UNIVERSITAT DE BARCELONA

Targeting the MAPK signalling pathway for breast cancer therapy

Carla Xena Bosch

Cancer, molecular biology and pharmacology
Master of Science in Bioinformatics and Biostatistics

Ivette Olivares Castiñeira
Carles Ventura Royo

June 2022

MASTER THESIS CARD

Title:	Targeting the MAPK signalling pathway for breast cancer therapy
Author:	Carla Xena Bosch
Tutor:	Ivette Olivares Castiñeira
SRP:	Carles Ventura Royo
Date of delivery:	June 2022
Studies:	Master of Science in Bioinformatics and Biostatistics
Area:	Cancer, molecular biology and pharmacology
Language:	English
Number of credits:	15
Keywords:	Breast cancer, MAPK pathway, RNA-Seq data, R pipeline

Abstract

Breast cancer is the most commonly diagnosed cancer in the world, with a prevalence that is only increasing. Therefore, it is urgent to find novel therapies for its treatment. The MAPK (mitogen activated protein kinase) signalling pathway is one of the most well-studied pathways in cancer biology, with its hyper-activation accounting for more than 40% of all cancer cases in humans.

In this work, after a bibliographical research of breast cancer, the MAPK signalling pathway, its implication with carcinogenesis and current therapies, an R pipeline for RNA-Seq data analysis of TNBC (Triple Negative Breast cancer) samples treated or not with an MAPK signalling pathway inhibitor (Trametinib) is performed.

Differentially expressed genes identified in the basal-like subtype include CALCB, CCL28, IVL, KRT6A, ADD2, HSD17B4 and PSCA, among others. In the mesenchymal subtype, FAM83A, GPX3, KRT13, KRT6A and TACSTD2 were found. The only gene differentially expressed in both subtypes was KRT6A.

The genes found in the two basal-like cell lines were linked to similar processes: nuclear division, chromosome segregation, sister chromatid segregation and mitotic sister chromatid segregation.

In the mesenchymal subtype, the most enriched categories found were blood vessel development, blood vessel morphogenesis, tissue morphogenesis and angiogenesis.

These findings may imply that treating TNBC with an MAPK inhibitor may affect the cancer cells' proliferation rate, notably in the basal-like subtype.

Therefore, it is concluded that inhibiting the MAPK signalling pathway is a promising technique for treating breast cancer, so further investigation is needed.

Resum

El càncer de mama és el càncer més diagnosticat al món, amb una prevalença en augment. Per tant, és urgent trobar noves teràpies per al seu tractament. La via de senyalització MAPK (proteïna quinasa activada amb mitogen) és una de les vies més ben estudiades en biologia del càncer, ja que la seva hiperactivació representa més del 40% de tots els casos de càncer en humans.

En aquest treball, després d'una investigació bibliogràfica del càncer de mama, la via de senyalització MAPK, la seva implicació amb la carcinogènesi i les teràpies actuals, una pipeline en R ha estat desenvolupada per a l'anàlisi de dades d'ARN-Seq de mostres de TNBC (càncer de mama triple negatiu) tractades o no amb un inhibidor de la via MAPK (Trametinib).

Els gens expressats de manera diferenciada identificats en el subtipus basal inclouen CALCB, CCL28, IVL, KRT6A, ADD2, HSD17B4 i PSCA, entre d'altres. En el subtipus mesenquimal, s'hi ha trobat els gens FAM83A, GPX3, KRT13, KRT6A i TACSTD2. L'únic gen expressat de manera diferencial en ambdós subtipus és el KRT6A.

Els gens trobats a les dues línies cel·lulars de tipus basal estan relacionats amb processos similars: divisió nuclear, segregació de cromosomes, segregació de cromàtides germanes i segregació de cromàtides germanes mitòtiques.

En el subtipus mesenquimal, les categories més enriquides trobades han estat el desenvolupament de vasos sanguinis, la morfogènesi de vasos sanguinis, la morfogènesi dels teixits i l'angiogènesi.

Aquestes troballes poden implicar que el tractament de TNBC amb un inhibidor de MAPK pot afectar la taxa de proliferació de cèl·lules canceroses, sobretot en el subtipus basal.

Per tant, es conclou que la inhibició de la via de senyalització MAPK és una tècnica prometedora per tractar el càncer de mama, per la qual cosa cal investigar més.

Contents

List of Figures

List of Tables

Chapter 1

Summary

Breast cancer is the most commonly diagnosed cancer in the world, with a prevalence that is only increasing. Therefore, it is urgent to find novel therapies for its treatment. The MAPK (mitogen activated protein kinase) signalling pathway is one of the most well-studied pathways in cancer biology, with its hyper-activation accounting for more than 40% of all cancer cases in humans.

This work is divided in two parts. Firstly, a bibliographical research of breast cancer, the MAPK signalling pathway and its implication with carcinogenesis is performed. Afterwards, an overview of therapeutic targets within this pathway and current therapies have been researched.

Secondly, an R pipeline for RNA-Seq data analysis of TNBC (Triple Negative Breast cancer) samples treated or not with an MAPK signalling pathway inhibitor (Trametinib) is performed.

The data are comprised by two different regimes of Trametinib treatment and three different TNBC cell lines, which are part of two different TNBC subtypes.

Differentially expressed genes have been identified in both basal-like and mesenchymal subtypes, with KRT6A being the only gene differentially expressed in both subtypes.

The two basal-like cell lines analysed present differentially expressed genes linked to similar processes: nuclear division, chromosome segregation, sister chromatid segregation and mitotic sister chromatid segregation.

In the mesenchymal subtype, the most enriched categories are blood vessel development, blood vessel morphogenesis, tissue morphogenesis and angiogenesis.

These findings may imply that treating TNBC with an MAPK inhibitor may affect the cancer cells' proliferation rate, notably in the basal-like subtype.

It is concluded that inhibiting the MAPK signalling pathway is a promising technique for treating breast cancer, which warrants further investigation.

Chapter 2

Introduction

2.1 Context and justification

Breast cancer is the most commonly diagnosed cancer in the world [19] and its prevalence is increasing. Therefore, it is of the utmost importance to find new drugs and mechanisms of actions to treat it and improve its survival rates.

Cancer is characterized as a complex disease caused by coordinated alterations of multiple signalling pathways. MAPK cascades are key signalling components that control fundamental processes such as cell proliferation, differentiation, and stress responses. MAPK (mitogen-activated protein kinase) signalling is one of the most well-studied pathways in cancer biology, with its hyper-activation accounting for more than 40% of all cancer cases in humans [9].

Given its high prevalence in cancers, great efforts have been made to develop specific inhibitors against oncogenic mutants in the last decades. Several of these drugs already been approved for cancer treatment and others are being tested in clinical trials.

Therefore, the aim of this Master Thesis is to introduce breast cancer, study the MAPK signalling pathway and its implication with carcinogenesis and to analyse, using statistical tools, datasets of samples treated with a drug targeting the MAPK pathway. Finally, it will be possible to conclude if targeting the MAPK pathway is indeed a good approach for treating breast cancer and, finally, if it is worthwhile to investigate it further.

2.2 Objectives

1. Bibliographic research of the deregulation of the MAPK pathway and its relationship with breast cancer.
   1.1 Study of breast cancer.
   1.2 Study of the MAPK pathway.

1.3 Study of the implication of the MAPK pathway in carcinogenesis.

2. Research and analysis of therapeutic targets within the MAPK pathway for breast cancer treatment.

   2.1 Study of current targets and therapies.

   2.2 Search and selection of drugs' targeting the MAPK pathway datasets in GEO.

   2.3 Development of a pipeline for data analysis of potential new treatments using R.

2.3 Approach and methods

This study is developed using a theoretical-practical approach.

Firstly, a theoretical part is developed, where a bibliographic research is conducted. In it, we study breast cancer, the MAPK pathway and its relationship with carcinogenesis.

