393	8.600099	8.806535	...	8.725447
...	...	...	...	...	...
ENSG00000284746	6.738703	6.738703	6.738703	...	6.738703

Table 2.5: DESeq2 vsd matrix structure.

2.5.3 Bar Plot of Principal Component Loading

As shown in Equation 2.2, PCs are decided by the weights/loading of each variable, and variables have high positive/negative loadings on each PC contribute most strongly to each PC. To evaluate the proportion of DEG that have large contributions on a few first PCs like $PC1$ and $PC2$ (and $PC3$), DEGs were ranked by their absolute PC loading in a bar plot, respectively. These plots can provide a direct view of the overall distribution of PC loading values of each DEG. By this bar plot, we can judge if only a small part of DEGs contribute strongly to PCs and have insights into which specific DEGs are more essential for distinguishing samples from phenotypes. DEGs with the top 20 absolute value of $PC1$ and $PC2$ loading were also ranked by the absolute value of their loading value and plotted in another bar plot. The higher the individual element loading, the stronger its association to the respective PC. With this bar plot, the contribution of each gene can be compared.

2.5.4 Correlation Coefficient Heatmap

The correlation heatmap was used to present the Pearson's correlation between samples based on expression levels of their DEGs. The origin used to generate the plot the correlation heatmap is also the *vst* data stored in Table 2.5. The Pearson's correlation between two samples is calculated by Equation 2.3 [39]. In this equation, r_{xy} is the Pearson's correlation between two samples x and y; n is the sample size, which is the number of DEGs in our case; x_i and y_i denotes the value corresponds to i-th DEG; $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and analogously for $\bar{y}$.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.3}$$

2.5.5 Gene Expression Pattern Heatmap

The gene expression pattern heatmap (GEPH) is another common visualization method in DEGA. In this plot, the rows are DEGs, and the columns are samples; both clustering of DEGs and samples

determined by Euclidean distance (ED) calculated from Equation 2.4 can be checked. Index u denotes the u-th DEG in the computation of the ED between two samples X and Y; index u denotes the u-th sample when computing the ED between two DEGs X and Y. This plot is also derived from the *vsd* DF. In the GEPH, samples are annotated with their experimental conditions. By the GEPH, we can check if the expression levels of different groups of genes of various samples different can be clustered according to the experimental conditions of these samples or not.

$$D_{EU} = \sqrt{\sum_{u}(X_u - Y_u)^2} \tag{2.4}$$

2.6 Geneset Enrichment Analysis

Although the analysis on single genes could elucidate the development trends of macrophages' immune system interacting with flat- or textured-surface SIMs, we cannot get a general conclusion on what gene sets are dominant in DEGA of macrophages. A gene set is a group of genes that share an identical biological function, chromosomal location, or regulation. Thus, for further exploration of the macrophage RNA-seq data, the priority would be the geneset enrichment analysis (GSEA). GSEA was first proposed by Aravind et al. in 2005 [40]. This method employs a weighted Kolmogorov–Smirnov (KS) like a statistic to determine the degree to which a reference gene set is overrepresented at the top or bottom of the entire ranked list of the genes to be analyzed. GSEA has demonstrated its power in cancer-related data sets like lung cancer. It revealed many common biological pathways in two independent studies on patient survival in lung cancer, even though the single-gene analysis result has indicated little similarities between them. By GSEA, we could get more clues if the reference gene sets from the pathway and gene ontology databases are overrepresented in our DEG list. To this end, GSEA [41], a freely available software package where the GSEA method is embodied, was selected to do the analysis. GSEA method is also the basis of the Connectivity map (CMap) [42], which is used to explore the relationship between disease, cell physiology, and treatment.

2.6.1 Input and Data preparation

The input of GSEA is either composed of a *.gct* file, a *.cls* file or a single *.RNK* file. The *.gct* file indicated in Table 2.6, and the *.cls* file shown in Table 2.7 contains the information of the expression level of DEGs and phenotype labels of each sample, respectively. The *.gct* file includes the following information, the number of DEGs and samples and the normalized expression level value of DEGs from each sample. GSEA needs the count value of genes in RNA-seq data normalized before inputting since the normalized data is more robust for the downstream analyzed. It is fairer to compare the abundance of a gene between samples. The normalization can be completed by the GenePattern DESeq2 model [43] or R package DESeq2. The normalization DESeq2 implies the median ratio method.

In our case, GSEA was focused on only DEGs, and the DEGs were gained from the normalized data of all the genes. Only select the data of DEGs will cause bias in the median-ratio normalization of DESeq2. Thus, a pre-ranked gene list is contained in a *.RNK* file was input to do the geneset enrichment. A *.RNK* file shown in Table 2.8 is composed of two columns, the first column is the HUGO symbol of genes, and the second column is the metric to depict their ranking. The matrix can be the Log2FC value gained from DEGA, which was applied in our study. To reduce the number of DEGs input into the GSEA and get a more accurate result, a cutoff of Log2FC 1.5 was set to select DEGs. The output of the GSEA respot will not denote phenotypes if the ranking list is pre-ranked. Instead, it will employs *na_ pos* and *na_ neg* to indicated the phenotypes. For example, DEGs in the ranking list, which are with the positive score, is the up-regulated genes in the Text96h phenotype, thus, in the report, the *na_ pos* is the phenotype Text96h, and the *na_ neg* is the phenotype flat96h. The gene sets with a positive score in the phenotype text96h

mean that those gene sets are most enriched in them [44]. MSigDB **Hallmark** [45] was chosen as reference gene set.

#1.2						
3465	32					
Name	Description	001-003	005-001	009-003	...	023-004
ENSG00000000460	n/a	5.8915	8.2926	6.0122	...	11.4939
ENSG00000001036	n/a	86.4083	92.2556	96.1952	...	148.4627
ENSG00000001497	n/a	25.5297	22.8048	16.0325	...	11.4939
...	...	...	...	...	...	...
ENSG00000284681	n/a	6.7387	6.8527	6.8998	...	7.0178
ENSG00000284746	n/a	0	0	6.8540	...	3.8757

Table 2.6: *.gct* file.

32	2	1					
#	Flat96h	Text96h					
Flat96h	Flat96h	Flat96h	Flat96h	...	Text96h	Text96h	Text96h

Table 2.7: *.cls* file.

NPAS2	3.899411
SPATA17	3.729204
TMEM92	3.461559
...	...
ATP10A	-4.700706
HSPA1B	-24.031256

Table 2.8: *.RNK* file.

2.6.2 Output

The analysis report will be presented in the HTML format available in website. The contents in the report are enrichment score (ES), normalized enrichment score (NES), false discovery rate (FDR), nominal p-value, and visualization on them.

Enrichment Score

Enrichment score(ES) is a primary result, which determines if a gene set is overrepresented at the top or the bottom of the ranked gene list. Firstly, the genes in the expression data set D that includes N genes and k samples are ranked to form the ranked gene list $L = \{g_1, ..., g_N\}$ according to their ranking score, $r_j = r(g_j)$ generated by the ranking matrix, e.g. signal-to-noise ratio (S2N; the default measure in GSEA). Then, the ranked gene list L will be walked through, if the gene is in a reference gene set S, the priori that contains N_H genes from the selected MSigDB, the running sum will be increased, otherwise, it will be decreased. In the calculation, an exponent p is used to control of the weight.

The ES, the maximum deviation from zero of P_{hit} - P_{miss}, is evaluated by the group of genes in S (P_{hit}; Equation 2.5) weighted by their ranking score and the fraction of genes not included by S (P_miss; Equation 2.5) present up to a given position i in L. Due to the constant step size of the walk in L, the ES starts and ends with 0. If the ES is positive, the gene set enrichment (GSE) is at the start of L; and a negative ES means that the GSE is at the bottom of L. The

larger the absolute value of ES is, the larger the enrichment degree of the priori S is. In addition, an enrichment plot (EP) in the GSEA report could indicate how ES of a gene set changes when walking through the ranked gene list L.

The exponent p is also an important parameter in the calculation of ES score. If $p = 0$, ES(S) reduces to the standard KS statistic; when $p = 1$ (default), genes in S are weighted by their ranking scores normalized by the sum of the ranking scores over all of the genes in S. In our case, we set $p = 1$ for DEGs of the macrophages dataset.

$$P_{hit}(S,i) = \sum_{\substack{g_j \in S \\ j \leq i}} \frac{|r_j|^p}{N_R},\ where\ N_R = \sum_{g_j \in S} |r_j|^p \tag{2.5}$$

$$P_{miss}(S,i) = \sum_{\substack{g_j \notin S \\ j \leq i}} \frac{1}{(N - N_H)} \tag{2.6}$$

$$ES(S,i) = P_{hit}(S,i) - P_{miss}(S,i) \tag{2.7}$$

For each gene in the ranking gene list L, the ES on this position is equal to $ES(S,i)$ calculated by Equation 2.7. The final result of ES is the largest deviation from zero across all the positions in the ranked gene list.

Normalized Enrichment Score

Normalized enrichment score (NES) is a statistic to exam gene set enrichment results. It is calculated by scaling the actual ES with by the average of ESs gained in a number of permutations as indicated in Equation 2.8. In each permutation, the original phenotype labels of samples are assigned at random, and the genes from D will be also reordered to determine a new ES.

$$NES = \frac{actual\ ES}{mean(ESs\ against\ all\ permutations\ of\ the\ dataset)} \tag{2.8}$$

Nominal P Value

After repeating 1000 (default value in GSEA) permutations, a histogram of the corresponding ESs ES_{NULL} will be created. Using the positive/negative portion of the observed ES distribution from ES_{NULL} can estimate the nominal P value for S. The nominal P value denotes the statistical significance of the ES.

False Discovery Rate

For each S and 1000 fixed permutations π of the phenotype labels, reorder the genes in L and determine ES(S, π). Normalize the ES(S, π) and the observed ES(S) by the mean of the ES(S, π) to yield the normalized scores NES(S, π) and NES(S).

For a given $NES(s) = NES^* \leq 0$, its FDR q value is equal to the ratio of the percentage of all (S,π) with $NES(S,\pi) \leq NES^*$ divided by the percentage of observed S with $NES(S) \leq NES^*$, and the similarly if $NES(S) = NES^*0$.

2.6.3 Criteria

In our study, we select gene sets with nominal P value ≤ 0.05, FDR (Q) ≤ 0.25 as significantly enriched gene sets (SEGSs). Even though previous research suggests that selecting SEGSs with FDR ≤ 0.05 could provide a more robust analysis result [44], more lenient thresholds like 0.25 can be used if there are not massive enriched gene sets. GSEA [41] also supports applying this value since a stringent FDR like 0.05 may lead to the overlook of potential SEGSs, and an 25% FDR Q indicates a 75% validity of the result, which is reasonable to find candidate gene sets to

propose hypothesis for the further research. However, given the reference dataset of GSEA are curated from different sources with various methods, these reference dataset may lack coherence. Thus, GSEA results need to be verified by further studies like the quantity of the expression level of specific cytokines.

2.7 Connectivity Map Query

Connectivity Map (CMap) is an online tool that uses cellular responses to perturbagen to find relationships between diseases, genes, and therapeutics [42]. The CMap data aims to extract a signature that represents genes turned on or off upon treatment with chemical or genetic perturbagens. Chemical perturbagens consist of small molecule compounds, including drugs and tool compounds. Genetic perturbagens include libraries of CRISPR/Cas9 constructs, short hairpin RNAs (shRNAs), and open reading frames (ORFs) used to edit, knockdown, or overexpress genes, respectively [42]. More than one million gene expression signatures of various cell types with different perturbagens are contained in the CMap database. CMap query can compare for the similarity between the user-supplied gene list to its database.

2.7.1 Reference Dataset: Touchstone

Touchstone was selected as the reference dataset. It is a high-throughput transcript abundance reference dataset generated by L1000 assay measures [46]. Compared to L1000, RNA-seq suffers from technical complexity in library preparation and the inability to detect non-abundant transcripts without deep sequencing. Also, the reagent cost for L1000 is considerably less than the cost of RNA-Seq. Moreover, the profiling of gene expression level of RNA-seq and L1000 were highly correlated. Thus, RNA-seq was replaced by L1000 in the CMap project. The "**L**" in **L1000** refers to the landmark gene whose expression is most informative to characterize the transcriptome, and it is measured directly in the assay, and "**1000**" denotes the 978 "landmark" genes from human cells. The L1000 assay directly measured or inferred the expression levels of 12,328 genes, among these genes, 10,174 genes were identified as genes which can be directly measured or well-inferred from the landmark genes. This subset is referred to as the Best Inferred Gene (BING) space, comprised of 978 landmarks and 9,196 well-inferred genes. Only expression profiles of BING space are contained in the Touchstone and are used to determine the similarity between the reference database and the user-supplied gene list.

2.7.2 Input

In our study, DEGs identified by DESeq2 were used for CMap query. To investigate the differences and similarities between different donors' immune response, DEGA was also executed on the RNA-seq data of each donor, respectively. The selection of DEGs input in the CMap query mainly depended on the Log2FC value. Since the dynamic range of the gene expression value in our macrophages is quite wide, if we just select the top and bottom 50 DEGs of each donor, the DEGs are more easily selected with a different Log2FC value. Also, to make our CMap result more biologically meaningful, a different threshold of Log2FC (starting from 1.2) were tested to filter the DEGs used in the query. In the past decade, the combination of the fold-change (FC) and the p-value were used to promote the accuracy of DEGA [47–49]. Both of *padj* and *FC* could decrease false positives. Thus, the threshold of Log2FC was set as large as possible, but make sure there are at least ten up-regulated and down-regulated DEGs input into the query. To improve the recognizability of input genes, these genes from the DEGA result were pre-filtered by the BING gene space of CMap. Then, the top and bottom genes from the ranked DEG list were input to the query.

2.7.3 Output

The output report contains the connectivity score (CS) of each perturbagen. The higher the CS a perturbagen has, the more similar it is with the input genome. By CMap, we can predict which perturbagen can be related to our DEGs. The CS of a query compared to the CMap database is computed by the following steps. More details can be found on the website [50].

First, the weighted connectivity score (WTCS) represents a similarity measure based on the weighted Kolmogorov-Smirnov enrichment statistic (ES), which is also the basic algorithm of GSEA [41], will be gained. WTCS is calculated as Equation 2.9, where r is a particular signature of the reference database, ES_{up} is the enrichment of q_{up} in r and ES_{down} is the enrichment of q_{down} in r. It ranges between -1 and 1, and it is a composite, bi-directional version of ES.

$$W_{qr} = \begin{cases} (ES_{up} - ES_{down})/2 & \text{if } \mathrm{sgn}(ES_{up}) \neq \mathrm{sgn}(ES_{down}) \\ 0 & \text{otherwise} \end{cases} \tag{2.9}$$

Then, to compare the connectivity score (CS) across cell types and perturbagen types, WTCSs are normalized to account for global differences. The normalized CS (NCS) is computed as described in Equation 2.10, where $w_{c,t}$ is the WTCS of a query list compared to cell type c and perturbagen type t, $\mu^{+}_{c,t}$ and $\mu^{-}_{c,t}$ means of the raw positive and negative WTCS, respectively. This computation procedure is similar to that used for NES in GSEA.

$$NCS_{c,t} = \begin{cases} w_{c,t}/\mu^{+}_{c,t} & \text{if } sgn(w_{c,t}) < 0 \\ w_{c,t}/\mu^{-}_{c,t} & \text{otherwise} \end{cases} \tag{2.10}$$

It is also useful to judge if the connectivity between user-supplied q and a signature r is significantly different from that observed between other queries and r. Tau (τ) compares an observed NCS between a query q and a signature r to all other queries in a reference database and r; it is computed by Equation 2.11. $|ncs_{i,r}|$ denotes the NCS for signature r relative to the i-th query in the reference compendium of queries (Qref), N is the number of queries in Qref [51]. Qref is comprised of queries obtained from exemplar signatures of Touchstone. τ ranges from -100 to 100. A τ of 95 indicates that only 5% of reference perturbagens showed stronger connectivity to the user-supplied query. A low τ would suggest connections are not unique. A positive τ indicates a similarity between the query and the signature perturbed by a molecule; while a negative τ means that the two signatures are opposing (i.e. genes that are decreased by treatment with the perturbagen are increased in the query and vice versa).

$$\tau_{q,r} = sgn(ncs_{q,r}) \frac{100}{N} \sum_{i=1}^{N} [|ncs_{i,r}| < |ncs_{q,r}|] \tag{2.11}$$

The last step is the summarization of results observed in individual cell types. This will be helpful when figuring out connections that are across cell lines or when one is unsure which cell line to examine. A cell-summarized CS is obtained using a maximum quantile statistic shown in Equation 2.12, where $ncs_{p,c}$ is a vector of NCS for perturbagen p, relative to query q, across all cell lines in which p was profiled, and Q_{hi} and Q_{low} are upper and lower quantiles, respectively.

$$NCS_{p} = \begin{cases} Q_{hi}(ncs_{p,c}) & \text{if } |Q_{hi}(ncs_{p,c})| \geq |Q_{low}(ncs_{p,c})| \\ Q_{low}(ncs_{p,c}) & \text{otherwise} \end{cases} \tag{2.12}$$

The heatmap tool provided by CMap can summarize CS of a perturbagen across cell lines and rank summarized CSs of different queries by the median value. The recommended CS value to choose highly correlated perturbagens is +/-90. This value was adjusted to 75 - 80 based on the query result of our macrophages dataset and the heatmap.

2.8 Gene Network

Gene network construction was fulfilled by two steps: the generation of induced network module by ConsensusPathDB and the extension of the network module by Cytoscape.

2.8.1 ConsensusPathDB

ConsensusPathDB (CPDB) is an online database system for the integration of human functional interactions [52]. CPDB was first reported by Atanas et al. [53], and has been updated in 2013 [54]. It currently integrates 215541 unique functional interactions (protein–protein interactions, biochemical reactions, gene regulatory interactions) and 4601 pathways from overall 30 databases [54]. CPDB contains an induced network module which uses reference interactions to build a network according to user-supplied genes. It allows users to analyse DEGA result in terms of gene set analysis and metabolism set analysis. In our study, CPDB is used to build the induced network based on DEGs gained from DEGA. DEGs were filtered by various *Log2FD* threshold to test which cutoff is suitable to construct a network, and only binary protein interactions were selected. DEGs with different Log2FC cutoffs were tested and used to generate the gene network to make sure that there are enough DEGs contained in the gene network and provide as much information as possible while there will not be too many DEGs contained in gene networks too complicated to study.

One of the advantages of CPDB is using intermediate nodes to improve the connectivity between genes. The intermediate gene is not from the user-supplied seed gene list, while it connects two or more seed genes with each other [54]. Even though intermediate genes may not be regulated on the transcriptional level and originated from the user-supplied gene list, it could be related to the phenotype under study. By the usage of the intermediate gene, there will be significantly more interactions within the induced network. Thus, intermediate genes can improve inter-gene connectivity. In addition, the intermediate gene may also reveal the underlined mechanism of the induced gene network module. For example, if an intermediate node, which represents a transcription factor (TF), is connected to a group of seed genes through interactions, this suggests that the TF may be dysfunctional, possibly due to a mutation that does not necessarily impact the TF's expression. Z-score is used to quantify the significance of the association between an intermediate gene and the seed genes it connects. Intermediate genes with a z-score larger than the threshold 20 are allowed to generate the network in our study.

2.8.2 Cytoscape, and CyTargetLinker

The resulting network was subsequently imported and extended by Cytoscape and CyTargetLinker. TFs were added to the network to form a gene regulatory network (GRN), which is a collection of regulatory interactions between TFs and their target genes.

TFs are proteins with unique abilities and attributes that are not common in other types of proteins [55]. They directly or indirectly bind to DNA and often work in pairs or networks to regulate particular regulatory pathways. TFs are important for all eukaryotic biochemical systems. They modulate gene expression and drive regulatory programs or networks that maintain cells in dynamical microenvironment changes. Some of them also interact with ligands or hormones. The research on TFs may help decipher the complex regulatory programs that enable a single genome to specify hundreds of phenotypically distinct cell types. Thus, the research on TFs is essential for studies such as cancer therapy, stem cell differentiation, and so forth. Understanding of TFs and the elements and processes that impact their activity is one of the goals of modern life science research. TFs from the transcription factor target database Transcription Factor encyclopedia (TFe), which is a smaller scale manual literature curation project containing 1531 human well-studied TF target interactions respectively [55].

Chapter 3

Result

In this part, we will first check the PCA plot of all the samples from 4 groups, including Flat24h, Text24h, Flat96h, and Text96h to view the global clustering of samples from the macrophage dataset. Then, based on the result given by *DESeq*2, differentially expressed genes (DEGs) with $padj \leq 0.01$ of two comparison groups, Text24h VS Flat24h, Text96h VS Flat96h are selected to plot visualizations. By this, we can investigate the effects of the structure of breast SMIs on the gene expression level of the macrophages cultured on them for 24 hours and 96 hours, respectively.

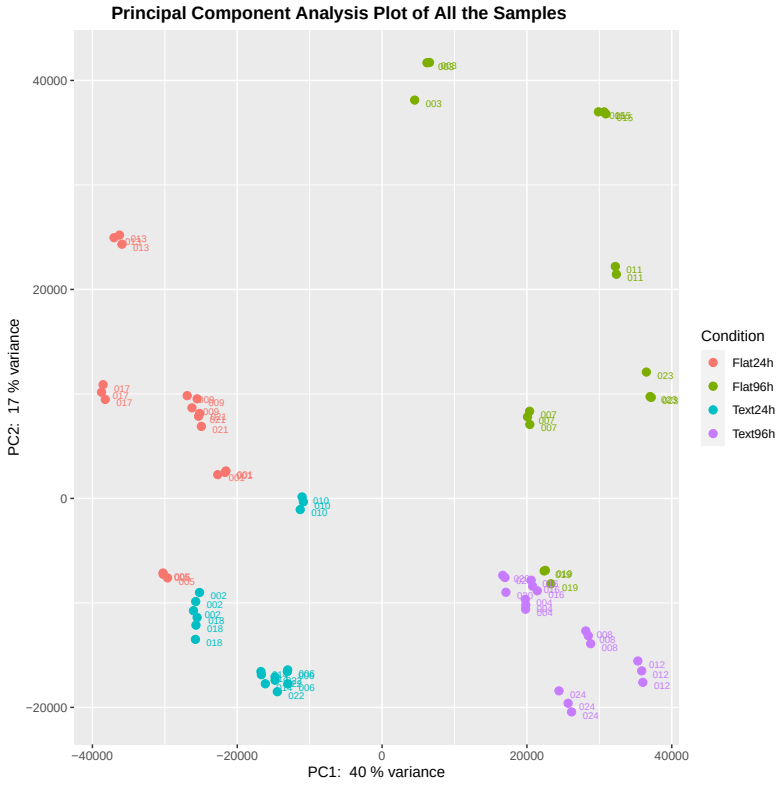


Figure 3.1: PCA plot of samples from Flat24h, Text24h, Flat96h, and Text96h.

3.1 Principal Component Analysis

3.1.1 PCA Plot of All the Samples

The PCA plot shown in Figure 3.1 is generated from the raw count matrix normalized by the median-ratio method provided by *DESeq*2. Genes whose total expression in all the samples smaller than 60 were analogous to genes that are not expressed and filtered out. The clustering of samples from these 4 groups can be found in Figure 3.1. Samples cultured on the flat surface (Flat24h and Flat96h) and samples cultured on the textured surface (Text24h and Text96h) were clearly distinguished from each other on $PC1$ direction. The dividing line between them is about $PC1 = 0$. Even though samples cultured on the surfaces for 24 hours (Flat24h and Text24h) and 96 hours (Flat96h and Text96h) distributed on different regions in the plot, there was no clear dividing line between them on $PC2$ direction. Sample 005 from the group Flat24h was clustered with Flat96h, and Sample 019 from Text24h was mainly composed of samples from Text96h.

3.1.2 Text24h VS Flat24h

Visualizations of the comparison group Text24h VS Flat24h is shown in Figure 3.2, which were derived from the 1587 DEGs detected by *DESeq*2. Figure 3.2(a) illustrates the distribution of Log2FC value all the genes corresponding to their expression level. Five dots in the red circles denote the genes with five smallest *padj*, which are recognized as the five most statistically significant DEGs. Among them, HSPA6 was the one whose expression level changed most intensely, and it has been significantly down-regulated in the Flat24h phenotype. Based on the study of Fagone et al. [56], variations in the expression of heat shock protein (HSP) gene family could be related to the polarization of macrophages. They found that HSP6A and other five genes showed significant up-regulation in M1 cells compared to unpolarized macrophages; however, no manifest changes in these HSP genes found in M2 cells in contrast to unpolarized macrophages. The indicative down-regulation of HSPA6 on Text24h compared to Flat24h may demonstrate that more macrophages from Flat24h than Text24h were polarized to M1 cells. As shown in Figure 3.2(b), a classification on the samples from Flat24h and Text24h could be found. Samples from Flat24h could be distinguished on the $PC2$ direction. The correlation coefficient heatmap, Figure 3.2(c), demonstrates that samples from Flat24h and Text24h cluster per their experimental conditions. This also confirms that macrophages from these two experimental conditions can be differentiated according to the expression level of DEGs gained from *DESeq*2. The expression pattern heatmap on DEGs provides clustering results on both of the DEGs and the samples. Two groups of DEGs can be found in Figure 3.2(d); they expressed variously when macrophages cultured on the flat and textured surfaces. In conclusion, the visualizations demonstrate that DEGs' expression levels detected by *DESeq*2 can distinguish samples from Flat24h and Text24h.

DEGs with high positive/negative loadings on each PC contribute most strongly to each PC. As indicated in Figure 3.1 and 3.3.b, macrophages per phenotype could be distinguished by PC1 direction. Thus, more investigation on how DEGs linearly combined the PC1 can provide more clues about which specific genes are important to distinguish the flat from the textured phenotype. Similarly, samples from the Flat24h group were also separated from each other in the PC2 direction. This may indicate that genes with higher PC2 loading are possible factors classifying samples from the Flat24h phenotype. To this end, it is important to evaluate the proportion of DEGs that have large contributions to a few first PCs like $PC1$ and $PC2$. To better investigate which genes had the largest effects on the classification of samples from these two groups, $PC1$ and $PC2$ loading of each DEGs were computed and compared in Figure 3.3.a and .b, which indicated that there was only a small part of genes possessed a relatively large PC loading value. Figure 3.3.c ranks the genes according to the absolute value of their PC loading. The genes in the bar plot are the genes with top 20 $PC1$ and $PC2$ loading. Their detailed information can be found in Table 3.1 and 3.2, respectively. The top 20 DEGs were selected since the absolute PC loading value of the first DEG was also double that of the 20th DEGs. Also, the PC loading value of the last DEGs almost did not change. Thus, the comparison was mainly executed among these 20 DEGs.

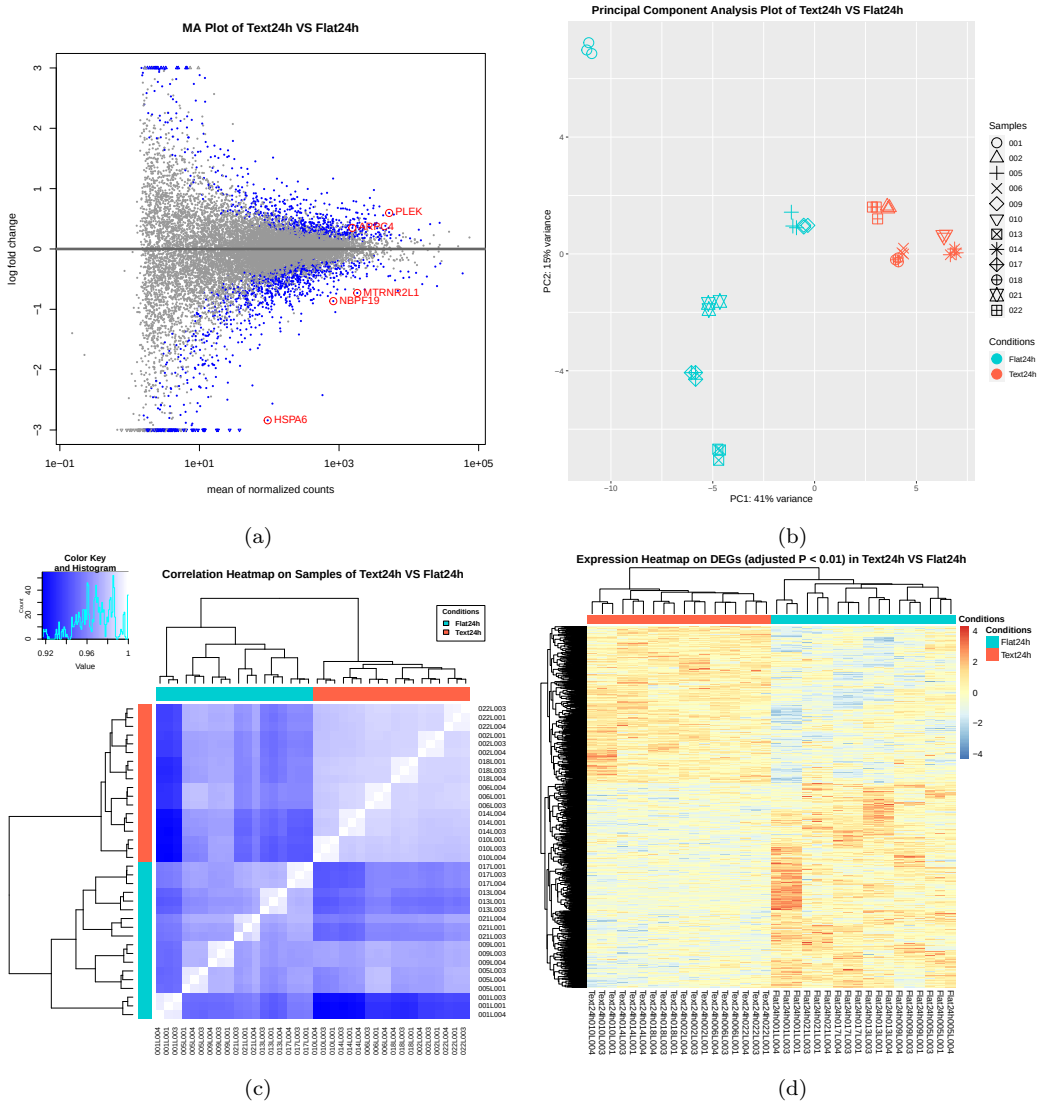


Figure 3.2: Visualization of comparison group Text24h VS Flat24h only with differentially expressed genes. (a) MA plot of Text24h VS Flat24h. (b) Principal Component Analysis plot of Text24h VS Flat24h. (c) Correlation Heatmap on Samples of Text24h VS Flat24h. (d) Expression Heatmap on DEGs (adjusted P $<$ 0.01) in Text24h VS Flat24h. (e) Bar plot of PC1 loading. (f) Bar plot of PC2 loading.

Based on the information of genes with high $PC1$ and $PC2$ loading, we could conclude that gene expression levels revealed that the textured-surface breast SMI is more likely to induce biomarkers' upregulation of breast cancer (BRCA) and the downregulation of tumor suppressors like TGFBI compared to flat-surface breast SMI. As indicated by Table 3.1, NCAPH positively contributed most to $PC1$ that divided Flat24h from Text24h. It was found upregulated in human cancer types, including BRCA and prostate cancer [57], and it demonstrated upregulation in Text24h according to its $Log2FC$ value. TGM2, also known as TG2, which is associated with drug resistance and metastasis in breast and pancreatic cancer cells [58], also expressed more in Text24h compared to Flat24h. TGFBI, which has been proved to be a tumor cell metastasis suppresses in vivo [59], slightly down-regulated in Text24h. MT1G was also found down-regulated in Text24h. Although there is no clear conclusion on whether MT1G is a tumor suppressor of BRCA or not, the study

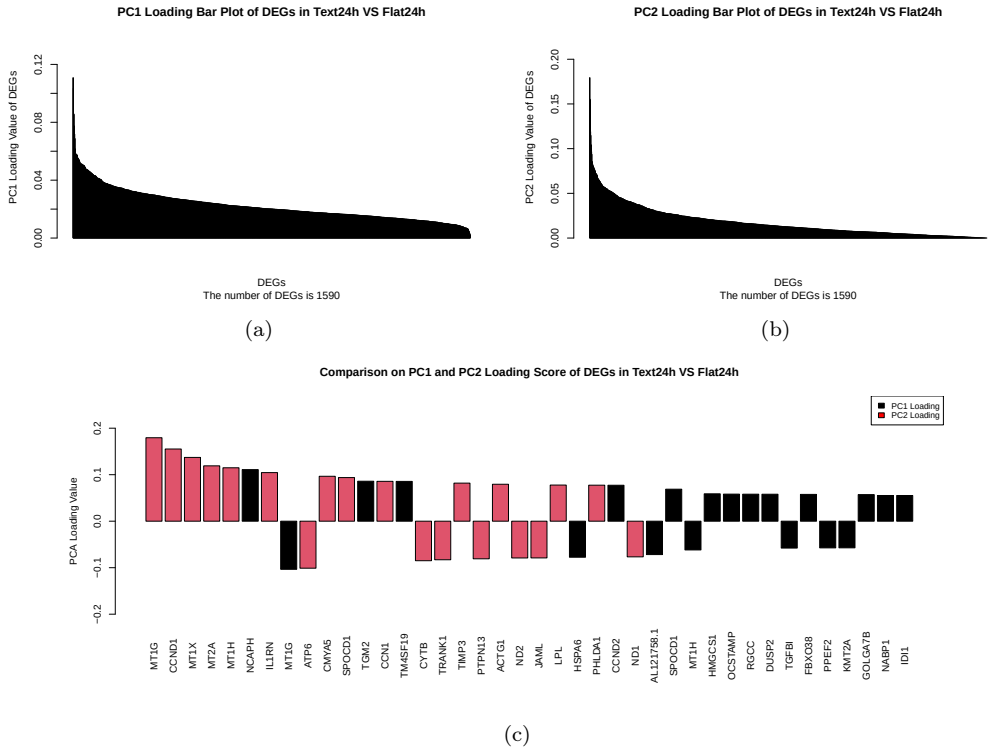
conducted by Rohit et al. demonstrated that MT1 clusters (including MT1G) downregulated in estrogen receptor (ER) BRCA cells. The comparison result of the gene expression levels of MT1 cluster in ERα+ and ERα− BRCA cells to those in one normal cell line, human mammary epithelial cells (HMEC), is illustrated by Figure 3.4.B [60]. As a result indicated, MT1G expressed less in ERα+ (MCF7, BT474) and ERα− (BT20) compared to the normal cell line (HMEC). In a nutshell, tumor suppressors like TGM2, TGFBI, and genes down-regulated in BRCA were found with lower expressions in the textured surface. Also, NCAPH that promotes BRCA is expressed more on the textured surface compared to the flat surface. Thus, macrophages from the Text24h phenotype are more likely to induce breast disease like BRCA.

According to the PCA plot of Text24h VS Flat24h, $PC2$ dimension mainly distinguished samples from Flat24h, i.e., genes that impacted $PC2$ most may differentially be expressed in Flat24h. As presented by Table 3.2, four of top five genes in the $PC2$ loading ranking list MT1G, MT1X, MT1H, and MT2A are from MT cluster. Metallothioneins (MT) are a family of metal-binding proteins that play an important role in cellular processes such as proliferation and apoptosis. As shown in Figure 3.5 [60], MT1G, MT1X, MT1H were downregulated in $ER\alpha+$ cell lines MCF7, BT474, and $ER\alpha-$ cell line BT20. Among them, MT1G had the strongest positive effect on $PC2$ loading among all the DEGs. Besides, four mitochondrially encoded (ME) genes MT-ATP6, CYTB, ND2, and ND1 negatively affected $PC2$ loading. Based on this information and the distribution of samples from Flat24h, the expression pattern of MT and those ME genes are possible factors that influence the clustering of samples from Flat24h. Since samples in Flat24h were from different donors; thus, these mentioned genes may express significantly differently from donor and donor.

In conclusion, based on what we found in the PC loading of DEGs, there is no strong correlation between the formation of fibrosis or inflammation and the DEGs in Text24h VS Flat24h. However, according to the results related to research about biomakers and tumor suppressors of BRCA, we could conclude that biomakers of BRCA were upregulated in Text24h, and tumor suppressors were downregulated in Text24h. Besides, samples from Flat24h were separated on $PC2$ dimension, and MT and ME clusters impact $PC2$ strongly. Thus, expression patterns of such genes in macrophages are possibly different from donor to donor.