On the second part of the project, we will search for current therapeutic targets and therapies within the MAPK pathway for breast cancer and finally we will investigate the potential of a new drug targeting it. For that, we will develop a pipeline using R.

This constitutes the best approach, since it is necessary to first understand the disease and its relationship with the MAPK pathway before being able to perform analyses of drugs targeting it using R. Finally, we will discuss our results.

2.4 Work planning

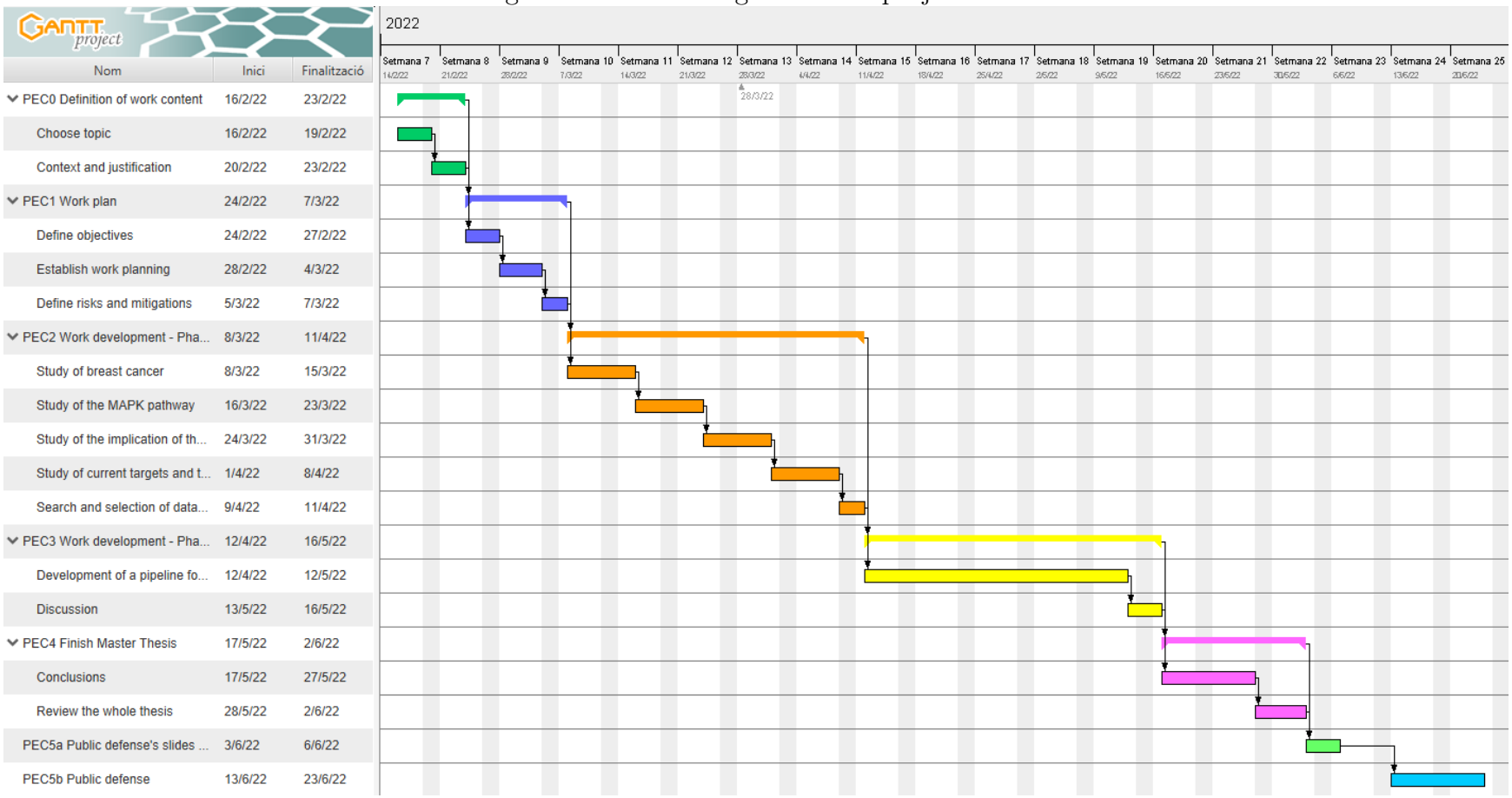
The following section details the activities planned to achieve the objectives defined.

1. Bibliographic review: it starts from a general framework of reference on the disease in question, the MAPK pathway and their relationship. This activity is related to objectives 1.1, 1.2 and 1.3.

2. Search of therapeutic targets: it will have a theoretical component, for which a bibliographic research will be done (objective 2.1) and a practical one, the selection of datasets in GEO (objective 2.2).

3. Pipeline development for the analysis: the methods for the analysis will be decided based on the dataset selected and will be developed in R Studio. This activity is related to specific objective 2.3.

Figure 2.1: Gantt diagram of the project.



As a whole, this study contains a series of milestones marked by the deliveries or continuous evaluation tests (PEC) that will serve for control within the established deadlines.

Milestone	**Date**
PEC0: Definition of the work's content	23/02/2022
PEC1: Work plan	07/03/2022
PEC2: Work development, phase 1	11/04/2022
PEC3: Work development, phase 2	16/05/2022
PEC4: Finish Master Thesis	02/06/2022
PEC5a: Public defense's slides preparation	06/06/2022
PEC5b: Public defense	23/06/2022

Table 2.1: Milestones of the project

A risk analysis of the project is carried out in order to identify possible changes in the planning. The possible risk factors are the following:

- Time risks, due to an over or under-estimation of the deadlines to undertake each of the tasks. This risk has minimized by reviewing the complexity of the tasks and setting realistic goals.
- Overloads in the workplace, which may affect the time dedicated. For this reason, the flexibility of start or end dates of subtasks has been considered.
- Technological risks. These have been addressed by working with tools I am comfortable with, such as RStudio.

2.5 Brief summary of contributions

The main objective of this project is to search for therapeutic targets in the MAPK pathway against breast cancer and to confirm whether drugs targeting it are effective through analysis of public datasets in GEO, obtaining comprehensive and quality results. The product obtained will be the analysis itself and the written report that includes the entire procedure followed.

In Chapter 3, State of the art, a literature review will be performed, with three subsections. Firstly, breast cancer will be introduced, then the MAPK pathway and finally its relationship with carcinogenesis.

In Chapter 4, Methodology, a short bibliographic review of current targets and therapies will be presented, followed by a practical section about the search of public domain datasets in GEO and the pipeline development in R. The data for analysis and experimental design, the data preprocessing and data exploration will be explained.

In Chapter 5, Results, the outcomes of the analysis will be shown. That includes the differential expression analysis, the results annotation and the enrichment analysis.

In Chapter 6, Discussion, the results obtained will be discussed and in Chapter 7, Conclusions, a closure will be provided. That will be followed by possible future lines of investigation and the planning follow-up.

At the end of the document the bibliography and the appendices can be found.

Chapter 3

State of the art

3.1 Breast cancer

Breast cancer is a disease in which cells in the breast grow out of control. It mostly affects women but men can suffer it too. There are different kinds of breast cancer, and they depend on which cells in the breast turn into cancer. Breast cancer can begin in different parts of the breast. A breast is made up of three main parts: lobules, ducts, and connective tissue. The lobules are the glands that produce milk. The ducts are tubes that carry milk to the nipple. The connective tissue (which consists of fibrous and fatty tissue) surrounds and holds everything together. Most breast cancers begin in the ducts or lobules. Breast cancer can spread outside the breast through blood vessels and lymph vessels. When breast cancer spreads to other parts of the body, it is said to have metastasised. [3]

In 2020, there were 2.3 million women diagnosed with breast cancer and 685 000 deaths globally. As of the end of 2020, there were 7.8 million women alive who were diagnosed with breast cancer in the past 5 years, making it the world's most prevalent cancer. Breast cancer occurs in every country of the world in women at any age after puberty but with increasing rates in later life. Breast cancer mortality changed little from the 1930s through to the 1970s. Improvements in survival began in the 1980s in countries with early detection programmes combined with different modes of treatment to eradicate invasive disease. Age-standardized breast cancer mortality in high-income countries dropped by 40% between the 1980s and 2020. Countries that have succeeded in reducing breast cancer mortality have been able to achieve an annual breast cancer mortality reduction of 2-4% per year. If an annual mortality reduction of 2.5% per year occurs worldwide, 2.5 million breast cancer deaths would be avoided between 2020 and 2040.