Symbol	Gene Description	PC1 Loading	Log2FC
NCAPH	Non-SMC Condensin I Complex Subunit H	0,110888464	1,16712806
MT1G	Metallothionein 1G	0,110888464	-2,422086236
TGM2	Transglutaminase 2	0,085949662	0,845623487
TM4SF19	Transmembrane 4 L Six Family Member 19	0,085586598	0,976994448
HSPA6	Heat Shock Protein Family A (Hsp70) Member 6	-0,077463944	-2,839588759
CCND2	Cyclin D2	0,077306564	0,915886456
AL121758.1	-	-0,071808648	-4,998020564
SPOCD1	SPOC Domain Containing 1	0,068823353	0,994828878
MT1H	Metallothionein 1H	-0,061807311	-2,562766049
HMGCS1	3-Hydroxy-3-Methylglutaryl-CoA Synthase 1	0,058854596	0,603824082
OCSTAMP	Osteoclast Stimulatory Transmembrane Protein	0,05821451	0,951556342
RGCC	Regulator Of Cell Cycle	0,058065789	0,793787212
DUSP2	Dual Specificity Phosphatase 2	0,057995529	0,79991803
TGFBI	Transforming Growth Factor Beta Induced	-0,057799291	-0,76223664
FBXO38	F-Box Protein 38	0,057628089	0,73644157
PPEF2	Protein Phosphatase With EF-Hand Domain 2	-0,057391242	-2,261531755
KMT2A	Lysine Methyltransferase 2A	-0,057299924	-1,07267816
GOLGA7B	Golgin A7 Family Member B	0,057144104	0,901477412
NABP1	Nucleic Acid Binding Protein 1	0,055264135	0,557850557
IDI1	Isopentenyl-Diphosphate Delta Isomerase 1	0,055239723	0,671775803

Table 3.1: DEGs with Top 20 PC1 loading of Text24h VS Flat24h



(a) (b)

(c)

Figure 3.3: PC loading bar plot of Text24h VS Flat24h.

3.1.3 Text96h VS Flat96h

Visualizations on the DEGA result of Text96h VS Flat96h is shown in Figure 3.6. Even though these two phenotypes did not have intersections in $PC1$ direction, the PCA plot (Figure 3.6.b) cannot indicate a clear classification of samples from these two groups. Samples from Text96h are distributed much more densely on the $PC2$ dimension than samples from Flat96h, which is similar to what has been shown in the PCA plot of Text24h VS Flat24h(Figure 3.2(b)). However, the distance between Sample 007 and 019 (Flat96h) from other samples in Flat96h is larger than that between these samples and the Text96h group, and they were more likely to be classified into the Text96h group. This is consistent with what the PCA plot of all the samples (Figure 3.1) showed. This can also be found in the correlation heatmap (Figure 3.6(c)), i.e., Sample 007 and 019 were more correlated to samples in Text96h rather than those from Flat96h. Nevertheless, Figure 3.2.d shows that Sample 007 and 019 were still clustered together with samples from Flat96h rather than Text96h based on the expression heatmap on DEGs. The MA plot (Figure 3.6.a) indicates that the five most statistically significant DEGs are EBP, RPL37A, CYBB, MTRNR2L1, and HNRNPUL2.

Among them, MTRNR2L1 is also one of the five most significant DEGs of the comparison group Text24h VS Flat24h. MTRNR2L1 encodes human MT-RNR2-like 1 whose functions yet unknown. According to Expression Atlas, MTRNR2L1 showed higher expression level under the situation of comparing non-triple-negative BRCA samples to normal breast organoids samples [61]. The expression level of MTRNR2L1 of textured samples was significantly lower than that of flat sampels, however, since there are not many studies on the functions of this human MT-RNR2-like 1, following validation is needed to verify the effects of if the higher expression of human MT-RNR2-like 1 on flat samples can be related to the induction of non-triple-negative BRCA.

As demonstrated in Figure 3.7.a and .b, the number of DEGs in Text96h VS Flat96h is 3278, this is almost double of that of Text24h VS Flat24h. Similarly, there are only a small part of genes possess strong (positive or negative) effects on $PC1$ or $PC2$ loading. Illustration of the comparison

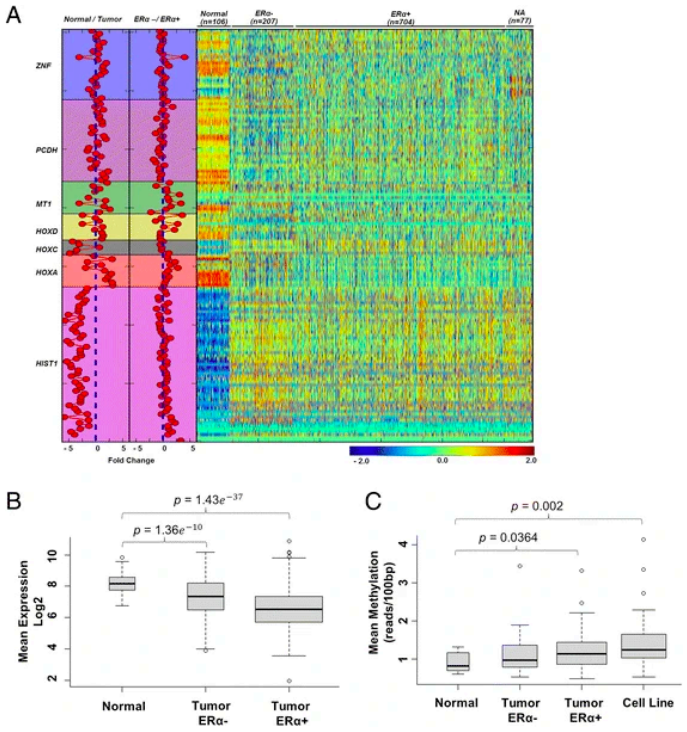


Figure 3.4: Lower expression and higher methylation of MT1 gene cluster for $ER\alpha+$ compared to $ER\alpha-$ and normal patient samples. [60]

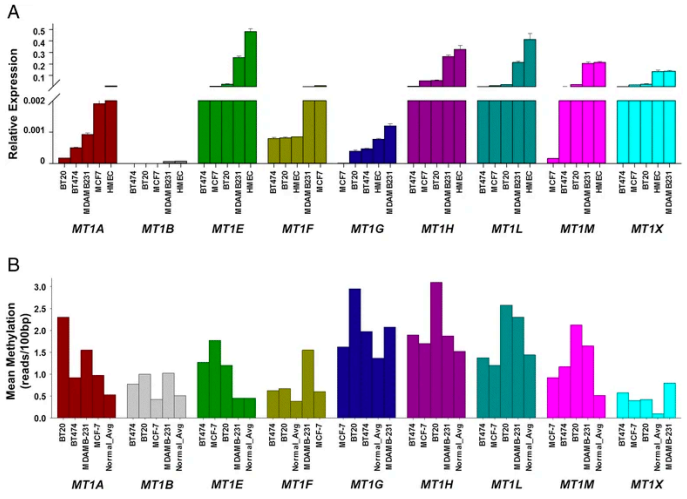


Figure 3.5: Basal level expression and DNA methylation for most of the genes in MT1 gene cluster show lower expression and higher methylation in breast cancer cell lines. [60]

between top 20 $PC1$ and $PC2$ loading shown in Figure 3.7(c) indicates that CRABP2 and ATP8A1 are the genes which encompass the largest contribution to $PC1$ and $PC2$, respectively. According to what has been shown in the PCA plot (Figure 3.6.b), genes contributed to $PC1$ most are possible factors that separate samples per phenotype, and their information can be viewed in Table 3.4. As indicated by the Log2FC value in Table 3.4, the expression level of possible inflammatory breast carcinoma biomarkers has been up-regulated in Text96h. CRABP2 has demonstrated its ability to promote invasion and metastasis of $ER-$ BRCA in vitro and in vivo [62], and it has

Symbol	Gene Description	PC2 Loading	Log2FC
MT1G	Metallothionein 1G	0,179481509	-2,422086236
CCND1	Cyclin D1	0,155249769	0,930278252
MT1X	Metallothionein 1X	0,137148999	-1,417218855
MT2A	Metallothionein 2A	0,119030774	-1,046174129
MT1H	Metallothionein 1H	0,114875073	-2,562766049
IL1RN	Interleukin 1 Receptor Antagonist	0,1045028	0,721016913
ATP6	Mitochondrially Encoded ATP Synthase Membrane Subunit 6	-0,101080942	-0,444650876
CMYA5	Cardiomyopathy Associated 5	0,096582389	-1,616909317
SPOCD1	SPOC Domain Containing 1	0,093804999	0,994828878
CCN1	-	0,085715492	-4,064713384
CYTB	Mitochondrially Encoded Cytochrome B	-0,084987039	-0,46955064
TRANK1	Tetratricopeptide Repeat And Ankyrin Repeat Containing 1	-0,082954389	-1,193816542
TIMP3	TIMP Metallopeptidase Inhibitor 3	0,081890324	0,669565508
PTPN13	Protein Tyrosine Phosphatase, Non-Receptor Type 13	-0,080884834	-1,295785926
ACTG1	Actin Gamma 1	0,07943415	0,570421155
ND2	Mitochondrially Encoded NADH: Ubiquinone Oxidoreductase Core Subunit 2	-0,079076738	-0,692647263
JAML	Junction Adhesion Molecule Like	-0,078973504	-0,700043119
LPL	Lipoprotein Lipase	0,077716839	0,704534318
PHLDA1	Pleckstrin Homology Like Domain Family A Member 1	0,077451331	0,654726961
ND1	-	-0,076691397	-0,451804169

Table 3.2: DEGs with Top 20 PC2 loading of Text24h VS Flat24h

been found up-regulated in Text96h. TGM2(TG2), which is associated with drug resistance and metastasis in breast and pancreatic cancer cells [58] also expressed more in Text96h compared to Flat96h. Besides, the increase of MT2A, a potential breast carcinogenesis [63], was also found in Text96h. RHOC is the only gene that is related to inflammation in breast tissue among all the top genes contribute to $PC1$ of Text96h VS Flat96h. The protein encoded by RHOC is thought to be important in cell locomotion, and the over-expression of it is associated with tumor cell proliferation and metastasis in breast disease and inflammatory breast carcinoma [64]. In summary, DEGs that contributed largely to the $PC1$ of Text96h VS Flat96h can be related to BRCA and its biomarkers including CRABP2, TGM2, MT2A have been found up-regulated in Text96h samples in our study. This may present that macrophages exposed to textured surface for 96h are more possible to cause the BRCA than those exposed to flat surface. This conclusion is also similar to that of the comparison group of Text24h VS Flat24h.

According to what has been shown in the PCA plot (Figure 3.6.b), genes contributed to $PC1$ most are possible factors that separate samples per phenotype, and their information can be viewed in Table 3.4. As indicated by the Log2FC value in Table 3.4, the expression level of possible inflammatory breast carcinoma biomarkers has been up-regulated in Text96h. CRABP2 has demonstrated its ability to promote invasion and metastasis of $ER-$ BRCA in vitro and in vivo [62], and it has been found up-regulated in Text96h. TGM2(TG2), which is associated with drug resistance and metastasis in breast and pancreatic cancer cells [58] also expressed more in Text96h compared to Flat96h. Besides, the increase of MT2A, a potential breast carcinogenesis [63], was also found in Text96h. RHOC is the only gene that is related to inflammation in breast tissue among all the top genes contribute to $PC1$ of Text96h VS Flat96h. The protein encoded by

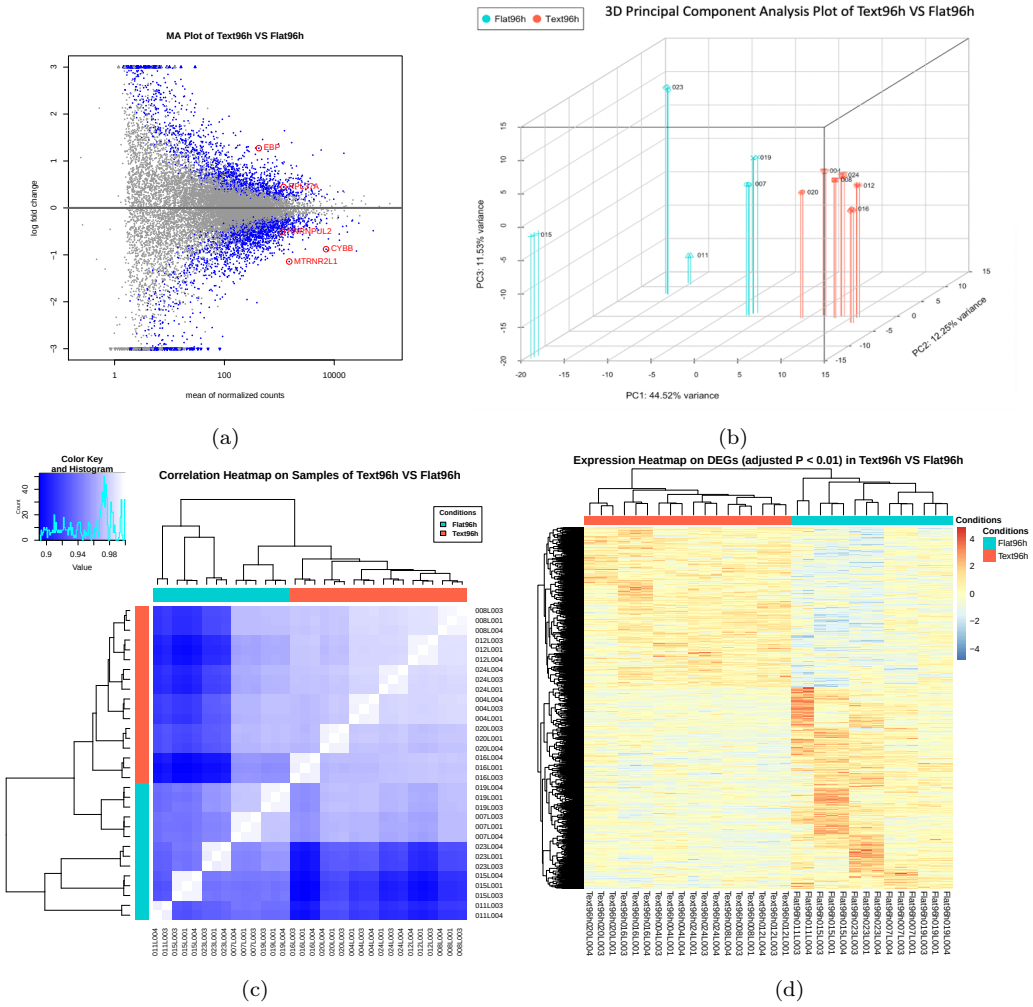
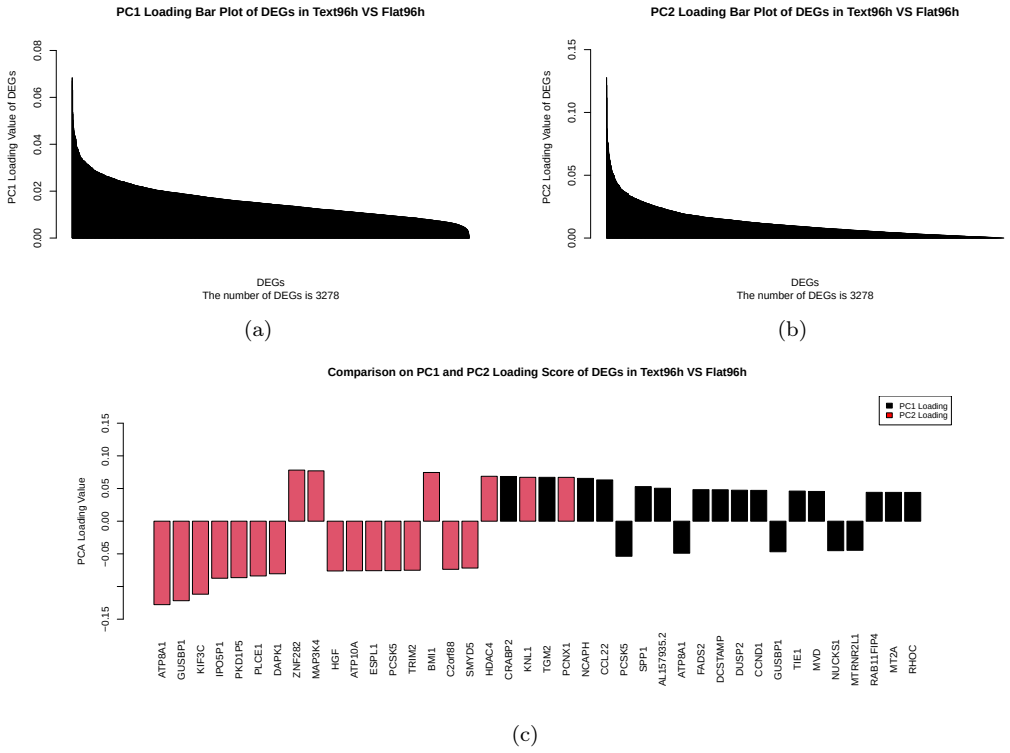


Figure 3.6: Visualization of comparison group Text96h VS Flat96h only with differentially expressed genes. (a) MA plot of Text96h VS Flat96h. (b) Principal Component Analysis plot of Text96h VS Flat96h. (c) Correlation Heatmap on Samples of Text96h VS Flat96h. (d) Expression Heatmap on DEGs (adjusted P $<$ 0.01) in Text96h VS Flat96h. (e) Bar plot of PC1 loading. (f) Bar plot of PC2 loading.

RHOC is thought to be important in cell locomotion, and the over-expression of it is associated with tumor cell proliferation and metastasis in breast disease and inflammatory breast carcinoma [64]. In summary, DEGs that contributed largely to the $PC1$ of Text96h VS Flat96h can be related to BRCA and its biomarkers including CRABP2, TGM2, MT2A have been found up-regulated in Text96h samples in our study. This suggests that macrophages exposed to textured surface for 96h are more possible to cause the BRCA than those exposed to flat surface. This conclusion is also similar to that of the comparison group of Text24h VS Flat24h.

Different from genes with higher $PC2$ loading in Text24h VS Flat24h, $PC2$ of Text96h VS Flat96h is not affected by genes from specific clusters. Nevertheless, some of these genes can be related to the regulation of inflammation response. For example, suppression of motor protein KIF3C, which was down-regulated in Text96h, has been proved to be able to inhibit growth and metastasis of tumors in BRCA by inhibiting TGF-β signaling [65], and this gene expressed less in Text96h. There was also the down-regulation of PLCE1 in Text96h. The down-regulated PLCE1 could contribute to the decrease of the expression of proinflammatory cytokines IL-6, TNF-α, and



(a) (b)

(c)

Figure 3.7: PC loading bar plot of Text96h VS Flat96h.