3.1.1 Symptoms and diagnosis

It is important to note that breast pain is usually not a symptom of breast cancer, so women need to be in the look for other signs. Usually, the first manifestation of breast cancer is a

lump in the breast or an area where the skin has become thicker. Since the majority of breast lumps are not cancerous, some women may decide to ignore them, which is not an appropriate approach. It is imperative to have any lumps checked by a doctor, since early detection of breast cancer is key for a good prognosis.

The following symptoms prompt a visit to a general practitioner [23]:

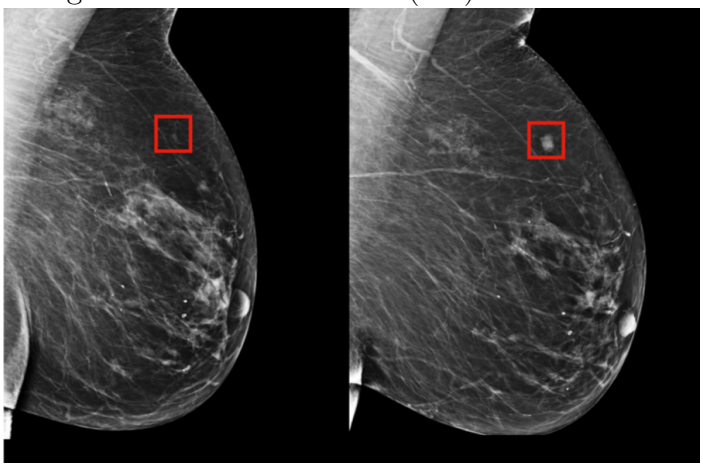
- A change in the size or shape of one or both breasts.
- A change in the look or feel of the skin, such as puckering or dimpling, a rash or redness.
- A lump or swelling in either of the armpits.
- A change in the appearance of the nipple, such as becoming sunken into the breast.
- A new lump or area of thickened tissue in either breast that was not there before.
- A discharge of fluid from either of the nipples.
- A rash (like eczema), crusting, scaly or itchy skin or redness on or around the nipple.

It is important for each individual to know what their breasts look and feel like in different stages of their lives, so any changes and problems can be detected as soon as possible.

If breast cancer is suspected, several of the following procedures, among others, will be performed to determine the diagnosis [5]:

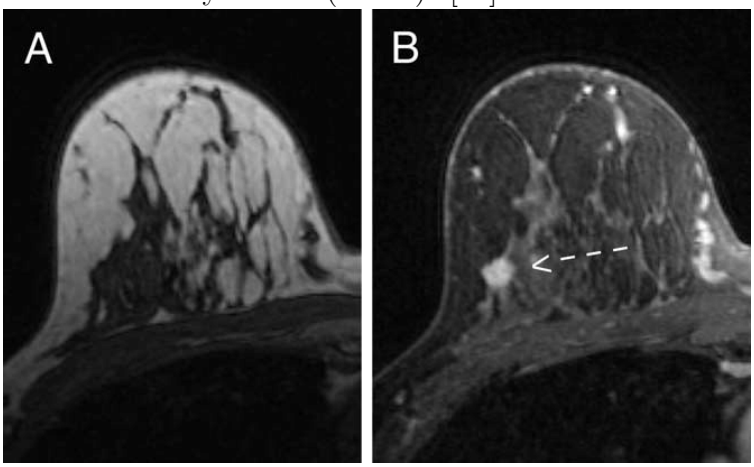
- Breast exam. The doctor will check both breasts and lymph nodes in the armpit. The objective is to feel any lumps or other abnormalities.
- Mammogram. It is basically an X-ray of the breast and they are commonly used to screen for breast cancer. If an abnormality is detected on a screening mammogram, the doctor may recommend another mammogram to further evaluate that abnormality.

Figure 3.1: Mammogram of a normal breast (left) and breast cancer (right). [16]



- Ultrasound. Ultrasounds use sound waves to create images of the breast. They may be used to determine whether a breast lump is a solid mass or a fluid-filled cyst. Ultrasounds are slightly better at predicting breast cancer than mammograms. [29]
- MRI. Breast magnetic resonance imaging consists of an MRI machine that uses a magnet and radio waves to create images of the breast's interior. Before that, an injection of dye may be received to improve the clarity of the images. It must be noted that MRI does not use radiation and it has a higher accuracy in predicting breast cancer than mammograms and ultrasounds, even though this unexpected specificity may lead to over-diagnosis.

Figure 3.2: MRI of breast cancer. (A) Without contrast, the cancer is not easily visible. (B) With contrast, the cancer is readily visible (arrow). [30]



- Biopsy. Removing a sample of breast cells for testing (biopsy) is the only definitive way to make a diagnosis of breast cancer. The doctor will use a specialized needle device guided by an imaging test, such as X-ray, to extract tissue from the suspicious area. Usually, a small metal marker is left at the site of the extraction so it is easier to identify the area on future imaging tests.

  Afterwards, the biopsy samples are sent to a laboratory for analysis, where it is determined by experts if the cells are cancerous or not. In addition, these samples are also analysed to determine the type of breast cancer cells, the aggressiveness of the cancer and whether the cells have hormone receptors. All of this helps decide on a treatment plan.

  There are several types of biopsies, the most used being [11]:

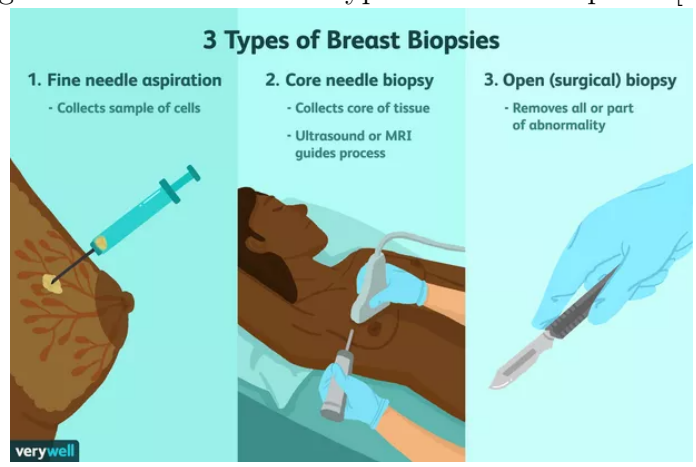
  – Fine needle aspiration (FNA) biopsy. This consists of a very thin needle being placed into the lump or area of concern. Afterwards, a a small sample of tissue or fluid is removed; no incision is needed. An FNA biopsy may be done to help see if the area is a cyst (fluid-filled sac) or a solid lump.

  – Core needle biopsy. In this case a large needle is guided into the lump or area of concern; no incision is needed either. Subsequently, small cylinders of tissue are

removed.

– Open biopsy. A cut is made in the breast and the surgeon removes part or all of the lump or area of concern. If the lump is small, deep and hard to find, a method called wire localization will be used. This consists in a very thin wire being inserted into the breast, with imaging guidance, prior to the surgery, to assist the surgeon in finding the area to excise.

Figure 3.3: Most common types of breast biopsies. [15]



3.1.2 Risk factors and prevention

Anything that enhances an individual's chances of developing an illness, such as breast cancer, is referred to as a risk factor. However, having one or more risk factors does not guarantee that the condition will be developed.

Therefore, two kinds of risk factors can be distinguished, the ones that are inherent and the ones that an individual can modify by changing habits.

Lifestyle-related risk factors for breast cancer comprise the following [26]:

- Drinking alcohol. Alcohol consumption is clearly connected to a higher risk of breast cancer. The danger rises in direct proportion to the amount of alcohol drank. Women who have one alcoholic drink per day face a small (7 to 10%) increase in risk compared to those who do not, whereas those who have two to three drinks per day face a 20 percent increase in risk. Alcohol has also been related to an increased risk of various cancers.
- Being overweight or obese. After menopause, being overweight or obese increases the risk of breast cancer.

  Before menopause, a woman's ovaries produce the majority of her oestrogen, with adipose tissue accounting for only a minor portion of the total. The majority of oestrogen comes from adipose tissue after menopause (when the ovaries stop producing oestrogen). After

menopause, having more fat tissue can enhance oestrogen levels and increase the risk of breast cancer.

Women who are overweight have greater insulin levels in their blood. Some malignancies, especially breast cancer, have been related to higher insulin levels.

- Not being physically active. Regular physical exercise appears to lower the incidence of breast cancer, especially in women after menopause. The biggest concern is the amount of activity required. According to some research, as little as a couple of hours each week may be beneficial, however more appears to be better.

  It's unclear how physical activity reduces the chance of breast cancer, although it could be because of its impact on body weight, inflammation, and hormone levels.

- Being childless. Breast cancer risk is slightly increased in women who have never had children or who had their first child after the age of 30. Breast cancer risk is reduced by having several pregnancies and being pregnant at a young age.

  Still, the impact of pregnancy on the risk of breast cancer is complicated. Breast cancer risk, for example, is increased for the first decade after having a child. As time passes, the risk decreases.

- Not breastfeeding. Breastfeeding appears to reduce the incidence of breast cancer in most studies, especially if it is continued for a year or more. However, this has been difficult to research, especially in nations like the United States, where long-term nursing is unusual.

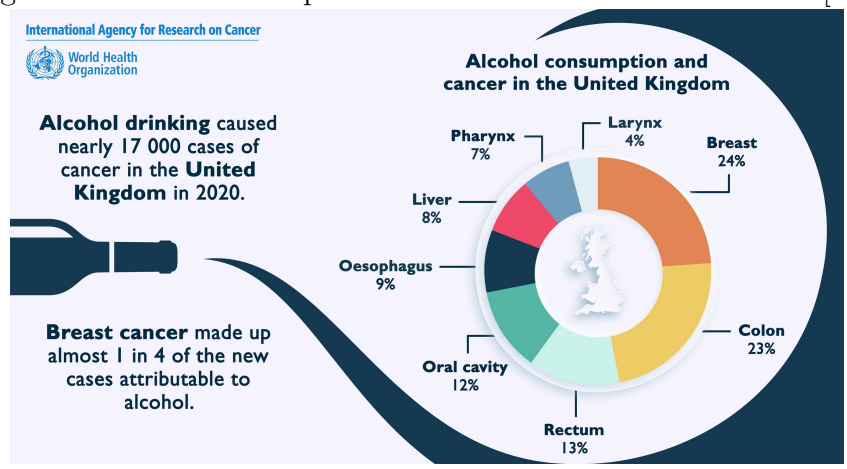
  Breastfeeding may reduce a woman's total number of menstruation cycles over her lifetime, which could explain this effect.

- Birth control. Hormones are used in several birth control methods, which may raise the risk of breast cancer.
  - Oral contraceptives (birth control pills): The majority of research have revealed that women who use oral contraceptives (birth control pills) have a slightly increased risk of breast cancer than women who do not. Within 10 years of stopping the tablets, the risk appears to return to normal.
  - Birth control shots: Some research suggests that receiving long-acting progesterone shots every three months for birth control may increase the risk of breast cancer, although this has not been proven in all trials.
  - Birth control implants, intrauterine devices (IUDs), skin patches, and vaginal rings are all examples of birth control devices. Hormones are also used in these types of birth control, which could potentially accelerate the growth of breast cancer. Although some research have revealed a link between the use of hormone-releasing IUDs and the risk of breast cancer, there have been few studies on the use of birth control implants, patches, and rings with the risk of breast cancer.

- Menopausal hormone therapy. For many years, oestrogen (often coupled with progesterone) has been used as menopausal hormone treatment to assist relieve symptoms of menopause and prevent osteoporosis (thinning of the bones).

  Other than maybe for short-term relief of menopausal symptoms, there aren't many compelling reasons to use post-menopausal hormone treatment. It seems to raise the risk of heart disease, blood clots, and strokes, in addition to the increased risk of breast cancer. It reduces the risk of colorectal cancer and osteoporosis, but the benefits must be balanced against the risks, especially since there are other ways to prevent and cure osteoporosis, and colon cancer can sometimes be prevented by screening.

Figure 3.4: Alcohol is a preventable breast cancer risk factor. [20]



The following are unchangeable breast cancer risk factors [24]:

- Being a woman. This is the leading cause of breast cancer. Breast cancer can strike men as well, but it is far more common in women than in men.
- Age. Breast cancer risk increases with age. The majority of breast cancers are diagnosed in women over 55.
- Inheriting gene mutations. About 5% to 10% of breast cancer cases are believed to be hereditary, meaning they are caused by gene alterations passed down from one parent to the other.

  A genetic mutation in the BRCA1 or BRCA2 gene is the most common cause of hereditary breast cancer. These genes aid in the production of proteins that repair damaged DNA in normal cells. Mutated versions of these genes can lead to abnormal cell growth, which can lead to cancer.

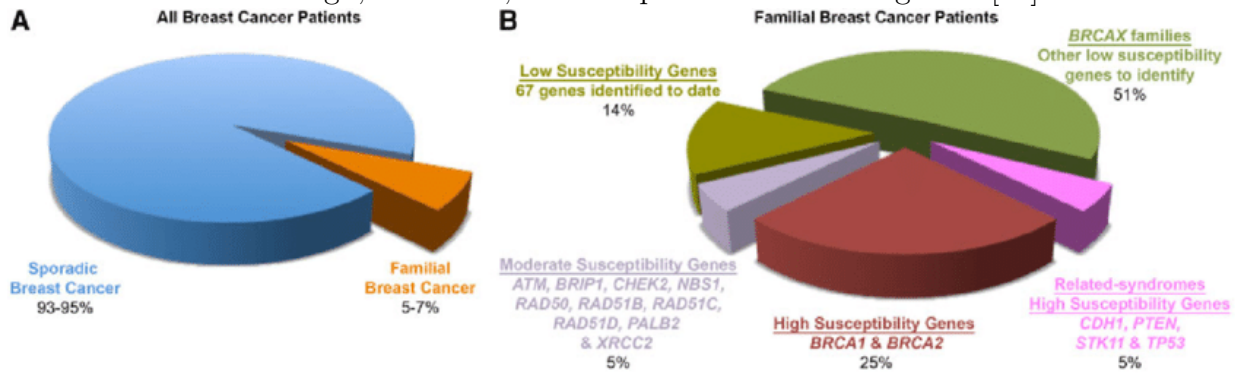
  Other common mutated genes are ATM, PALB2, TP53, CHEK2, PTEN, CDH1 and STK11.