IL-1α, and increased the expression of IL-10, which is an anti-inflammatory cytokine [66]. Other genes in the top 20 are not reported to be associated with immune responses of macrophages based on literal research. Thus, genes mediate inflammatory or anti-inflammatory reactions that may play a role in distinguishing samples from Flat96h on $PC2$ direction.

3.1.4 Discussion

Based on visualizations and the information of genes strongly impacting $PC1$ and $PC2$, we could find some similarities between the DEGA results of Text24h VS Flat24h and Text96h VS Flat96h. First of all, these two comparison groups encompass a same significant DEG MTRNR2L1. In both of these two groups, MTRNR2L1 were down-regulated in textured phenotype. Moreover, changes in expression level of this gene in Text96h VS Flat96h (Log2FC = -1,15) is larger than that (Log2FC = -0,73) in Text24h VS Flat24h. Another similarity between them is that both of their $PC1$ are mainly affected by genes that are possible biomarkers or tumor suppressors of BRCA. The differences in the expression levels in MT and ME clusters may indicate that after 24 hours, macrophages exposured on different surfaces may shown variations in energy supply model. Table 3.5 presents a summary based on the comparison of PC loading of DEGs, based on it, DEGs like NCAPH, TGM2, TGFBI, MT1G that are related to regulating BRCA are essential for the classification of samples from flat and textured phenotypes. The difference between the Text24h VS Flat24h and Text96h VS Flat96h is presented by genes that strongly impact $PC2$ of these two group. As we mentioned before, $PC2$ of Text24h VS Flat24 hare possibly affected by MT and ME clusters. On the contrary, genes that had a large contribution to $PC2$ in Text96h VS Flat96h are not from any specific clusters; however, KIF3C, HDAC4, and PLCE1 are associated with inflammatory responses.

Symbol	Gene Description	PC1 loading	Log2FC
ATP8A1	ATPase Phospholipid Transporting 8A1	-0,127750782	-1,741503281
GUSBP1	Glucuronidase, Beta Pseudogene 1	-0,121714466	-3,078532688
KIF3C	Kinesin Family Member 3C	-0,111739272	-2,151929322
IPO5P1	Importin 5 Pseudogene 1	-0,087318483	-3,11292227
PKD1P5	Polycystin 1, Transient Receptor Potential Channel Interacting Pseudogene 5	-0,08641264	-1,993340424
PLCE1	Phospholipase C Epsilon 1	-0,083803214	-3,147320686
DAPK1	Death Associated Protein Kinase 1	-0,080470865	-1,00742748
ZNF282	Zinc Finger Protein 282	0,078202593	-1,580745688
MAP3K4	Mitogen-Activated Protein Kinase 4	0,077048686	-1,214426657
HGF	Hepatocyte Growth Factor	-0,076200458	-3,207145176
ATP10A	ATPase Phospholipid Transporting 10A (Putative)	-0,07593819	-6,358317014
ESPL1	Extra Spindle Pole Bodies Like 1, Separase	-0,075781213	-2,805595759
PCSK5	Proprotein Convertase Subtilisin/Kexin Type 5	-0,075681296	-2,026757518
TRIM2	Tripartite Motif Containing 2	-0,075068796	-1,919088801
BMI1	BMI1 Proto-Oncogene, Polycomb Ring Finger	0,07455622	-1,035406377
C2orf88	Chromosome 2 Open Reading Frame 88	-0,073651655	-2,987104446
SMYD5	SMYD Family Member 5	-0,071724938	-1,484769899
HDAC4	Histone Deacetylase 4	0,068734776	-1,620810463
KNL1	Kinetochore Scaffold 1	0,067160901	-2,911058807
PCNX1	Pecanex Homolog 1	0,067106165	-1,16623115

Table 3.3: DEGs with Top 20 PC2 loading of Text96h VS Flat96h

Symbol	Gene Description	PC2 Loading	Log2FC
CRABP2	Cellular Retinoic Acid Binding Protein 2	0,068437131	1,362473429
TGM2	Transglutaminase 2	0,067135436	1,393346679
NCAPH	Non-SMC Condensin I Complex Subunit H	0,065514839	1,647398032
CCL22	C-C Motif Chemokine Ligand 22	0,063141742	1,164960954
PCSK5	Proprotein Convertase Subtilisin/Kexin Type 5	-0,053657857	-2,026757518
SPP1	Secreted Phosphoprotein 1	0,052930195	1,164960954
AL157935.3	-	0,050421253	1,213666593
ATP8A1	ATPase Phospholipid Transporting 8A1	-0,049009993	-1,741503281
FADS2	Fatty Acid Desaturase 2	0,048232727	1,015614152
DCSTAMP	Dendrocyte Expressed Seven Transmembrane Protein	0,048080864	0,891859066
DUSP2	Dual Specificity Phosphatase 2	0,04742501	1,304282324
CCND1	Cyclin D1	0,047174342	1,274666605
GUSBP1	Glucuronidase, Beta Pseudogene 1	-0,046791975	-3,078532688
TIE1	Tyrosine Kinase With Immunoglobulin Like And EGF Like Domains 1	0,046085248	0,964346055
MVD	Mevalonate Diphosphate Decarboxylase	0,045529004	1,465601239
NUCKS1	Nuclear Casein Kinase And Cyclin Dependent Kinase Substrate 1	-0,045110925	-0,900069369
MTRNR2L1	MT-RNR2 Like 1	-0,044532268	-1,149479552
RAB11FIP4	RAB11 Family Interacting Protein 4	0,044154353	1,147086048
MT2A	Metallothionein 2A	0,044098003	1,147086048
RHOC	Ras Homolog Family Member C	0,044014079	0,957148274

Table 3.4: DEGs with Top 20 PC1 loading of Text96h VS Flat96h

	Text24h VS Flat24h	Text96h VS Flat96h
Classification per Phenotype: PC1	DEGs mediate BRCA: NCAPH, TGM2, TGFBI, MT1G	DEGs mediate BRCA: CRABP2, TGM2, MT2A, RHOC
Classification per sample within Flat Phenotype: PC2	Metallothioneins (MT1) cluster: MT1G, MT1X, MT1H Mitochondrial-encoded genes: ATP6, CYTB, ND2, ND1	DEGs mediate inflammation: KIF3C, PLCE1, HDAC4

Table 3.5: Summary of genes with large PC loading value.

3.2 Geneset Enrichment Analysis

There is no gold standard for how to rank genes for the GSEA. Thus, two different ways have been used in our study to generate the *.RNK* file. The first was to rank all the genes based on the product of direction (sign) of Log2FC and logarithm of *padj* for each gene as shown in Equation 3.1 [44]. Some DEGs may possess an extreme Log2FC value due to that they have overall very low expression in most samples but high expression in one or a few samples, however, they are not truly DEGs and possess large p-value/adjusted p-value. Nevertheless, by the p-value/adjusted p-value ranking method, genes with extremely low p-value or *padj* are recognized as differentially expressed at a low FDR. The other method was ranking all the genes according to their Log2FC. This ranking method aimed to provide a GSEA result based on Log2FC values of these genes and a comparison to the p-value ranking method. DEGs with the largest positive Log2FC values were at the top of the list. The reference dataset was also the **H** dataset.

$$r = sign(Log2FC) * (-1) * log_{10}(padj) \tag{3.1}$$

The Log2FC value was decided by comparing a gene's expression level on the textured surface phenotype to that on the flat surface phenotype. Positive genes at the top of the list possess high expression levels in class textured surface group. Genes at the top of the list are more highly expressed in Text class (i.e., Text24h or Text96h) of samples, while genes at the bottom are highly expressed in Flat class (i.e., Flat24h or Flat96h). In GSEA result, gene sets with positive enrichment score are those up-regulated in the textured group (i.e., enriched at the top of the ranking list), gene sets with negative enrichment score are those down-regulated in the textured group (i.e., enriched at the bottom of the ranking list). There were more than 10,000 genes in the list; and GSEA was executed with the reference of hallmark (**H**) dataset to get a more general and non-redundant summary of enriched biological processes based on all genes. Hallmark gene sets summarize and represent specific, well-defined biological states or processes and display coherent expression [45]. These gene sets were computed based on overlaps between gene sets in other MSigDB collections. In this part, H_pos and H_neg denote the result gained from hallmark collection. Pathways with an FDR value smaller than 0.25 and p-value smaller than 0.05 will be selected as significantly enriched gene sets.

3.2.1 Text24h VS Flat24h

The Log2FC value was decided by comparing a gene's expression level on the textured surface phenotype to that on the flat surface phenotype. Positive genes at the top of the list possess high expression levels in class textured surface group. Genes at the top of the list are more highly expressed in Text class (i.e., Text24h or Text96h) of samples, while genes at the bottom are highly expressed in Flat class (i.e., Flat24h or Flat96h). In GSEA result, gene sets with positive enrichment score are those up-regulated in the textured group (i.e., enriched at the top of the ranking list), gene sets with negative enrichment score are those down-regulated in the textured group (i.e., enriched at the bottom of the ranking list). There were more than 10,000 genes in the list; and GSEA was executed with the reference of hallmark (**H**) dataset to get a more general and non-redundant summary of enriched biological processes based on all genes. Hallmark gene sets summarize and represent specific, well-defined biological states or processes and display coherent expression [45]. These gene sets were computed based on overlaps between gene sets in other MSigDB collections. In this part, H_pos and H_neg denote the result gained from hallmark collection. Pathways with an FDR value smaller than 0.25 and p-value smaller than 0.05 will be selected as significantly enriched gene sets.

GSEA result generated by ranking *padj* value provides more clues than that by ranking Log2FC. As shown in Table 3.6, including cholesterol homeostasis, oxidative phosphorylation (OXPHOS), Myc and so forth were enriched in Text24h. Hallmarks related to IFN-γ and $-\alpha$; P53, KRAS, and TNF-α signals; and xenobiotic metabolism were enriched in Text24h phenotype. Based on previous studies on macrophages' metabolism, the energy supply pattern is important for distinguishing M1 from M2 cells. M1 cells use glycolysis for rapid killing, while M2 macrophages

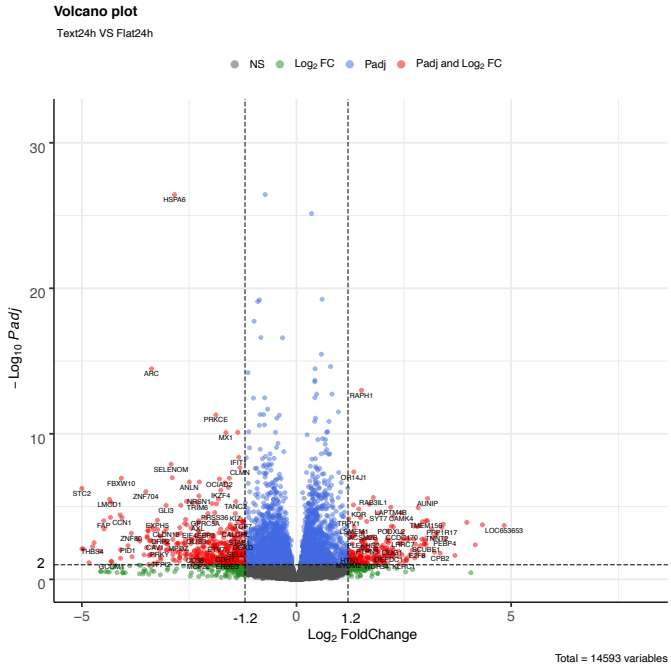


Figure 3.8: Volcano plot of Text24h VS Flat24h

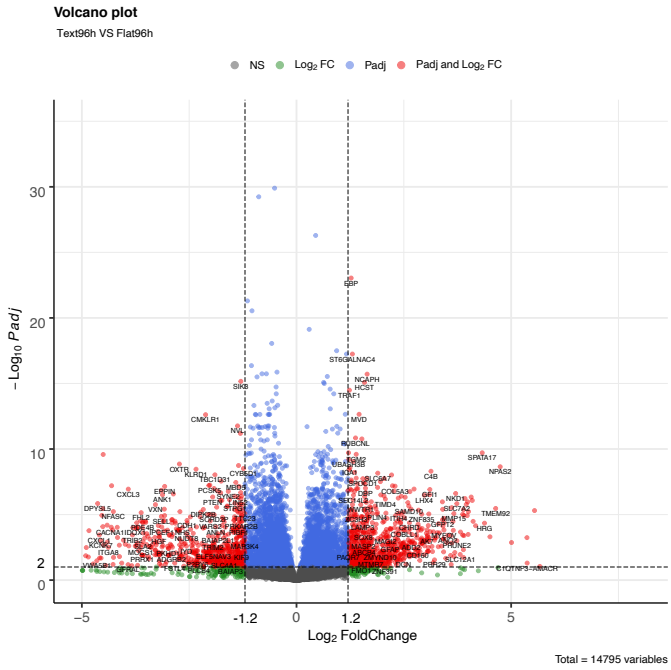


Figure 3.9: Volcano plot of Text96h VS Flat96h

rely on mitochondrial OXPHOS for continuously producing energy [67]. As for the enrichment of hallmarks and pathways related to inflammation modulation. The expression of IFN-γ and IFN-α, which are pro-inflammatory cytokines and identifiers of M1 macrophages, were enriched in flat phenotype. Pathways of P53, Myc, KRas, and TNFα were also found enriched in Flat24h. Among them, Myc and TNFα, whose deregulation will contribute to inflammation and immune

suppression [68]. These GSEA results are consistent with previous research on the expression level of pro-inflammatory cytokines and identifiers of M1 macrophages according to the studies of Thomas et al. and Maciej et al. [69, 70]. They reviewed the fibrosis formation mechanism of macrophages and the polarization of M1- and M2-like cells as shown in Figure 3.10 and 3.11. Thus, these hallmarks and pathways are the possible factors leading to the fibrosis formation on the textured surface compared to the flat surface. Besides, hallmark of heme metabolism was found enriched in Flat24h, which may indicate that macrophages from Flat24h are also polarized from M1 cells to M2 cells based on the fact that heme oxygenase-1 induction could drive the phenotypic shift to M2 macrophages [71].

As for the GSEA result based on ranking Log2FC value indicated in Table 3.7, there were similarities found with those shown in Table 3.6, including the enrichment in biological process or states related to the cholesterol homeostasis, IFN-α and -γ, P53 pathway, Kras and TNFα signaling, hypoxia. Even though these results demonstrated that macrophages from the Flat24h phenotype presented more characteristics of M1-like cells, there was also enrichment that showed M2 macrophages' features, such as angiogenesis [72] and Hedgehog signaling pathway [73] in samples from Flat24h. These results provided clues that both of M1-like cells and M2-like cells existed in the Flat24h phenotype, however, more macrophages were likely to be the M1-like cell type.

	GeneSet	Size	NES	p-value	FDR
H_pos	HALLMARK_CHOLESTEROL_HOMEOSTASIS	70	2.07	0.0	0.0
H_pos	HALLMARK_OXIDATIVE_PHOSPHORYLATION	199	1.85	0.0	0.002
H_pos	HALLMARK_ANDROGEN_RESPONSE	91	1.5	0.007	0.077
H_pos	HALLMARK_MTORC1_SIGNALING	198	1.415	0.002	0.11
H_pos	HALLMARK_MYC_TARGETS_V1	199	1.35	0.01	0.14
H_pos	HALLMARK_APICAL_JUNCTION	147	1.272	0.04	0.20
H_neg	HALLMARK_INTERFERON_GAMMA_RESPONSE	184	-1.946	0.0	0.0
H_neg	HALLMARK_P53_PATHWAY	185	-1.735	0.0	0.008
H_neg	HALLMARK_KRAS_SIGNALING_DN	89	-1.616	0.002	0.031
H_neg	HALLMARK_INTERFERON_ALPHA_RESPONSE	93	-1.576	0.005	0.036
H_neg	HALLMARK_TNFA_SIGNALING_VIA_NFKB	194	-1.540	0.0	0.042
H_neg	HALLMARK_E2F_TARGETS	197	-1.517	0.002	0.048
H_neg	HALLMARK_XENOBIOTIC_METABOLISM	155	-1.468	0.0	0.068
H_neg	HALLMARK_APOPTOSIS	151	-1.445	0.0098	0.073
H_neg	HALLMARK_UV_RESPONSE_UP	139	-1.424	0.016	0.08
H_neg	HALLMARK_HYPOXIA	171	-1.406	0.009	0.085
H_neg	HALLMARK_ADIPOGENESIS	182	-1.341	0.023	0.12
H_neg	HALLMARK_HEME_METABOLISM	169	-1.327	0.026	0.11

Table 3.6: GSEA result of Text24h VS Flat24h by ranking *padj*.

3.2.2 Text96h VS Flat96h

13827 genes were included in the *.RNK* file of Text96h VS Flat96h. The GSEA results of Text96h VS Flat96h generated by ranking *padj* are shown in Table 3.8, in which hallmarks related to Myc, IFN-α and -γ responses are enriched. Also, hallmarks G2/M checkpoint and mitotic spindle were found enriched in Flat96h. This may indicate a low proliferation of Flat96h macrophages.

3.2.3 Discussion

In summary, macrophages from Flat24h and Flat96h phenotype expressed more significant enrichment of pro-inflammatory factors, including IFN-α and -γ, TNF-α, which are the biomarkers of M1-like macrophages. On the contrary, there was no significant enrichment related pro-inflammatory processes in macrophages from Text24h or Text96h phenotype. The GSEA result

	GeneSet	Size	NES	p-value	FDR
H_pos	HALLMARK_CHOLESTEROL_HOMEOSTASIS	70	1.82	0.0	0.009
H_neg	HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	159	-1.76	0.0	0.024
H_neg	HALLMARK_INTERFERON_GAMMA_RESPONSE	184	-1.74	0.0	0.13
H_neg	HALLMARK_KRAS_SIGNALING_DN	93	-1.65	0.0	0.024
H_neg	HALLMARK_INTERFERON_ALPHA_RESPONSE	93	-1.54	0.005	0.069
H_neg	HALLMARK_ANGIOGENESIS	26	-1.51	0.043	0.075
H_neg	HALLMARK_UV_RESPONSE_DN	128	-1.50	0.005	0.070
H_neg	HALLMARK_UV_RESPONSE_UP	139	-1.42	0.021	0.130
H_neg	HALLMARK_HEDGEHOG_SIGNALING	24	-1.41	0.071	0.129
H_neg	HALLMARK_INFLAMMATORY_RESPONSE	171	-1.40	0.017	0.123
H_neg	HALLMARK_TNFA_SIGNALING_VIA_NFKB	194	-1.40	0.016	0.117
H_neg	HALLMARK_HYPOXIA	171	-1.39	0.020	0.111
H_neg	HALLMARK_APOPTOSIS	151	-1.38	0.026	0.109
H_neg	HALLMARK_P53_PATHWAY	184	-1.31	0.042	0.202

Table 3.7: GSEA result of Text24h VS Flat24h by ranking Log2FC.

	GeneSet	Size	NES	p-value	FDR
H_neg	HALLMARK_G2M_CHECKPOINT	194	-1.61	0.001	0.054
H_neg	HALLMARK_E2F_TARGETS	195	-1.54	0.001	0.072
H_neg	HALLMARK_MITOTIC_SPINDLE	192	-1.5	0.0	0.093
H_neg	HALLMARK_MYC_TARGETS_V1	200	-1.47	0.003	0.086
H_neg	HALLMARK_INTERFERON_GAMMA_RESPONSE	178	-1.42	0.006	0.109
H_neg	HALLMARK_INTERFERON_ALPHA_RESPONSE	92	-1.37	0.032	0.15

Table 3.8: GSEA result of Text96h VS Flat96h by ranking adjusted p value.

	GeneSet	Size	NES	p-value	FDR
H_pos	HALLMARK_CHOLESTEROL_HOMEOSTASIS	70	1.82	0.002	0.005
H_pos	HALLMARK_APICAL_SURFACE	28	1.46	0.044	0.157

Table 3.9: GSEA result of Text96h VS Flat96h by ranking log2FC.

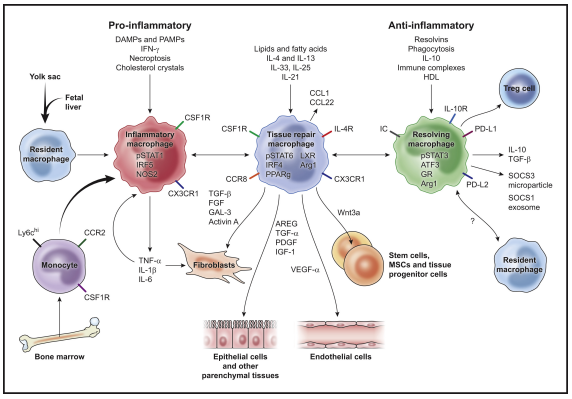


Figure 3.10: Macrophages in tissue repair, regeneration, and fibrosis [69].

shows that the metabolism model of macrophages cultured on a textured surface is oxidative metabolism that is the metabolism model of M2-like macrophages. Thus, in our case, macrophages samples from flat breast SMIs and textured SMIs represented characteristics of M1 and M2 macrophages, respectively. M2 type is the fibrotic macrophages that lead to wound healing and tissue repair [70]. The flat surface caused higher expressions of pro-inflammatory cytokines, this may

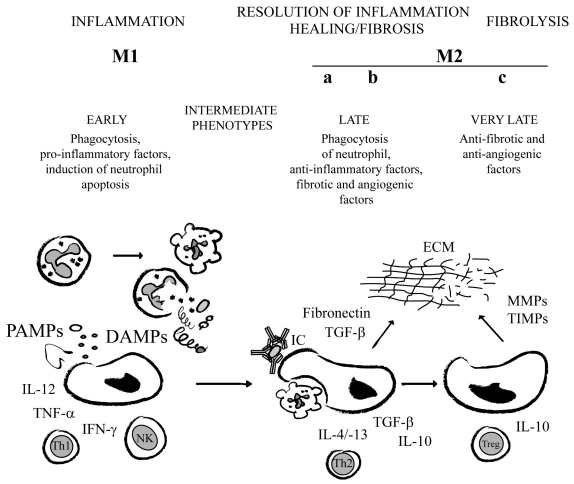


Figure 3.11: Macrophages and fibrosis [70].

be the reason why flat breast SIMs more frequently caused fibrous capsule formation, which is consistent with the conclusion that a smooth implant surface is more likely to induce capsular contracture [74]. Moreover, M1- and M2-like macrophages are anti-tumor and pro-tumor cells, respectively, explaining why genes that possess larger PC loading value are tumor modulators.

Based on the results of GSEA on 24h and 96h, we could conclude that at the early stage (24 hours) of exposing macrophages to flat and textured surfaces, gene sets are related to macrophages' metabolism, which indicates their polarization is enriched. More gene sets of pathways and cytokines that modulate the immune and inflammation responses inside macrophages were found at 24 hours enriched than at the late stage (96 hours). In summary, pathways including Myc, NFkB, and cytokines INF-γ and -α may critical in the different expression levels of inflammatory-related genes between flat phenotype and textured phenotype. The differences in gene expression level may cause various M1/M2 polarization models and lead to the different proportions between M1- and M2-like macrophages, which is the possible factor that affects the distinct extent of capsular contracture on the flat and textured surface.

3.3 Gene Network

3.3.1 Transcription Factor Extended Gene Network

As the GSEA results showed that macrophages exposed to the flat and the textured surfaces represented different M1/M2 polarization, to better investigate how DEGs cooperated and lead to various polarization models, the gene networks are composed of most up-regulated or down-regulated DEGs were created. Transcription factors (TFs) were also extended to the gene network based on literature research. In the gene networks, edges represent protein interactions; grey circular nodes indicate genes; blue rectangles are genes retrieved from TFe, meaning the transcriptional control. Genes related to M1 and M2 macrophages are circled by green and red, respectively. In the TF extended gene network, the genes up-regulated or gene markers of M1 and M2 macrophages are circled by green and red, respectively.

Text24h VS Flat24h

As for Text24h VS Flat24h, DEGs with an absolute Log2FC value ≥ 2.9 were chosen to build the proper network shown in Figure 3.12. Biomarkers of M1 cells, including IL6; M2 cells biomarkers such as IGFs (IGFBP3, IGF2BP1), MMPs (MMP1, MMP3, MMP14), TGFB2 can be found from this network. According to previous studies, the expression level of IL6 has been proved highly up-regulated in M1 macrophages compared to M2 cells. Thus, the following research should be focusing on the protein level to check its transcription activities under the flat and the textured surfaces. Besides, CAV1 is also a centroid gene that connected different parts in this gene network. However, there were few studies about how it acts on the gene profile and behavior of macrophages. Since it is directly connected with MMP14, MMP1, and IGFBP3, it may directly modulate the expression of genes that are up-regulated in M2 cells and affect the M1/M2 polarization on different topography.

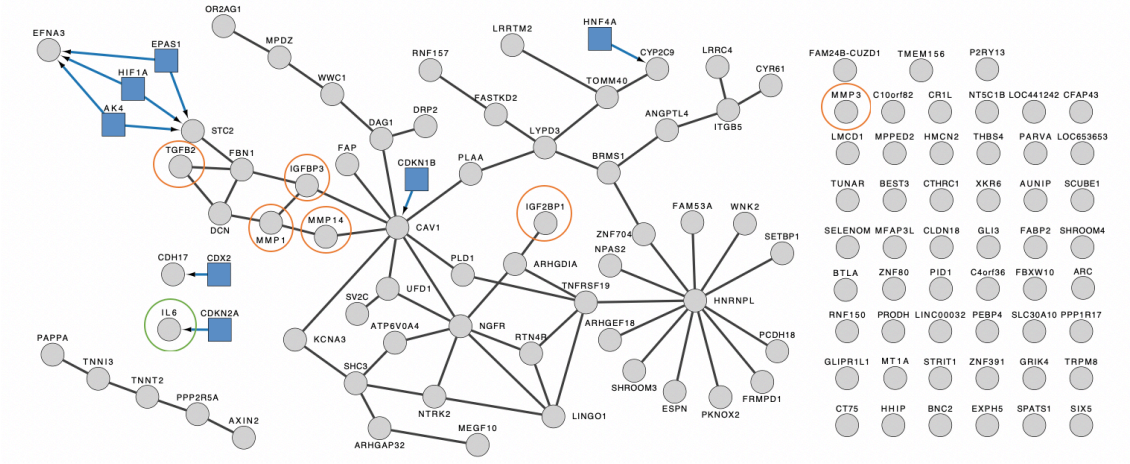


Figure 3.12: Transcription factor extended gene of Text24h VS Flat24h.

Text96h VS Flat96h

In Text96h VS Flat96h, DEGs with an absolute Log2FC value ≥ 3.5 were chosen to build the gene network illustrated by Figure 3.13. Based on literature research, the biomarkers of M1 cells including CXCL5, CCL11, and STAT1; M2 cells biomarkers such as IGFs (IGF1, IGFBP1), MMP15, FGF7, chemokine ligands (CXCL1 and CXCL3), and chemokine receptors (CXCR1, CXCR2), CD226 can be found from this network. However, the expression of STAT1, CXCLs, and CXCRs are environment-dependent; therefore, only based on this gene network, we could not

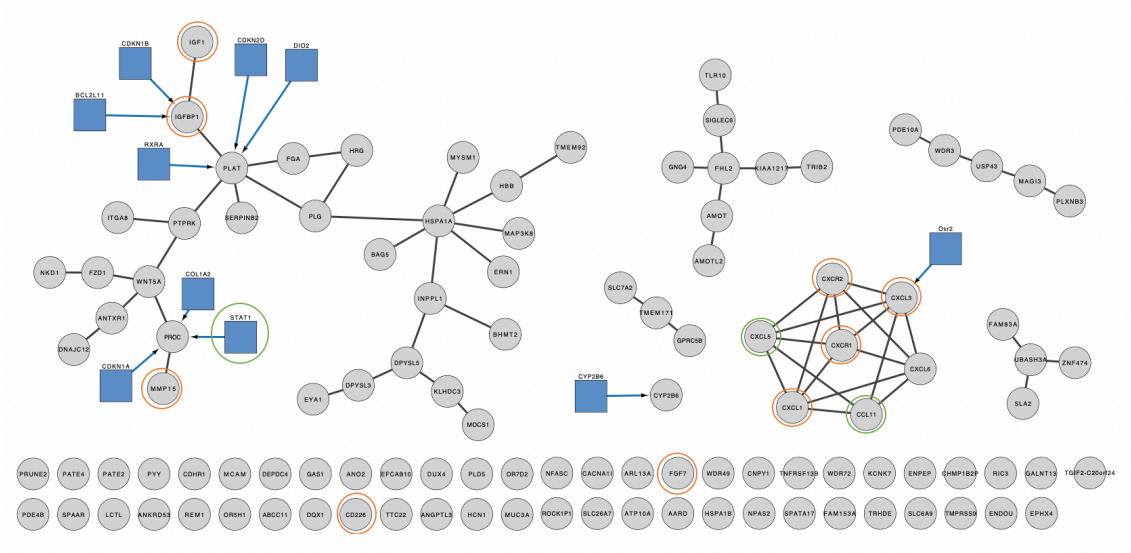


Figure 3.13: Transcription factor extended gene network of Text96h VS Flat96h.

predict how they will affect macrophages in polarization or cell cycle. Further investigation on their functions on macrophages cultured on different surfaces is needed to verify their impacts. Besides, we noticed WNT5A connected in this network, it is also interesting to research how this Wnt family gene react in macrophages.

3.4 CMap Query

To assess the gene landscape of different donors reacting to the flat and textured implant surface and which connectivity is between biological processes and expression profiles of macrophages from each donor exposed to flat and textured surface, DEGA was also performed each donor separately. DEGs from each donor were ranked by their Log2FC value and filtered by different cutoffs. DEGs at the top and bottom of the ranking list was used for CMap query. Before selection, DEGs were pre-filtered by BING space to ensure efficient input of the query. Perturbants with an absolute connectivity score (CS) between the donor-specified DEGs and CMap reference DEGs larger than 80 were selected as highly-related.

3.4.1 Text24h VS Flat24h

We observed between 25 to 305 DEGs with an absolute Log2FC ≥ 4.55 as compare textured surface to flat surface, as shown in Figure 3.14 among all the donors. In Text24h VS Flat24h, 38 DEGs, which were differentially expressed in all the donors, were selected by the threshold of 2.8. The threshold was adjusted to 2.8 since the number of up-regulated DEGs, which possessed an absolute Log2FC larger than 2.9, did not meet the lowest requirement (10) of CMap query input. According to the Heatmap analysis module provided by CMap query, single compounds Anisomycin and NSC-632839 were with the largest negative median CS (-84.81 and -85.40, respectively), the CS values of were listed in Table 3.10. Both of these two compounds are protein synthesis inhibitor, but they target different proteins and pathways.

Anisomycin is the inhibitor of the synthesis of ribosomal proteins (RPL10L, RPL11, RPL13A, RPL15, RPL19, RPL23, RPL23A, RPL26L1, RPL3, RPL37, RPL8, RSL24D1) and small nuclear ribonucleoprotein (SNU13) [75]. Anisomycin could also p38/JNK MAPK pathway [76]. The p38MAPK pathways are vital for regulating pro-inflammatory cytokines biosynthesis at the transcriptional level [77]; and they are strongly activated in vivo by environmental stresses and inflammatory cytokines [78]. Blockage of the p38MAPK pathway could obstacle the production of pro-inflammatory cytokines like TNF-α and IL-1 [79]. Also, the JNK pathway is considered to be a potential target for the therapy of inflammatory diseases. JNK regulates the synthesis of pro-inflammatory cytokines such as IL-2, IL-6, and TNF-α, which are biomarkers of M1-like macrophages [80]. It can promote transcript blockages involved in the fibrotic responses [81]. The connectivity between the gene landscape of macrophages and the p38/JNK MAPK pathways may demonstrate that the exposure to flat and textured surfaces lead to the variations in inflammation-related pathways like p38/JNK MAPKs. This result also reflects that the differences between macrophages exposed to the flat and textured surfaces mainly represent inflammation responses. More important, the study of Hao et al. demonstrated that Anisomycin could up-regulate the expression of fibrotic proteins, including E-ts1 [82], Pai-1 [83] and CTGF [84]. Based on their study, there were probably lower fibrosis-related protein expressions after macrophages were exposed to the textured rather than the flat surface after 24 hours.

IKK-2-inhibitor-V is the IKK inhibitor, NF_kB pathway inhibitor, and the protein kinase inhibitor of IKBKB [75]. IKK are necessary for rapidly activating NF_kB by proinflammatory signaling which are triggered by TNF_α or lipopolysaccharide (LPS). The negative CS between IKK-2-inhibitor-V and the macrophages genome suggests the up-regulation of IKK, IKBKB, or NF_kB in Text24h compared to Flat24h. Macrophages could be polarized to an immunosuppressive M2 phenotype by interleukin (IL)-1R and MyD88, which required IKBKB-mediated NF_kB activation [85]. However, NF_kB's functions on the regulation of inflammation and tumour are still complicated and environment-dependent. The increase in the expression of NF_kB can promote not only the inflammation but also the tumour [86, 87]. Thus, only the connectivity between the genome and the compounds could not provide enough clues to predict the variances between macrophages cultured on different implant surfaces.

NSC-632839 is a ubiquitin-specific proteases (USPs) inhibitor, which targets on SENP2, USP1, USP2, USP7 [88]. Among these target genes, SENP2 could inhibit the transcriptional activity of the Wnt/β-catenin pathway that is a major regulator of human fibrosis development and progres-

sion across organs [89]. Since unresolved inflammation frequently occurs during the transition from normal wound healing to chronic fibrosis, and interfering Wnt signaling could attenuate several experimental fibrosis models in vivo, Wnt signaling is a promising (pre)clinical therapeutic targets to suppress the formation of fibrosis [90]. There are few studies about USPs' functions; however, the inhibition of USPs is a potential and novel anticancer therapeutic strategy [91]. In summary, NSC-632839 may promote the transcription of Wnt pathway, resulting in more fibrosis; and the negative connectivity between the gene landscape of Text24h VS Flat24h and NSC-632839 may indicate that the textured surface may more anti-fibrosis compared to the flat.

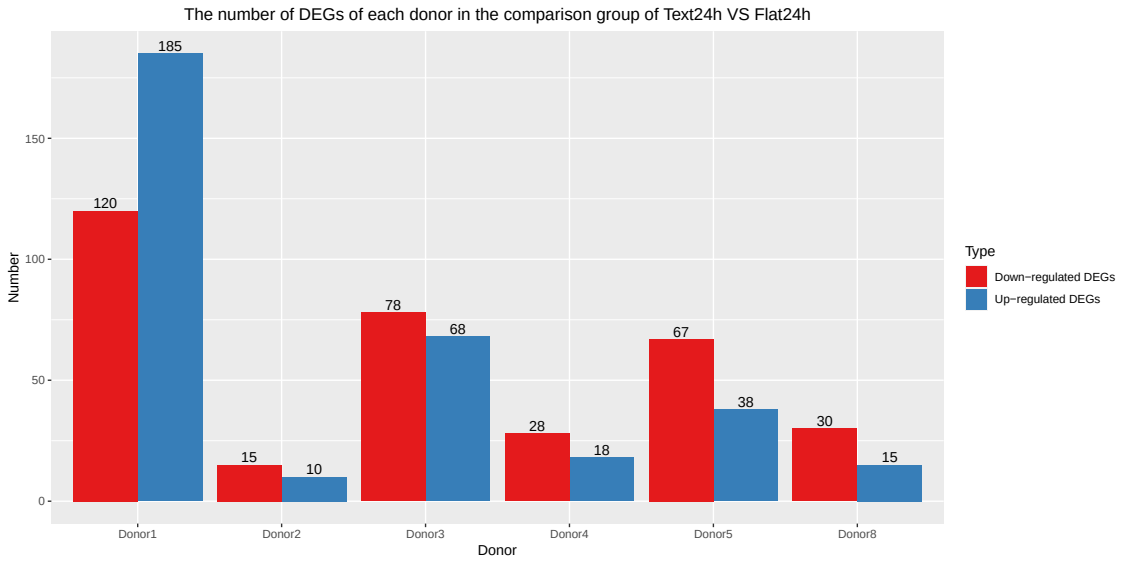


Figure 3.14: The number of DEGs of Text24h VS Flat24h.

Compound	Donor 1	Donor 2	Donor 3	Donor 4	Donor 5	Donor 8	All Donors	Median Tau Score
Anisomycin	-89.15	-23.64	-13.80	-78.06	0	-84.72	-90.81	-78.06
IKK-2-inhibitor-V	-78.30	1.73	-1.90	-77.44	-10.39	-85.16	-76.30	-76.30
NSC-632839	-1.25	-19.83	-87.70	-89.90	0	-75.51	-94.49	-75.51

Table 3.10: CMap query result of Text24h VS Flat24h.