- Dense breasts. Fatty tissue, fibrous tissue, and glandular tissue make up the breasts. On a mammography, breasts with more glandular and fibrous tissue and less fatty tissue appear denser. Breast cancer is more likely in women with dense breasts on mammograms than in women with ordinary breast density. Unfortunately, dense breast tissue can make tumours more difficult to detect on mammograms.
- Early menstruation. Because they began menstruation early (particularly before the age of 12), women who have had more menstrual cycles have a slightly greater risk of breast cancer. A prolonged lifetime exposure to the hormones oestrogen and progesterone may be to blame for the increased risk.
- Late menopause. Women who have more menstrual cycles after menopause (usually beyond age 55) have a slightly greater risk of breast cancer. It's possible that the increased risk is due to a longer lifetime exposure to the hormones oestrogen and progesterone.
- A family history of breast cancer. It's worth noting that the majority of women who develop breast cancer have no family history of the disease. Women with close blood relatives who have had breast cancer are at an increased risk:

  Having a first-degree family with breast cancer (mother, sister, or daughter) almost doubles a woman's risk. Her chance increases by roughly threefold if she has two first-degree relatives. Breast cancer is also more likely in women who have a father or sibling who has had the disease.
- A personal history of breast cancer. A woman who has had cancer in one breast is more likely to develop cancer in the other breast or another region of the same breast. This risk is increased for younger women with breast cancer, despite the fact that it is low overall.
- Being taller. Taller women have a higher risk of breast cancer than shorter ones, according to numerous studies. The reasons for this aren't entirely apparent, but they could have something to do with early-life issues like diet, as well as hormonal or genetic elements.
- Having certain benign breast conditions. Some of these conditions have a stronger link to the risk of breast cancer than others. Some examples are fibrosis and/or simple cysts, mild hyperplasia, usual ductal hyperplasia, fibroadenoma and lobular carcinoma in situ, among others.
- Previous radiation to the chest. Women who received chest radiation therapy when they were younger for another malignancy have a considerably increased risk of breast cancer. This risk is determined by their age at the time they were exposed to radiation. Women who received radiation as a teen or young adult, when their breasts were still developing, are at the greatest danger. Radiation therapy does not appear to raise the risk of breast cancer in older women (beyond the age of 40 to 45).
- Exposure to diethylstilbestrol (DES). Some pregnant women were given an oestrogen-like medicine called DES from the 1940s to the early 1970s in the hopes of lowering their odds of suffering a miscarriage. These women have a slightly higher risk of breast cancer than

others.

Figure 3.5: Distribution of breast cancer patients. a Familial breast cancer represents a minor percentage of all breast cancer patients. b Proportion of familial breast cancer patients due to germ line mutations in high, moderate, and low penetrance cancer genes. [12]



Breast cancer cannot be completely avoided. However, there are several things that may be done to reduce the risk of having it [25].

- Maintaining a healthy weight. Both increasing body weight and weight gain as an adult have been associated to an increased risk of breast cancer after menopause. By combining the food intake with physical activity,it is recommended that people maintain a healthy weight throughout their life and avoid excessive weight gain.
- Being physically active. Many studies have linked moderate to intense physical activity to a lower risk of breast cancer, so it's crucial to do some exercise on a regular basis. Adults should acquire at least 150 to 300 minutes of moderate intensity or 75 to 150 minutes of high intensity activity per week, spread out over the week if possible. It's ideal if the 300-minute mark can be reached or exceeded.
- Avoiding or limiting alcohol. Alcohol consumption, even in little amounts, has been associated to an increase in risk of suffering breast cancer, therefore, it is best not to consume any alcohol. People who do drink should limit themselves to one alcoholic drink per day.

People who are at a higher risk of developing breast cancer (due to a strong family history of the disease or a known hereditary gene mutation that raises breast cancer risk, for example) have several alternatives to help reduce their chances of developing the disease (or finding it early):

- Genetic counselling and testing. Genetic counselling should be discussed with a doctor to see to decided on whether or not to be tested if there are reasons to think that a gene change that increases breast cancer risk may have been inherited.
- Medicines. Tamoxifen and raloxifene are drugs that block oestrogen's activity in breast tissue. Tamoxifen may be administered even if a woman has not reached menopause,

but raloxifene and aromatase inhibitors are exclusively prescribed for menopausic women. Because all of these medications might have negative side effects, it's critical to understand the potential advantages and risks before using one.

- Prophylactic (preventive) surgery. Surgery to remove the breasts (prophylactic mastectomy) may be an option for the small percentage of women who have an extremely high risk of breast cancer, such as from a BRCA gene mutation. Another alternative is to remove the ovaries, which are the body's primary source of oestrogen. While surgery can reduce the incidence of breast cancer, it can't entirely eradicate it, and it comes with its own set of risks.
- Close observation. Some doctors may recommend close observation for women who are at an elevated risk of breast cancer but do not wish to take medications or have surgery. This strategy could involve the following:
  - Breast exams and risk assessments should be done more frequently (every 6 to 12 months, for example).
  - Breast cancer screening should begin at a young age, with yearly mammograms.
  - Another screening test, such as a breast MRI, could be added.

  While this method does not reduce the incidence of breast cancer, it may aid in the early detection of the disease, when it is more treatable.

3.1.3 Classification

Histological subtypes

It must be understood whether the tumor is limited to the epithelial component of the breast or has infiltrated the surrounding stroma, and whether the tumor appeared in the mammary ducts or lobes for morphological analysis of breast cancer.

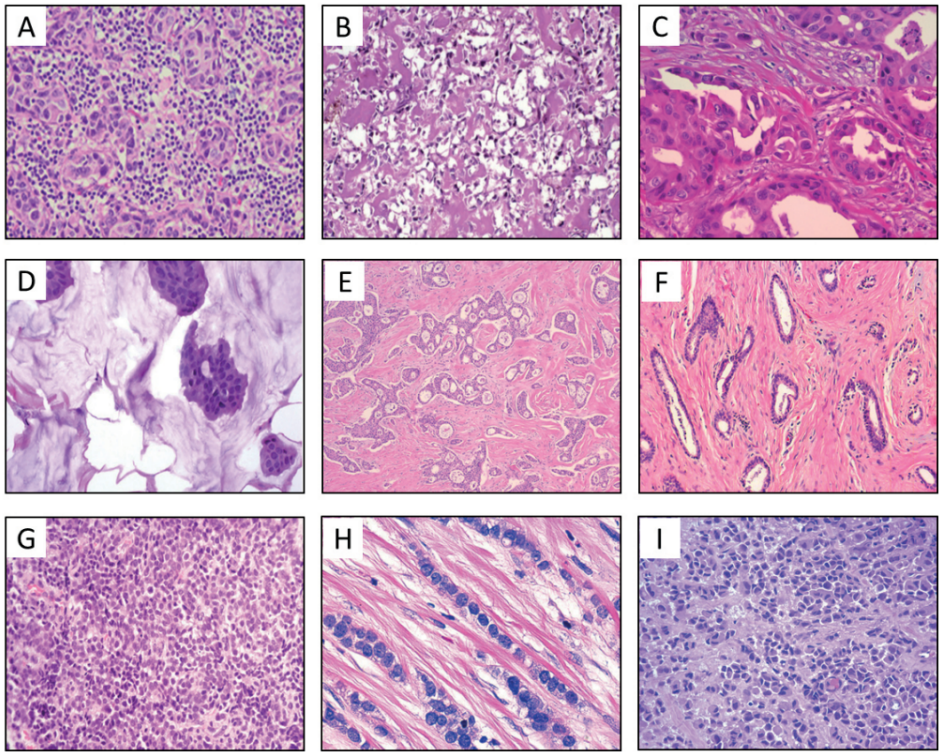
Invasive ductal carcinoma (IDC) accounts for 50 to 80 percent of newly diagnosed breast cancer cases; the other instances are classed as invasive lobular carcinoma (ILC). IDCs can be classed as "no specific type" if they lack enough morphological traits to be classified as a specific histological type; they can also be defined as a "special type" if they have enough distinguishing characteristics and exhibit specific cellular and molecular behaviour.

- Invasive ductal carcinoma no specific type (IDC-NST) IDC-NST is the most frequent histological subtype, accounting for 40 percent to 75 percent of all invasive breast carcinomas. It usually has a wide range of morphological and clinical variance. Tumor cells are poorly differentiated, with protruding nucleoli and many mitoses. In more than half of the instances, necrosis and calcifications can be detected.
- Invasive lobular carcinoma. It is the second most common physiologically different carcinoma, accounting for roughly 5% to 15% of all newly diagnosed cases and primarily affecting women of advanced age. The presence of tiny tumor cells with little atypia,

consistently dispersed throughout the stroma in a concentric pattern, characterizes the classic form of the ILC.