3.4.2 Text96h VS Flat96h

We observed between 36 to 228 DEGs with an absolute Log2FC ≥ 5.5 as compare textured surface to flat surface as shown in Figure 3.15 among all the donors. In the comparison group of Text96h VS Flat96h, 39 DEGs which were differentially expressed in all the donors were selected by the threshold of 3.5. According to the Heatmap analysis module provided by CMap query, single compounds emetine and homoharringtonine were with the largest negative median CS (-83.21 and -89.63, respectively), the CS values of were listed in Table 3.11.

Both of these two compounds can suppress BRCA cells. Emetine is the protein synthesis inhibitor of RPS2. In the study conducted by Yun et al., they investigated the effect of RPS2 by using RAW 264.7 murine macrophage cells [92]. The results showed that RPS-2 RPS2 could activate RAW 264.7 in different pathways, including NF_kB signal and MAPKs pathway that can regulate inflammation and the formation of tumors. Sun et al. showed that emetine treatment could antagonize Wnt/β-catenin signaling, induce apoptosis, and suppress the migration, invasion, and sphere formation of BRCA cells [93]. Homoharringtonine is an apoptosis stimulant and a protein synthesis inhibitor targeting on RPL3. The combination of RPL3 and 5-FU has been

demonstrated to be a promising strategy for chemotherapy of lung cancers lacking functional p53 that are resistant to 5-FU [94]. Homoharringtonine is also an approved anti-leukemia drug that can suppress triple negative BRCA growth by rapidly reducing anti-apoptotic protein abundance [95]. In a conclusion, the negative connectivity between the Emetine and Homoharringtonine represents that the textured are more likely to induce BRCA in macrophages after a 96-hour exposure in vitro.

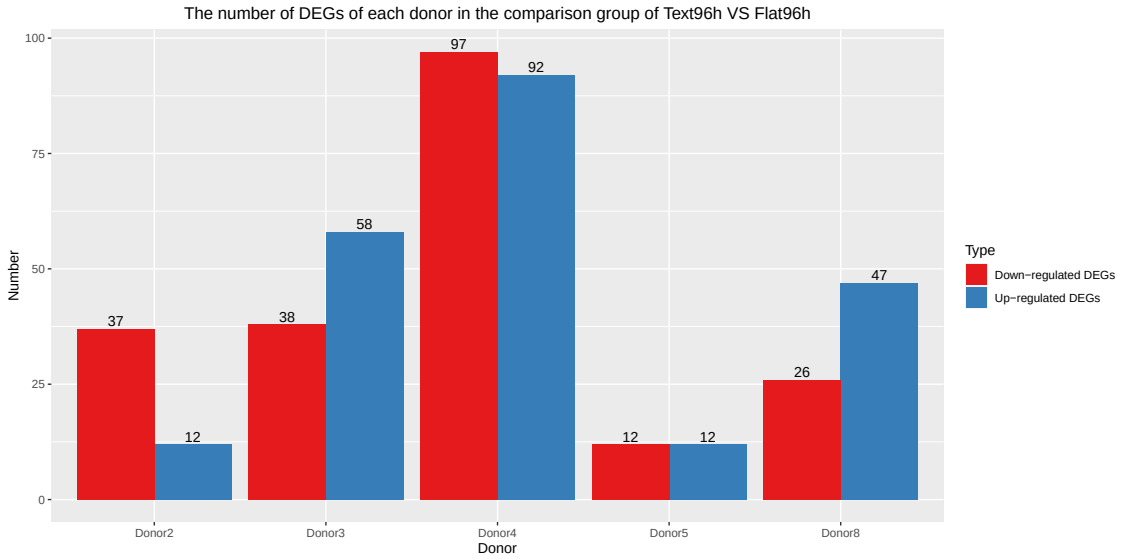


Figure 3.15: The number of DEGs of Text96h VS Flat96h.

Compound	Donor 2	Donor 3	Donor 4	Donor 5	Donor 8	All Donors	Median Tau Score
Emetine	-92.66	-67.89	-73.75	-93.79	0	-92.74	-83.21
Homoharringtonine	-91.40	-89.99	-89.28	-86.64	0	-95.52	-89.63

Table 3.11: CMap query result of Text96h VS Flat96h.

3.4.3 Discussion

Based on the CMap prediction result, after exposure to different surfaces for 24 hours, the expression profile of macrophages cultured on the textured surface presented variations in P38/JNK and Wnt pathways compared to that of flat-surface phenotype. Besides, negative CS may reflect that the textured surface is more anti-fibrosis. Differently, after 96 hours, the distinction between these two phenotypes was mainly reflected in the induction of BRCA, showing that macrophages under the textured condition could promote tumor. In summary, these differences can be related to the characteristics of M1/M2 macrophages; thus, the CMap query result also indicates that different breast implant surfaces may result in different M1/M2 polarization, fibrosis formation, and BRCA induction models.

Chapter 4

Discussion and Further Work

4.1 Discussion

Our study provides a complete DEGA pipeline, from the quality control step of RNA-seq dataset to the generation of the gene network from most up-/down-regulated DEGs.

In this project, a new insight into the DEGA of macrophages exposed to the breast SMIs with various surface structures is presented. Based on the analysis of two comparison groups Text24h VS Flat24h and Text96h VS Flat96h, we could conclude that DEGs can distinguish macrophages cultured on the flat-surface implant from those cultured on the textured-surface implant. In addition, genes strongly impacted PC1 of both Text24h VS Flat24h and Text96h VS Flat96h are mainly possible biomarkers and tumor suppressors of breast cancer (BRCA). More critical, biomarkers were upregulated in Text24h, and the tumor suppressors were downregulated. This indicates that the textured SMI may be more likely to induce breast diseases like BRCA and suggests that macrophages cultured on the textured surface presented the characteristics of M2-like macrophages. However, there is no evidence that genes that possessed large PC loading values can be associated with inflammation responses. Besides, MT and ME genes are possible factors that affect the classification of samples within Flat24h, which means that these genes are differentially expressed from donor to donor. As for Text96h VS Flat96h, even though there is not a clear cluster between samples, the expression pattern of DEGs indicated that groups of genes did differentially expressed in Text96h VS Flat96h, and their expression levels can separate samples from these two groups.

The GSEA result proved that macrophages cultured on flat SMIs represented enriched pathways of pro-inflammatory cytokines, including IFN-α and -γ, and TNF-α on the flat breast SMI surface. These pro-inflammatory factors are the bio-markers of M1-like macrophages and can form capsular contracture and chronic wounding. On the contrary, more characteristics of M2-like macrophages, such as oxidative phosphorylation, can be found from samples cultured on the textured surfaces. Based on this information and literature research, M1/M2 macrophages polarization may be the key causing different macrophages' immune responses when they are cultured on the flat and textured surface. The control of the quantity of M1- and M2-like cells may be clinical access to suppressing the fibrous capsule.

CMap query result provides possible perturbagens that are relatively highly-related to the DEG landscape of each donor and landscape of genes, which are differentially expressed in all the donors. The gene expression pattern of macrophages cultured for 24 hours is opposite to that of the patterns treated by anisomycin and NSC-63283. These two compounds can impact p38/JNK MAPK and Wnt pathways that are important to regulate the expression of pro-inflammatory cytokines. Differently, the gene expression pattern of macrophages cultured for 96 hours showed high connectivity with patterns perturbed by two single compounds, emetine, and homoharringtonine that can suppress BRCA. Based on these findings, we can infer that at the early stage (24 hours), the significant distinction between the flat-surface macrophages and textured-surface

macrophages is the expression level of pro-inflammatory factors. In contrast, at the late stage, the main difference is the induction of tumors.

In conclusion, our study proved that the macrophages exposed to the flat (Allergan Smooth surface) and the textured (Mentor Siltex surface) could present different expression profiles, this result is consistent with the study of Giuseppe et al. [11]. Besides, GSEA and CMap results also give a further assumption on the formation of fibrosis on breast implants, macrophages cultured on flat surfaces are more likely to develop to M1 macrophages cells, this maybe the main cell factor that cause the fibrosis and capsular contracture (CC). Compared to macrophages cultured on the flat surface, those on the textured surface possessed more characteristics of M2 macrophages. An important difference between M1 and M2 cells is that M2 cells are more tumor-associated. This may be the reason why most of genes that possess higher PC1 loading were genes related to tumor.

4.2 Further Work

Diabetic mice are often used for studies on wound macrophages and delayed wound healing since they share several characteristics with human chronic wounds. The research investigating wound macrophages show that their function is not properly regulated in diabetic mice compared to wild type ones, which was caused by a prolonged M1 macrophage presence, leading to inefficient transition to the M2 phenotype [96]. Based on this finding and results of our study, shorten M1 macrophage presence at the early stage is a promising way to reduce the occurrence of fibrosis or CC when transplanting macrophages to breast SMIs in vitro. Compounds like anisomycin and NSC-63283 indicated by CMap query result is possible to enhance the transformation efficiency between M1 and M2 macrophages. Thus, for the future work, the first priority would be verifying effects of these two mentioned compounds in vitro to investigate the M1/M2 proportion and the occurrence rate of fibrosis and CC.

Another important part of future work is the validation and the exploration of the gene network of macrophages shown in Figure 3.12 and 3.13. Even though M1 and M2 macrophages' biomarkers can be found in these gene networks, how interactions between DEGs affect immune responses of macrophages are yet evident. Validating such interactions can be completed by assessing the quantity of a gene's expression after regulating others connected to that gene.

Based on the CMap query result, no all the donors exhibited reactions to compounds like anisomycin and NSC-63283; besides, CC is not found in all the patients or costumers accepting breast augmentation surgery based on the previous investigation. Thus, it is also essential to compare the transcriptome and DEG landscape of each donor to detect which characteristics are possessed by the immune system that can adjust M1/M2 proportion and avoid CC. Combining techniques like machine learning may provide a possible pre-clinical way to predict which patient or client will suffer from the complication caused by breast implants.

As mentioned, there is no gold standard on how to rank all the genes or part of genes that are recognized as DEGs to do GSEA. Thus, further studies can also focus on whether a ranking method is more accurate than the other. This can be done by the following cell engineering experiment and protein quantity assessment of those cytokines or pathways from the enrichment analysis result of our study.

Bibliography

[1] Marlene Johnson. Breast implants: history, safety, and imaging. *Radiologic technology*, 84(5):439M–520M, 2013. 1

[2] Litong Ji, Tie Wang, Lining Tian, Hongjiang Song, and Meizhuo Gao. Roxatidine inhibits fibrosis by inhibiting nf-κb and mapk signaling in macrophages sensing breast implant surface materials. *Molecular medicine reports*, 21(1):161–172, 2020. 1

[3] Hannah Headon, Adbul Kasem, and Kefah Mokbel. Capsular contracture after breast augmentation: an update for clinical practice. *Archives of plastic surgery*, 42(5):532, 2015. 1

[4] Maria Elsa Meza Britez, Carmelo Caballero LLano, and Alcides Chaux. Periprosthetic breast capsules and immunophenotypes of inflammatory cells. *European journal of plastic surgery*, 35(9):647–651, 2012. 1

[5] Honghua Hu, Anita Jacombs, Karen Vickery, Steven L Merten, David G Pennington, and Anand K Deva. Chronic biofilm infection in breast implants is associated with an increased t-cell lymphocytic infiltrate: implications for breast implant–associated lymphoma. *Plastic and reconstructive surgery*, 135(2):319–329, 2015. 1

[6] Devin B Lowe and Walter J Storkus. Chronic inflammation and immunologic-based constraints in malignant disease. *Immunotherapy*, 3(10):1265–1274, 2011. 1

[7] S Barr, E Hill, and A Bayat. Current implant surface technology: an examination of their nanostructure and their influence on fibroblast alignment and biocompatibility. *Eplasty*, 9, 2009. 1, 2

[8] David Franklyn Williams. *The Williams dictionary of biomaterials*. Liverpool University Press, 1999. 1

[9] S Barr, EW Hill, and A Bayat. Functional biocompatibility testing of silicone breast implants and a novel classification system based on surface roughness. *Journal of the mechanical behavior of biomedical materials*, 75:75–81, 2017. 1, 2

[10] Georg Wick, Cecilia Grundtman, Christina Mayerl, Thomas-Florian Wimpissinger, Johann Feichtinger, Bettina Zelger, Roswitha Sgonc, and Dolores Wolfram. The immunology of fibrosis. *Annual review of immunology*, 31:107–135, 2013. 1

[11] Giuseppe Cappellano, Christian Ploner, Susanne Lobenwein, Sieghart Sopper, Paul Hoertnagl, Christina Mayerl, Nikolaus Wick, Gerhard Pierer, Georg Wick, and Dolores Wolfram. Immunophenotypic characterization of human t cells after in vitro exposure to different silicone breast implant surfaces. *PloS one*, 13(2), 2018. 1, 4, 45

[12] Giulia Daneshgaran, Daniel Gardner, Annie Chen, David Perrault, Maxwell B Johnson, Solmaz Niknam-Bienia, Vinaya Soundararajan, Alexander Fedenko, Regina Y Baker, and Alex K Wong. Differential gene expression in capsules derived from smooth and textured silicone implants. *Plastic and Reconstructive Surgery–Global Open*, 7(8S-1):35–36, 2019. 2

[13] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009. 4

[14] Ugrappa Nagalakshmi, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. The transcriptional landscape of the yeast genome defined by rna sequencing. *Science*, 320(5881):1344–1349, 2008. 4

[15] Clarissa M Koch, Stephen F Chiu, Mahzad Akbarpour, Ankit Bharat, Karen M Ridge, Elizabeth T Bartom, and Deborah R Winter. A beginner's guide to analysis of rna sequencing data. *American journal of respiratory cell and molecular biology*, 59(2):145–157, 2018. 5

[16] Vijender Chaitankar, Gökhan Karakülah, Rinki Ratnapriya, Felipe O Giuste, Matthew J Brooks, and Anand Swaroop. Next generation sequencing technology and genomewide data analysis: Perspectives for retinal research. *Progress in retinal and eye research*, 55:1–31, 2016. 5

[17] Simon Andrews. Fastqc: a quality control tool for high throughput sequence data. "http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/", 2010. [Online; accessed 02-April-2020]. 6, 7

[18] bioinformatics.babraham. Index of/projects/fastqc/help/3 analysis modules. "https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help", 2020. [Online; accessed 2-April-2020]. 7, 10

[19] Philip Ewels, Måns Magnusson, Sverker Lundin, and Max Käller. Multiqc: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048, 2016. 7

[20] Felix Krueger. Trim galore! "http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/", 2015. [Online; accessed 01-April-2020]. 8

[21] NCBI. Standard nucleotide blast. "https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome", 2020. [Online; accessed 16-July-2020]. 8

[22] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):10–12, 2011. 8

[23] Broad Institute. Flow cell images. "http://www.broadinstitute.org/files/shared/illuminavids/sequencingSlides.pdf", 2020. [Online; accessed 03-April-2020]. 9, 10

[24] Bushnell B. Bbmap filterbytile.sh. "sourceforge.net/projects/bbmap/", 2020. [Online; accessed 03-April-2020]. 10

[25] Wikipedia contributors. Awk — Wikipedia, the free encyclopedia, 2020. [Online; accessed 17-April-2020]. 11

[26] Claire R Williams, Alyssa Baccarella, Jay Z Parrish, and Charles C Kim. Empirical assessment of analysis workflows for differential expression analysis of human samples using rna-seq. *BMC bioinformatics*, 18(1):38, 2017. 12

[27] Juliana Costa-Silva, Douglas Domingues, and Fabricio Martins Lopes. Rna-seq differential expression analysis: An extended review and a software tool. *PloS one*, 12(12):e0190152, 2017. 12

[28] Charlotte Soneson, Michael I Love, and Mark D Robinson. Differential analyses for rna-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, 4, 2015. 12, 14

[29] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic rna-seq quantification. *Nature biotechnology*, 34(5):525–527, 2016. 12, 13

[30] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550, 2014. 12, 14

[31] Daehwan Kim, Ben Langmead, and Steven L Salzberg. Hisat: a fast spliced aligner with low memory requirements. *Nature methods*, 12(4):357–360, 2015. 13

[32] Genome reference consortium. "https://www.ncbi.nlm.nih.gov/grc", 2020. [Online; accessed 03-April-2020]. 14

[33] Ensembl. Grch38. "https://www.ensembl.org/Homo_sapiens/Info/Annotation", 2020. [Online; accessed 03-April-2020]. 14

[34] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Nature Precedings*, pages 1–1, 2010. 14

[35] Wikipedia contributors. Ma plot — Wikipedia, the free encyclopedia, 2020. [Online; accessed 14-March-2020]. 15

[36] Wikipedia contributors. Deseq2: Ma plot. "https://www.rdocumentation.org/packages/DESeq2/versions/1.12.3/topics/plotMA", 2020. [Online; accessed 25-September-2020]. 15

[37] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016. 15

[38] Steven M Holland. Principal components analysis (pca). *Department of Geology, University of Georgia, Athens, GA*, pages 30602–2501, 2008. 16

[39] Wikipedia contributors. Pearson correlation coefficient — Wikipedia, the free encyclopedia. "https://en.wikipedia.org/w/index.php?title=Pearson_correlation_coefficient&oldid=944456717", 2020. [Online; accessed 13-March-2020]. 16

[40] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005. 17

[41] Gsea. "https://www.gsea-msigdb.org/gsea/index.jsp", 2020. [Online; accessed 03-April-2020]. 17, 19, 21

[42] CLUE CONNECTOPEDIA. What is the connectivity map? "https://clue.io/connectopedia/cmap_overview", 2020. [Online; accessed 03-April-2020]. 17, 20

[43] Genepattern. Genepattern deseq2 module. "https://cloud.genepattern.org/gp/pages/index.jsf?lsid=urn:lsid:broad.mit.edu:cancer.software.genepattern.module.analysis:00362:1", 2020. [Online; accessed 03-April-2020]. 17

[44] Jüri Reimand, Ruth Isserlin, Veronique Voisin, Mike Kucera, Christian Tannus-Lopes, Asha Rostamianfar, Lina Wadi, Mona Meyer, Jeff Wong, Changjiang Xu, et al. Pathway enrichment analysis and visualization of omics data using g: Profiler, gsea, cytoscape and enrichmentmap. *Nature protocols*, 14(2):482–517, 2019. 18, 19, 34

[45] GSEA. Hallmark. "https://www.gsea-msigdb.org/gsea/msigdb/genesets.jsp?collection=H", 2020. [Online; accessed 16-July-2020]. 18, 34

[46] CLUE CONNECTOPEDIA. What is the l1000 assay? "https://clue.io/connectopedia/what_is_l1000", 2020. [Online; accessed 03-April-2020]. 20