- Medullary carcinoma. A subtype of invasive breast cancer that accounts for about 5% of all occurrences and is linked to better clinical outcomes and lower rates of axillary lymph node involvement. It mainly affects people between the ages of 30 and 40, and it's often linked to BRCA1 gene abnormalities.
- Metaplastic carcinoma. This histological subtype is defined by the presence of a dominant component of metaplastic differentiation, accounting for about 1% of all cases and affecting mostly women after menopause. This type of tumors has aggressive biological behavior and frequently involves lymph nodes.
- Apocrine carcinoma. It accounts for roughly 1% to 4% of all cases, with strong apocrine differentiation accounting for at least 90% of tumor cells. This subtype has a high histological grade and a poor prognosis, and it affects people of all ages, although it is more common in postmenopausal women.
- Mucinous carcinoma. It is a kind of breast cancer that accounts for 2% of all newly diagnosed cases and is also known as colloid, gelatinous, mucous, and mucoid carcinoma. This subtype has been linked to a better prognosis and is most common in women over the age of 60.
- Cribriform carcinoma. A special subtype associated with a favourable prognosis, affecting individuals around the age of 50 and accounting for about 1 to 3.5 percent of all breast cancer occurrences. There is essentially no evidence of regional or distant metastases in cribriform cancer. This subtype is characterized by islands of homogeneous tumor cells with low-grade atypia.
- Tubular carcinoma. A well-differentiated subtype that affects women between the ages of 50 and 60 and accounts for around 2% of all newly diagnosed cases. The majority of tubular carcinomas are linked to a variety of premalignant proliferative lesions. The growth of prominent tubules distinguishes this subtype.
- Neuroendocrine carcinoma. It accounts for roughly 0.5 to 5% of all breast cancer incidences and is more common in people over the age of 50. The markers chromogranin A and synaptophysin are regularly expressed in more than 50% of neoplastic cells in this type of tumor, which is comparable to neuroendocrine tumors of the gastrointestinal tract and lung.

Figure 3.6: Morphological variants representative of the main subtypes of invasive breast carcinomas. (A) medullary carcinoma; (B) metaplastic carcinoma; (C) apocrine carcinoma; (D) mucinous carcinoma; (E) cribriform carcinoma; (F) tubular carcinoma; (G) neuroendocrine carcinoma; (H) invasive lobular carcinoma (ILC); and (I) pleomorphic lobular carcinoma. [14]



Molecular classification

In the year 2000, a benchmark article enhanced awareness of the molecular subtypes of breast cancer. Perou et al. explained in "Molecular Portraits of Human Breast Tumours" that breast cancer might be divided into molecular subtypes based on gene expression patterns [14].

The groups of genes responsible mainly for the segregation of the molecular subtypes of breast cancer are genes related to the expression of estrogen receptors (ER), progesterone receptors (PR), HER2 (Human epidermal growth factor receptor 2), and cell proliferation regulator (Ki-67).

In the stratification of these molecular entities, the immunohistochemical (IHC) panel containing these four biomarkers (ER/PR/HER2/Ki-67) has been found to be effective and significant.

The four subtypes identified are as follows:

- Luminal A. This is the most frequent molecular subtype, accounting for around half of all newly diagnosed breast cancer cases. According to the most recent update from St. Gallen in 2013, the immunohistochemical profile of this subtype was classified as follows: ER+ ($\geq$1%), high PR expression ($\geq$20%), HER2- ($\leq$10%), and low Ki-67 ($<$14%) levels.

These tumors include a variety of low histological grade variants, such as IDC-NST, tubular, cribriform, mucinous, and classic ILC.

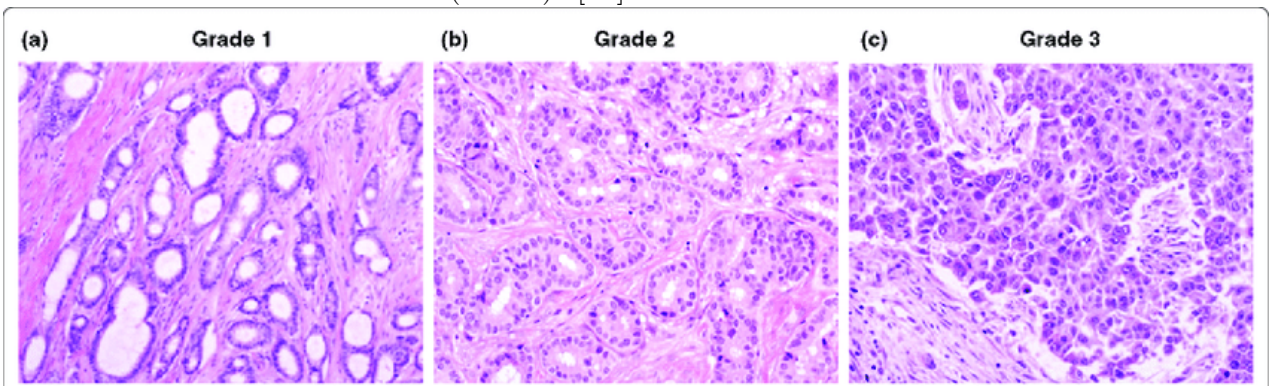
- Luminal B. Approximately 20% to 30% of invasive breast cancer cases are Luminal B. This subtype can be classified immunophenotypically as Luminal B (HER-): ER+ ($\geq$ 1%), PR- or <20%, HER2- ($\leq$ 10%) and high levels of Ki-67 ($\geq$ 20%); or Luminal B (HER2+): ER+ ($\geq$ 1%), HER2+ (>10%) and any level of PR and Ki-67. This molecular subtype, which includes most of the IDC-NST, has a moderate histological grade and is thought to be the most aggressive form of ER+ breast cancer. The key molecular difference between the two luminous subgroups is the higher expression of cell proliferation-related genes.
- HER2+. It accounts for 15% to 20% of all newly diagnosed breast cancer cases. This subtype is defined by a high level of HER2 expression (>10%), ER and PR negativity (<1% and <20%, respectively), and a high level of Ki-67 expression (>20%). HER2 over-expression is almost always found in the pleomorphic ILC variant. The increased expression of the HER2 protein has been linked to cancers with a higher histological grade.
- Triple Negative (TNBC). This type of tumor accounts for ten to twenty percent of all breast cancer cases. This subtype is distinguished by a lack of expression of the hormone receptors ER (<1%) and PR (<20%), and the oncoprotein HER2 ($\leq$ 10%); furthermore, they are highly proliferative tumors, as measured by the Ki-67 index (>30%). The IDC-NST histological type is found in the majority of TN tumors. They include, however, medullary, metaplastic, and apocrine carcinoma forms. Patients with BRCA1 mutations and young women are more likely to develop these tumors, which have a higher histological grade.

Histological grade

When cancer cells are taken from the breast and examined in the lab, they are given a grade. The grade is determined by how similar the cancer cells are to normal cells and is determined by the Nottingham Grading System. The grade is used to anticipate the prognosis and to determine which therapies are most effective. [28]

- Grade 1 or well differentiated. The cells have a lower growth rate and resemble normal breast cells.
- Grade 2 or moderately differentiated. The cells are proliferating at a speed of and look like cells somewhere between grades 1 and 3.
- Grade 3 or poorly differentiated. The cancer cells differ from normal cells in appearance, and they will likely develop and spread more quickly.

Figure 3.7: Histological grade of breast cancer as assessed by the Nottingham Grading System.(a) A well-differentiated tumor (grade 1) that demonstrates high homology to the normal breast terminal duct lobular unit, tubule formation (>75%), a mild degree of nuclear pleomorphism, and low mitotic count. (b) A moderately differentiated tumor (grade 2). (c) A poorly differentiated (grade 3) tumor with a marked degree of cellular pleomorphism and frequent mitoses and no tubule formation (<10%). [22]



Stage

After a person is diagnosed with breast cancer, doctors will try to determine whether the disease has spread and, if so, how far it has spread. This is referred to as staging. A cancer's stage refers to how much cancer is present in the body. It aids in determining the severity of the malignancy and the best treatment options. [27]

The American Joint Committee on Cancer (AJCC) TNM system is the most often used staging classification for breast cancer. The most recent version, effective January 2018, uses the following information:

- The tumor's extent (size) (T).
- The spread to surrounding lymph nodes (N).
- The spread to other parts of the body (metastasis) (M).
- Estrogen Receptor (ER) status.
- Progesterone Receptor (PR) status.
- HER2 status.
- The cancer's grade (G).

The earliest stage breast cancers are stage 0 (carcinoma in situ). It then progresses from stage I (1) through IV (4). The lower the number, the less cancer has spread, in general. A higher number, such as stage IV, indicates that the cancer has progressed farther. An earlier letter inside a stage denotes a lower stage. Although each person's cancer journey is unique, malignancies at similar stages frequently have similar outcomes and are treated in similar ways.

Breast cancer staging has become more complicated than for other cancers due to the addition of information concerning ER, PR, and HER2 status, as well as grade. As a result, the patient's doctor can provide the specific stage of the cancer and what that means.

3.2 The MAPK signalling pathway

3.2.1 Cell signalling

Cells must be able to receive and analyze signals from outside their borders in order to respond to changes in their immediate environment. Individual cells frequently receive multiple impulses at the same time, which they then combine into a single action plan. Cells, on the other hand, aren't merely targets. They also send messages to nearby and faraway cells. [2]

Types of signals

Chemical signals make up the majority of cell signals. Prokaryotic organisms, for example, possess sensors that detect nutrients and guide them toward food sources. Growth factors, hormones, neurotransmitters, and extracellular matrix components are just a few of the chemical signals that multicellular organisms use. These compounds can have local impacts or have the ability to travel large distances.

Mechanical stimuli also have an effect on some cells. Sensory cells in the skin, for example, respond to touch pressure, whereas similar cells in the ear respond to sound wave movement. Furthermore, specific cells in the human vascular system detect fluctuations in blood pressure, which the body employs to keep the cardiac load stable.

Receptors

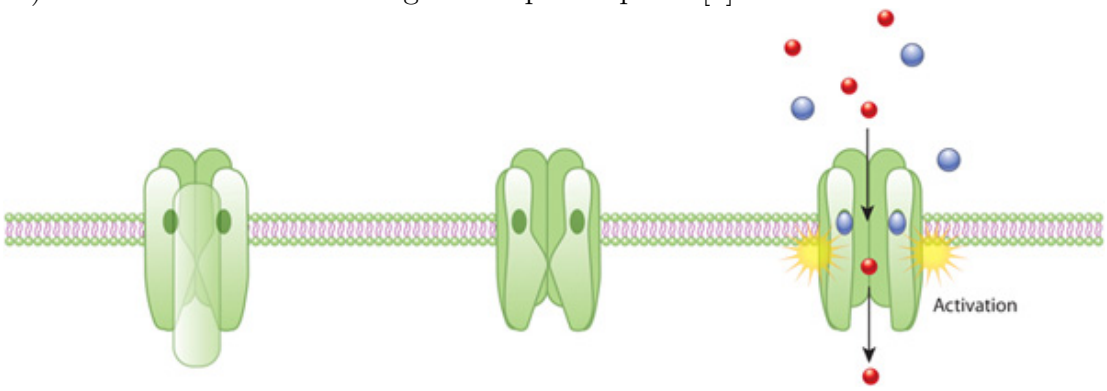
Receptor proteins in cells bind to signalling molecules and trigger a physiological response. Distinct chemicals elicit different responses from different receptors. Insulin receptors bind insulin, nerve growth factor receptors bind nerve growth factor, and so on. In fact, cells have hundreds of different receptor types, with differing populations of receptors in different cell types. Cells are sensitive to events in the atmosphere because receptors can respond immediately to light or pressure.

Receptors are transmembrane proteins that bind to signalling molecules outside the cell and then transmit the signal to internal signalling pathways via a series of molecular switches. G-protein-coupled receptors, ion channel receptors, and enzyme-linked receptors are the three types of membrane receptors.

Membrane receptors interact with both extracellular signals and intracellular molecules, allowing signalling molecules to influence cell activity without entering the cell. Because most signalling molecules are either too large or too charged to traverse a cell's plasma membrane, this is critical.

Not all receptors are found on the cell's exterior. Some can be found deep within the cell, even in the nucleus. Typically, these receptors bind to substances that can pass through the plasma membrane, such as nitrous oxide and steroid hormones like oestrogen.

Figure 3.8: An example of ion channel activation. In the plasma membrane, an acetylcholine receptor (green) forms a gated ion channel. The pore is closed when there is no external stimulus (center). When acetylcholine molecules (blue) connect to the receptor, a change occurs, allowing ions (red) to flow into the cell through the aqueous pore. [2]



Signal transduction pathways

Once a receptor protein receives a signal, it undergoes a conformational change, which triggers a cascade of metabolic processes within the cell. Intracellular signalling pathways, also known as signal transduction cascades, enhance the message by producing multiple intracellular signals for each attached receptor.

The creation of small molecules known as second messengers, which initiate and coordinate intracellular signalling pathways, can be triggered by the activation of receptors.

In conclusion, signals are often delivered to cells in chemical form via a variety of signalling molecules. When a signalling molecule binds to a suitable receptor on a cell surface, it triggers a series of events that not only transport but also amplifie the signal to the cell interior. Cells can also communicate with other cells by sending signalling molecules. Some chemical signals, such as neurotransmitters, travel only a short distance to their destinations, whereas others must travel much further.

3.2.2 Mechanism of action

MAPK cascades are key signalling components that control fundamental processes such cell proliferation, differentiation, and stress responses. These cascades transmit signals through sequential activation of three to five layers of protein kinases known as MAPK kinase kinase kinase (MAP4K), MAPK kinase kinase (MAP3K), MAPK kinase (MAPKK), MAPK and MAPK-activated protein kinases (MAPKAPK). The first three central layers form a basic core unit, although the last two layers appear in some cascades and change depending on cells and

stimuli. Based on the components in the MAPK layer, three primary MAPK cascades have been identified: ERK1/2, c-Jun N-terminal kinase (JNK), and p38 MAPK.

The JNK and p38 MAPK pathways have been linked to cell stress and apoptosis, but the ERK/MAPK signaling system, which is the most extensively researched MAPK signaling pathway, has been linked to cell proliferation and differentiation and plays a key role in the cell signal transduction network.[7]

ERK pathway

The Ras/Raf/MAPK (MEK)/ERK signaling pathway is the most important of all MAPK signal transduction pathways, and it is critical for tumor cell survival and progression. [7]

The ERK/MAPK pathway lies at the heart of the signaling network that controls cell growth, development, and division.