[47] Melissa J Peart, Gordon K Smyth, Ryan K van Laar, David D Bowtell, Victoria M Richon, Paul A Marks, Andrew J Holloway, and Ricky W Johnstone. Identification and functional significance of genes regulated by structurally different histone deacetylase inhibitors. *Proceedings of the National Academy of Sciences*, 102(10):3697–3702, 2005. 20

[48] Davis J McCarthy and Gordon K Smyth. Testing significance relative to a fold-change threshold is a treat. *Bioinformatics*, 25(6):765–771, 2009. 20

[49] Afshin Raouf, Yun Zhao, Karen To, John Stingl, Allen Delaney, Mary Barbara, Norman Iscove, Steven Jones, Steven McKinney, Joanne Emerman, et al. Transcriptome analysis of the normal human mammary cell commitment and differentiation process. *Cell stem cell*, 3(1):109–118, 2008. 20

[50] CLUE. Cmap algorithm. "https://clue.io/connectopedia/cmap_algorithms", 2020. [Online; accessed 16-July-2020]. 21

[51] Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452, 2017. 21

[52] ConsensusPathDB. Consensuspathdb. "http://cpdb.molgen.mpg.de/", 2020. [Online; accessed 15-June-2020]. 22

[53] Atanas Kamburov, Christoph Wierling, Hans Lehrach, and Ralf Herwig. Consensuspathdb—a database for integrating human functional interaction networks. *Nucleic acids research*, 37(suppl_1):D623–D628, 2009. 22

[54] Atanas Kamburov, Ulrich Stelzl, Hans Lehrach, and Ralf Herwig. The consensuspathdb interaction database: 2013 update. *Nucleic acids research*, 41(D1):D793–D800, 2013. 22

[55] Dimas Yusuf, Stefanie L Butland, Magdalena I Swanson, Eugene Bolotin, Amy Ticoll, Warren A Cheung, Xiao Yu Cindy Zhang, Christopher TD Dickman, Debra L Fulton, Jonathan S Lim, et al. The transcription factor encyclopedia. *Genome biology*, 13(3):1–25, 2012. 22

[56] Paolo Fagone, Michelino Di Rosa, Maria Palumbo, Corinne De Gregorio, Ferdinando Nicoletti, and Lucia Malaguarnera. Modulation of heat shock proteins during macrophage differentiation. *Inflammation Research*, 61(10):1131–1139, 2012. 24

[57] Feilun Cui, Jianpeng Hu, Zhipeng Xu, Jian Tan, and Huaming Tang. Overexpression of ncaph is upregulated and predicts a poor prognosis in prostate cancer. *Oncology letters*, 17(6):5768–5776, 2019. 25

[58] Norikatsu Miyoshi, Hideshi Ishii, Koshi Mimori, Fumiaki Tanaka, Toshiki Hitora, Mitsuyoshi Tei, Mitsugu Sekimoto, Yuichiro Doki, and Masaki Mori. Tgm2 is a novel marker for prognosis and therapeutic target in colorectal cancer. *Annals of surgical oncology*, 17(4):967–972, 2010. 25, 29

[59] Gengyun Wen, Michael A Partridge, Bingyan Li, Mei Hong, Wupeng Liao, Simon K Cheng, Yongliang Zhao, Gloria M Calaf, Tian Liu, Jun Zhou, et al. Tgfbi expression reduces in vitro and in vivo metastatic potential of lung and breast tumor cells. *Cancer letters*, 308(1):23–32, 2011. 25

[60] Rohit R Jadhav, Zhenqing Ye, Rui-Lan Huang, Joseph Liu, Pei-Yin Hsu, Yi-Wen Huang, Leticia B Rangel, Hung-Cheng Lai, Juan Carlos Roa, Nameer B Kirma, et al. Genome-wide dna methylation analysis reveals estrogen-mediated epigenetic repression of metallothionein-1 gene cluster in breast cancer. *Clinical epigenetics*, 7(1):13, 2015. 26, 28

[61] Expression Atlas. Expression atlas: Rna-seq of 17 breast tumor samples of three different subtypes and normal human breast organoids samples. "https://www.ebi.ac.uk/gxa/experiments/E-GEOD-52194/Results?geneQuery=ENSG00000256618", 2020. [Online; accessed 25-September-2020]. 27

[62] Xuefei Feng, Miao Zhang, Bo Wang, Can Zhou, Yudong Mu, Juan Li, Xiaoxu Liu, Yaochun Wang, Zhangjun Song, and Peijun Liu. Crabp2 regulates invasion and metastasis of breast cancer through hippo pathway dependent on er status. *Journal of Experimental & Clinical Cancer Research*, 38(1):1–18, 2019. 28, 29

[63] Anna Krześlak, Ewa Forma, Paweł Jóźwiak, Agnieszka Szymczyk, Beata Smolarz, Hanna Romanowicz-Makowska, Waldemar Różański, and Magdalena Bryś. Metallothionein 2a genetic polymorphisms and risk of ductal breast cancer. *Clinical and experimental medicine*, 14(1):107–113, 2014. 29

[64] GeneCards. RHOC ready-to-use reference sequences and annotations. "https://www.genecards.org/cgi-bin/carddisp.pl?gene=RHOC&keywords=RHOC", 2020. [Online; accessed 03-April-2020]. 29, 30

[65] Chengqin Wang, Chenggang Wang, Zhimin Wei, Yujun Li, Wenhong Wang, Xia Li, Jing Zhao, Xuan Zhou, Xun Qu, and Fenggang Xiang. Suppression of motor protein kif3c expression inhibits tumor growth and metastasis in breast cancer by inhibiting tgf-β signaling. *Cancer letters*, 368(1):105–114, 2015. 30

[66] WenHua Li, Yong Li, Ying Chu, WeiMin Wu, QiuHua Yu, XiaoBo Zhu, and Qiang Wang. Plce1 promotes myocardial ischemia–reperfusion injury in h/r h9c2 cells and i/r rats by promoting inflammation. *Bioscience Reports*, 39(7), 2019. 31

[67] Jan Van den Bossche, Jeroen Baardman, Natasja A Otto, Saskia van der Velden, Annette E Neele, Susan M van den Berg, Rosario Luque-Martin, Hung-Jen Chen, Marieke CS Boshuizen, Mohamed Ahmed, et al. Mitochondrial dysfunction prevents repolarization of inflammatory macrophages. *Cell reports*, 17(3):684–696, 2016. 35

[68] Roderik M Kortlever, Nicole M Sodir, Catherine H Wilson, Deborah L Burkhart, Luca Pellegrinet, Lamorna Brown Swigart, Trevor D Littlewood, and Gerard I Evan. Myc cooperates with ras by programming inflammation and immune suppression. *Cell*, 171(6):1301–1315, 2017. 36

[69] Thomas A Wynn and Kevin M Vannella. Macrophages in tissue repair, regeneration, and fibrosis. *Immunity*, 44(3):450–462, 2016. 36, 37

[70] Maciej Lech and Hans-Joachim Anders. Macrophages and fibrosis: How resident and infiltrating mononuclear phagocytes orchestrate all phases of tissue injury and repair. *Biochimica et biophysica acta (BBA)-molecular basis of disease*, 1832(7):989–997, 2013. 36, 37, 38

[71] Yuji Naito, Tomohisa Takagi, and Yasuki Higashimura. Heme oxygenase-1 and anti-inflammatory m2 macrophages. *Archives of biochemistry and biophysics*, 564:83–88, 2014. 36

[72] Nadine Jetten, Sanne Verbruggen, Marion J Gijbels, Mark J Post, Menno PJ De Winther, and Marjo MPC Donners. Anti-inflammatory m2, but not pro-inflammatory m1 macrophages promote angiogenesis in vivo. *Angiogenesis*, 17(1):109–118, 2014. 36

[73] Amy J Petty, Ang Li, Xinyi Wang, Rui Dai, Benjamin Heyman, David Hsu, Xiaopei Huang, and Yiping Yang. Hedgehog signaling promotes tumor-associated macrophage polarization to suppress intratumoral cd8+ t cell recruitment. *The Journal of clinical investigation*, 129(12), 2019. 36

[74] Chin-Ho Wong, Miny Samuel, Bien-Keem Tan, and Colin Song. Capsular contracture in subglandular breast augmentation with textured versus smooth breast implants: a systematic review. *Plastic and reconstructive surgery*, 118(5):1224–1236, 2006. 38

[75] clue.io. Touchstone. "https://clue.io/touchstone", 2020. [Online; accessed 25-September-2020]. 41

[76] Wei Xiong, Ljubomir Z Kojic, Lanjing Zhang, Shiv S Prasad, Robert Douglas, Yutian Wang, and Max S Cynader. Anisomycin activates p38 map kinase to induce ltd in mouse primary visual cortex. *Brain research*, 1085(1):68–76, 2006. 41

[77] Jeremy Saklatvala. The p38 map kinase pathway as a therapeutic target in inflammatory disease. *Current opinion in pharmacology*, 4(4):372–377, 2004. 41

[78] Ana Cuenda and Simon Rousseau. p38 map-kinases pathway regulation, function and role in human diseases. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1773(8):1358–1375, 2007. 41

[79] Yan Li, Wei Zhang, Jianxin Gao, Jiaqi Liu, Hongtao Wang, Jun Li, Xuekang Yang, Ting He, Hao Guan, Zhao Zheng, et al. Adipose tissue-derived stem cells suppress hypertrophic scar fibrosis via the p38/mapk signaling pathway. *Stem cell research & therapy*, 7(1):1–16, 2016. 41

[80] Praveen K Roy, Farzana Rashid, Jack Bragg, and Jamal A Ibdah. Role of the jnk signal transduction pathway in inflammatory bowel disease. *World journal of gastroenterology: WJG*, 14(2):200, 2008. 41

[81] Keren Grynberg, Frank Y Ma, and David J Nikolic-Paterson. The jnk signaling pathway in renal fibrosis. *Frontiers in Physiology*, 8:829, 2017. 41

[82] Guanghua Hao, Zhenhua Han, Zhe Meng, Jin Wei, Dengfeng Gao, Hong Zhang, and Nanping Wang. Ets-1 upregulation mediates angiotensin ii-related cardiac fibrosis. *International journal of clinical and experimental pathology*, 8(9):10216, 2015. 41

[83] Rui-Ming Liu. Oxidative stress, plasminogen activator inhibitor 1, and lung fibrosis. *Antioxidants & redox signaling*, 10(2):303–320, 2008. 41

[84] Toshifumi Mori, Shigeru Kawara, Mikio Shinozaki, Nobukazu Hayashi, Takashi Kakinuma, Atsuyuki Igarashi, Masaharu Takigawa, Toru Nakanishi, and Kazuhiko Takehara. Role and interaction of connective tissue growth factor with transforming growth factor-β in persistent fibrosis: A mouse fibrosis model. *Journal of cellular physiology*, 181(1):153–159, 1999. 41

[85] Thorsten Hagemann, Toby Lawrence, Iain McNeish, Kellie A Charles, Hagen Kulbe, Richard G Thompson, Stephen C Robinson, and Frances R Balkwill. "re-educating" tumor-associated macrophages by targeting nf-κb. *The Journal of experimental medicine*, 205(6):1261–1268, 2008. 41

[86] Michael G Dorrington and Iain DC Fraser. Nf-κb signaling in macrophages: dynamics, crosstalk, and signal integration. *Frontiers in immunology*, 10:705, 2019. 41

[87] Mi Hee Park and Jin Tae Hong. Roles of nf-κb in cancer and inflammatory diseases and their therapeutic approaches. *Cells*, 5(2):15, 2016. 41

[88] Ming-Jer Young, Kai-Cheng Hsu, Tony Eight Lin, Wen-Chang Chang, and Jan-Jong Hung. The role of ubiquitin-specific peptidases in cancer progression. *Journal of biomedical science*, 26(1):42, 2019. 41

[89] Qing-Feng Jiang, Yu-Wei Tian, Quan Shen, Huan-Zhou Xue, and Ke Li. Senp2 regulated the stability of β-catenin through wwox in hepatocellular carcinoma cell. *Tumor Biology*, 35(10):9677–9682, 2014. 42

[90] Olivier Burgy and Melanie Königshoff. The wnt signaling pathways in wound healing and fibrosis. *Matrix Biology*, 68:67–80, 2018. 42

[91] Tao Yuan, Fangjie Yan, Meidan Ying, Ji Cao, Qiaojun He, Hong Zhu, and Bo Yang. Inhibition of ubiquitin-specific proteases as a novel anticancer therapeutic strategy. *Frontiers in pharmacology*, 9:1080, 2018. 42

[92] Yun Chen, Ruigang Zhou, Lixing He, Fengyang Wang, Xin Yang, Ling Teng, Chengheng Li, Suya Liao, Yongjian Zhu, Yuhui Yang, et al. Okra polysaccharide-2 plays a vital role on the activation of raw264. 7 cells by tlr2/4-mediated signal transduction pathways. *International immunopharmacology*, 86:106708, 2020. 42

[93] Qi Sun, Qiuxia Fu, Shiyue Li, Junjun Li, Shanshan Liu, Zhongyuan Wang, Zijie Su, Jiaxing Song, and Desheng Lu. Emetine exhibits anticancer activity in breast cancer cells as an antagonist of wnt/β-catenin signaling. *Oncology reports*, 42(5):1735–1744, 2019. 42

[94] Annapina Russo, Assunta Saide, Roberta Cagliani, Monica Cantile, Gerardo Botti, and Giulia Russo. rpl3 promotes the apoptosis of p53 mutated lung cancer cells by down-regulating cbs and nfκb upon 5-fu treatment. *Scientific reports*, 6(1):1–13, 2016. 43

[95] Mohamad Yakhni, Arnaud Briat, Abderrahim El Guerrab, Ludivine Furtado, Fabrice Kwiatkowski, Elisabeth Miot-Noirault, Florent Cachin, Frederique Penault-Llorca, and Nina Radosevic-Robin. Homoharringtonine, an approved anti-leukemia drug, suppresses triple negative breast cancer growth through a rapid reduction of anti-apoptotic protein abundance. *American Journal of Cancer Research*, 9(5):1043, 2019. 43

[96] Paulina Krzyszczyk, Rene Schloss, Andre Palmer, and François Berthiaume. The role of macrophages in acute and chronic wound healing and interventions to promote pro-wound healing phenotypes. *Frontiers in physiology*, 9:419, 2018. 45

Appendix A

Abbreviations and acronyms

Abbreviation	Definition
BRCA	Breast Cancer
CC	Capsular Contracture
CMap	Connectivity Map
DEG	Differentially Expresses Gene
DEGA	Differentially Expresses Gene Analysis
DET	Differential Expression Tool
FDR	False Discovery Rate
GEPH	Gene Expression Pattern Heatmap
GSEA	Gene Set Enrichment Analysis
H	Hallmark
KS	Kolmogorov–Smirnov
Log2FC	Log2 Fold Change
MT	Metallothionein
ME	Mitochondrially Encoded
NES	Normalized enrichment score
OXPHOS	Oxidative Phosphorylation
padj	Adjusted P-value
PC	Principal Component
PC1	Principal Component 1
PC2	Principal Component 2
QC	Quality Control
SMI	Silicone Mammary Implant
TF	Transcription Factor

Table A.1: Abbreviations table in alphabetical order.

Appendix B

MultiQC status check on RNA-seq data processed by FilterByTile twice

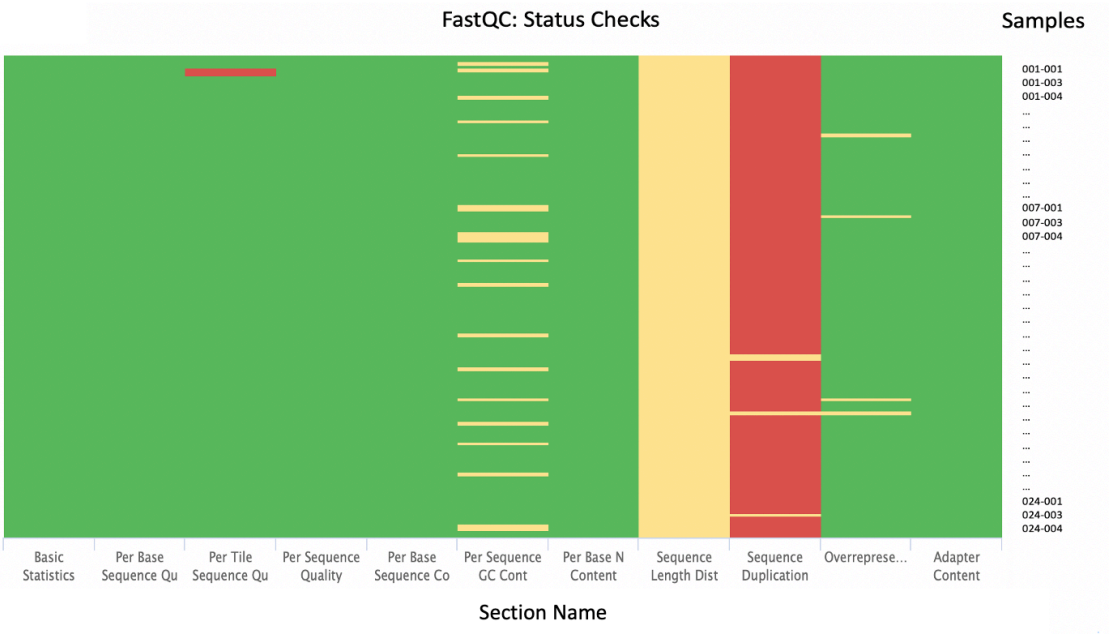


Figure B.1: MultiQC status check report on RNA-seq processed by FilterByTile twice.

Appendix C

Number of reads in the RNA-seq

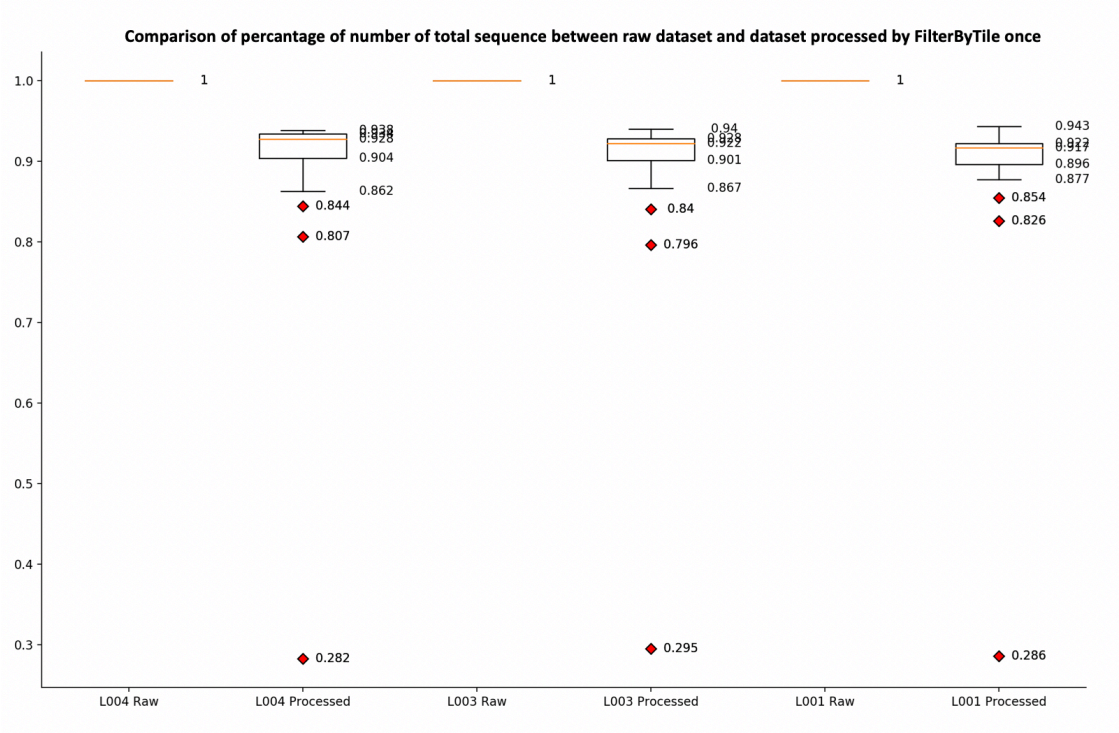


Figure C.1: Comparison of percentage of number of total sequence between raw dataset and dataset processed by FilterByTile once.

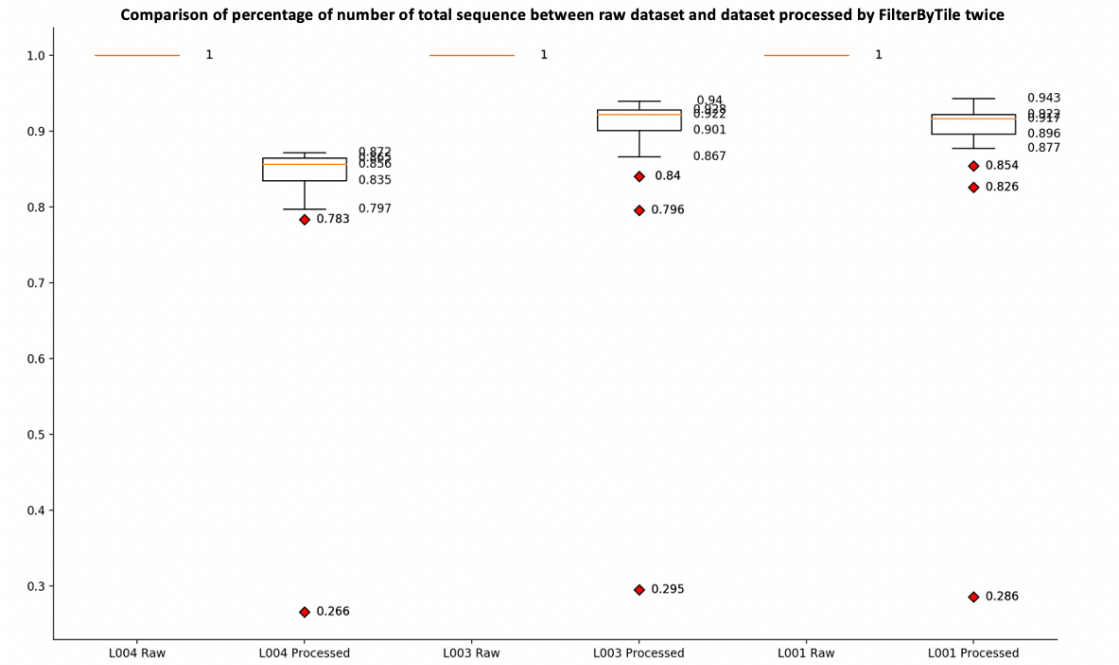


Figure C.2: Comparison of percentage of number of total sequence between raw dataset and dataset processed by FilterByTile twice.

Appendix D

Scree Plot of Variances of Principal Components

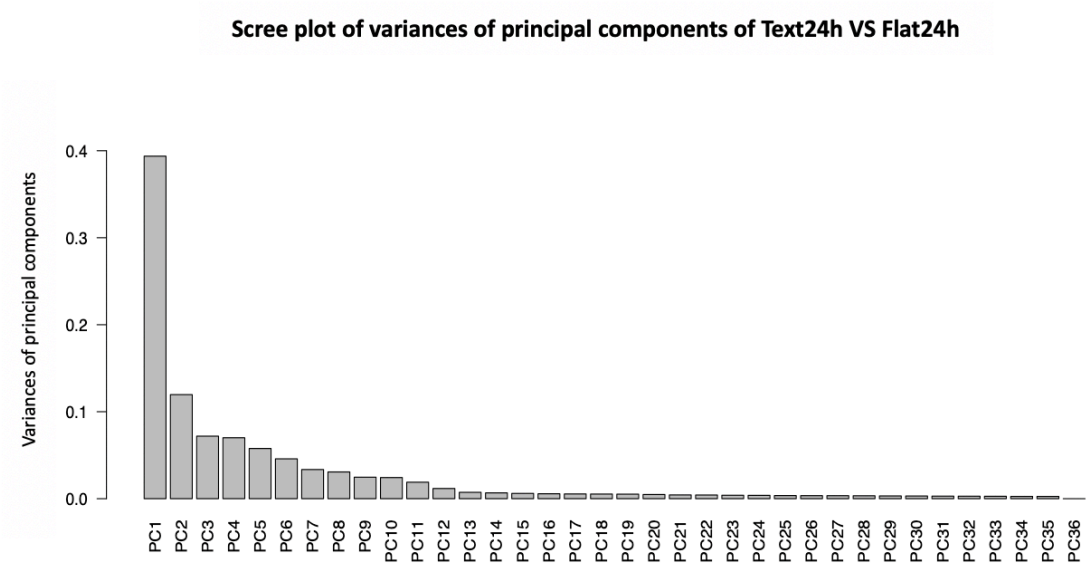


Figure D.1: Scree plot of variances of principal components of Text24h VS Flat24h.

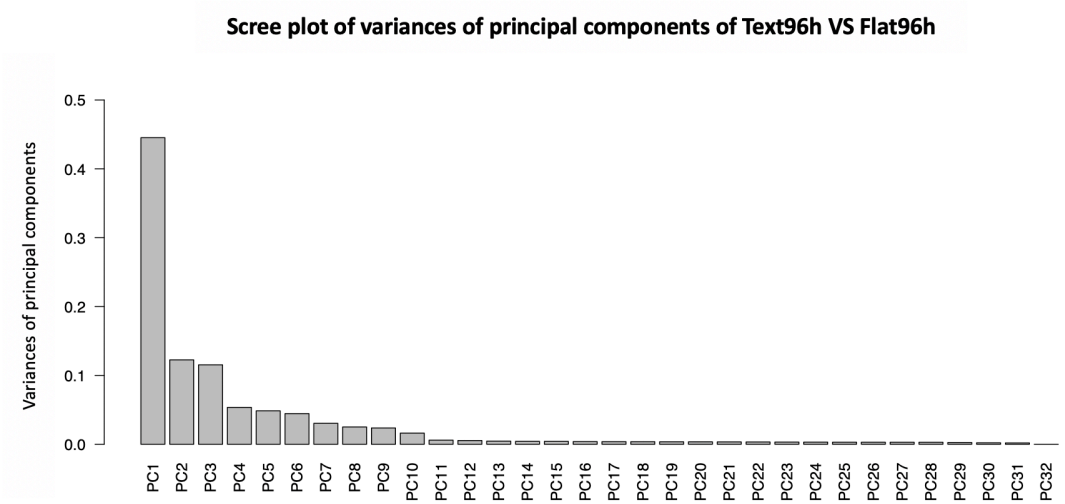
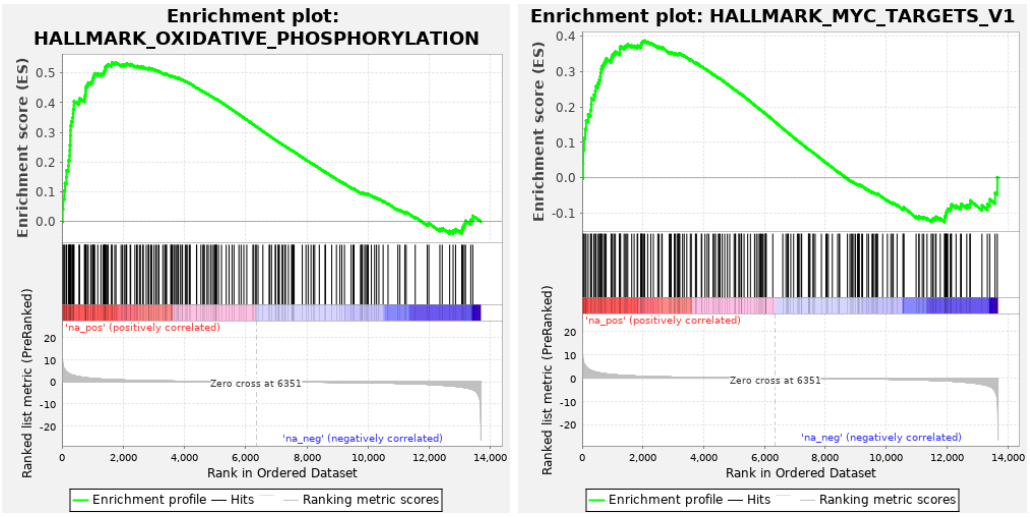


Figure D.2: Scree plot of variances of principal components of Text96h VS Flat96h.

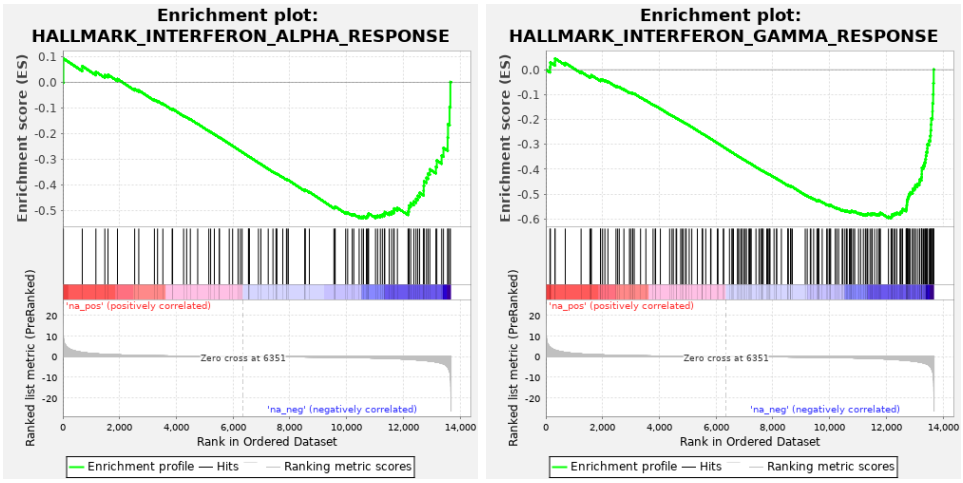
Appendix E

Gene Set Enrichment Score



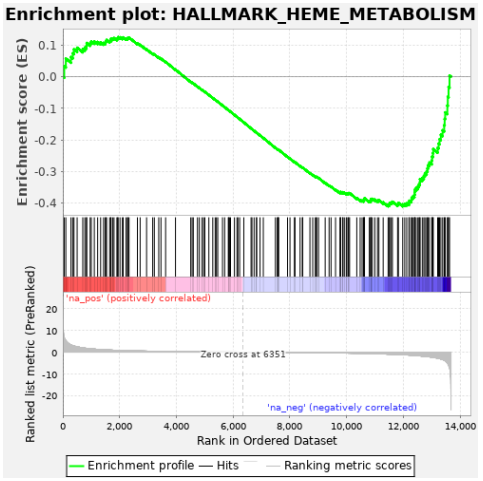
(a) (b)

Figure E.1: Enrichment score of hallmarks enriched in Text24h. (a) Hallmark oxidative phosphorylation. (b) Hallmark Myc targets V1.

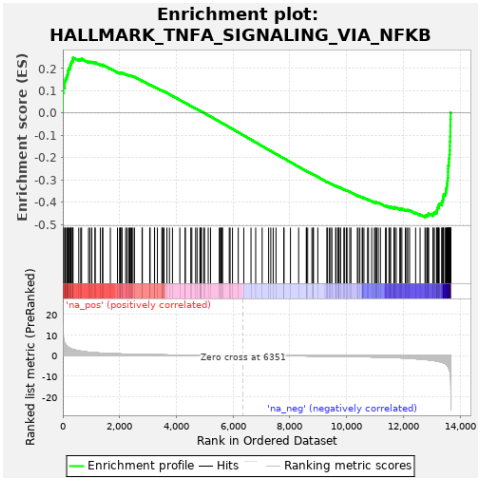


(a)

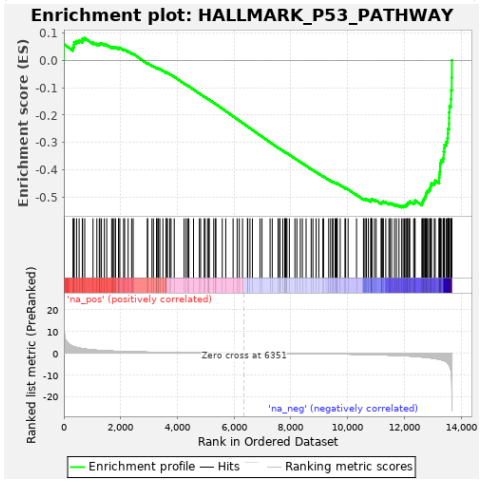
(b)



(c)



(d)



(e)

Figure E.2: Enrichment score of hallmarks enriched in Flat24h. (a) Hallmark interferon alpha response. (b) Hallmark interferon gamma response. (c) Hallmark heme metabolism. (d) Hallmark TNFα signaling via NF$_k$B. (e) Hallmark P53 pathway.

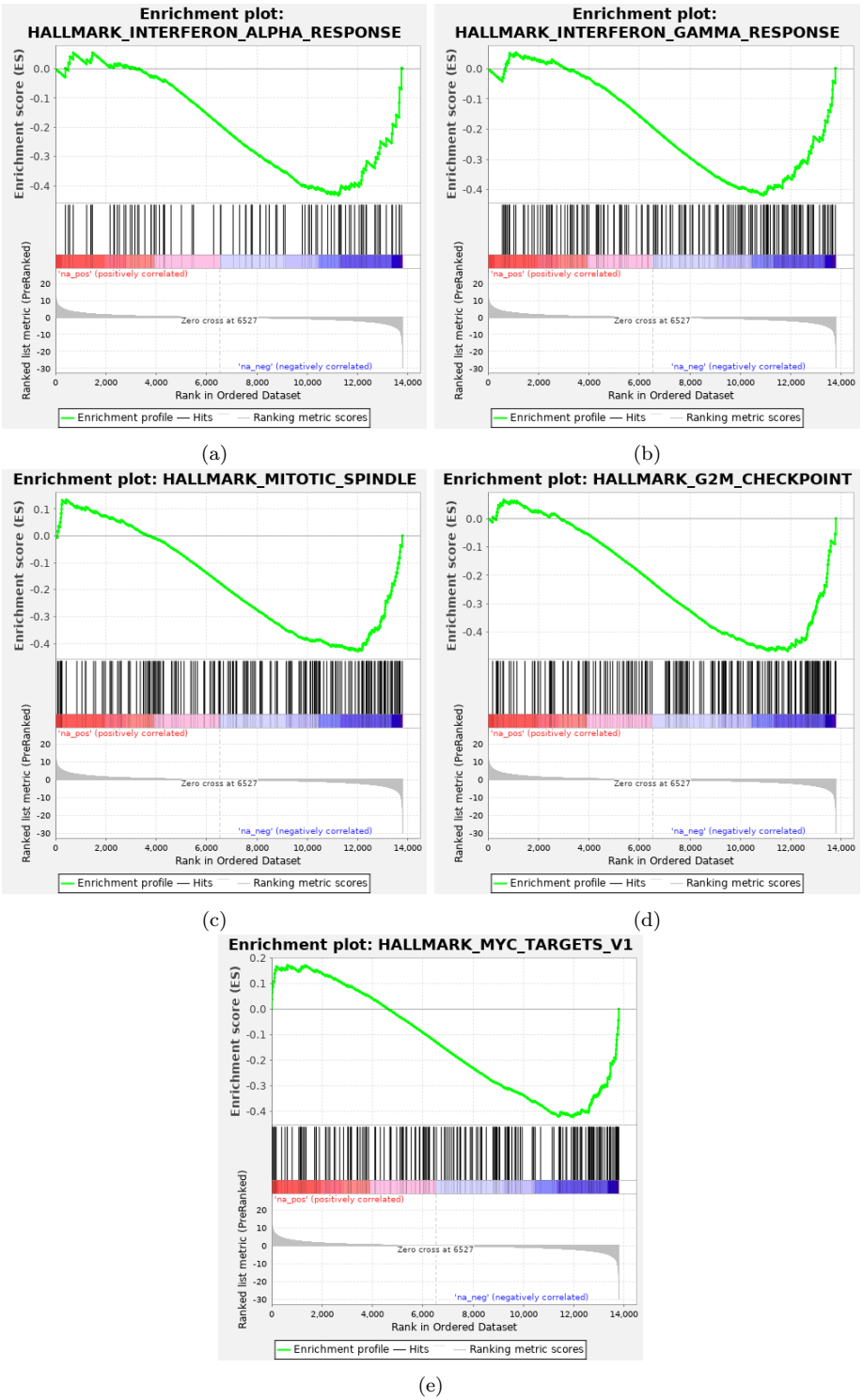


Figure E.3: Enrichment score of hallmarks enriched in Flat96h. (a) Hallmark interferon alpha response. (b) Hallmark interferon gamma response. (c) Hallmark mitotic spindle. (d) Hallmark G2M checkpoint. (e) Hallmark Myc targets V1.