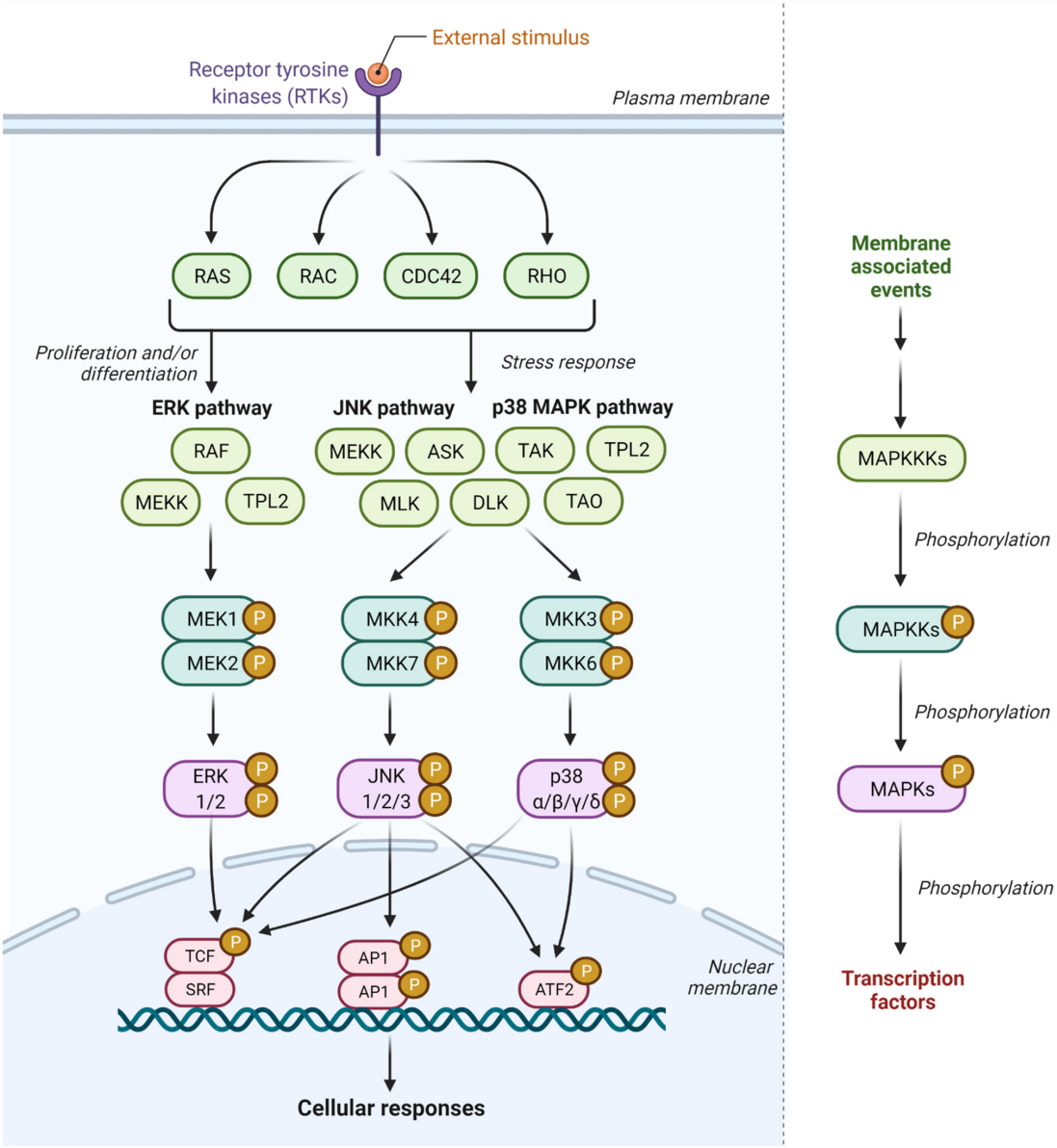
The binding of a ligand (such as a growth factor, cytokine, or hormone) to the extracellular region of two subunits of a receptor tyrosine kinase (RTK), a transmembrane protein, is the starting point. Once the subunits bind to a ligand, they form a dimer and their cytoplasmic domains are phosphorylated. When the RTK is activated, cytoplasmic adaptor proteins (not shown) can attach to it.

The adaptor proteins then attract guanine–nucleotide exchange factors (GEFs)recruit GEFs to the plasma membrane, where they activate a small G protein like RAS. GTPases that are also active in the PI-3K pathway are known as RAS proteins. RAS is typically inactive, binding guanosine diphosphate (GDP). Once GEF displaces GDP from RAF, allowing GTP to bind, RAS becomes transiently active; RAS then cleaves the bound GTP and becomes inactive again.

RAS activates a protein kinase known as a mitogen-activated protein kinase kinase kinase (MAPKKK). BRAF is the MAPKKK in this example, and it enables phosphorylation of the second protein kinase in the cascade, MAPKK (MEK).

MAPKKs have dual selectivity for tyrosine and serine/threonine amino acid residues, which is crucial for activating the third and final enzyme in the cascade, a MAPK (ERK). The MAPK requires the phosphorylation of two tyrosine residues and an almost-adjacent threonine residue to become activated. Completion of double phosphorylation allows MAPK to function as an enzyme and translocate to the nucleus, where it phosphorylates and activates transcription factors. [10]

Figure 3.9: JNK, p38 MAPK, and ERK pathways in MAPK signaling. [21]



JNK pathway

JNK is a member of the MAPK family of proteins that is mostly activated in response to stress. A succession of phosphorylation events activate JNK. Phosphorylation of MAPKKK (e.g., MEKK1-4, ASK1/2, TAK1, MLK2, DLK, and TAO1/2) promotes phosphorylation of MAPKK (e.g., MKK4 and MKK7), which causes dual phosphorylation of JNKs at the threonine (Thr183) and tyrosine (Tyr185) sites on the Thr-Pro-Tyr (TPY) motif. Activated JNKs subsequently phosphorylate Jun proteins (JunB, JunD, and c-Jun), causing them to dimerize with Fos proteins (c-Fos, FosB, Fra-1/2) to create transcription factor activator protein-1

(AP-1) and activate the target genes' transcriptional programs.

Activated JNKs can also regulate the transcription of c-Myc, p53, ETS Like-1 protein (ELK1), activating transcription factor 2 (ATF2), nuclear factor of activated T cells (NFAT), signal transducer and activator of transcription 1/3 (STAT1/3), paired box (PAX) genes, and BCL2 family proteins (e.g., BCL2, BCL-xL, BAD, BIM, and BAX). Multiple physiological processes, including cell proliferation and apoptosis, immunological effects, insulin signaling, and neuronal function, are eventually regulated by these signals. [21]

p38 pathway

Environmental stress (e.g., heat, osmotic, and oxidative stress) and genotoxic stress (e.g., ionizing radiation, ultraviolet (UV) light, and cytotoxic DNA damaging chemicals) primarily activate p38 MAPK. MKK3 and MKK6, as well as, to a lesser extent, MKK4, are the main upstream MAP2Ks implicated in p38 MAPK activation. p38 MAPK activation, like JNK activation, requires dual phosphorylation by MAP3Ks at the Thr-Gly-Tyr (TGY) motif. Once activated, p38 MAPKs translocate from the cytosol to the nucleus, where they activate downstream transcriptional targets such PAX6, ETS1, PRAK, MK3, RAR, AP-1, ATF1 and CHOP, which allows them to regulate cellular processes.

MSK1/2 can activate other transcription factors such as STAT1, NF-B, MEF-2, ELK1, and CREB through activating p38 MAPK. Furthermore, depending on their substrate selectivity, different isoforms of p38 MAPK might activate different types of downstream molecules. Gene expression, cell motility, transcription, and chromatin remodeling will all be regulated by these proteins. [21]

3.2.3 Pathologic deregulations in breast cancer

ERK

Mutations in MAPK effectors cause cancer-related changes in MAPK signalling, affecting the functionality and, as a result, the course of the signalling cascade in both constitutive activation and continuous signal transduction. The MAPK pathways, as previously stated, are a complex regulatory network made up of a number of crosstalking and compensating pathways that are involved in transducing distress and thereby influencing growth signals and cellular metabolism. These effectors are thought to play a role in cancer progression and therapeutic resistance, according to evidence.

The majority of solid tumors are defined by mutations in the RAS/RAF/MEK/ERK signaling pathway genes. BRAF (B-Raf proto-oncogene serine/threonine kinase) and RAS family genes (KRAS and NRAS) mutations are common, although MEK (MAP kinase-ERK kinase) or ERK (extracellular regulated MAP kinase) mutations are less common or rare, respectively. In addition to the primary signal transduction cascade members previously described, mutations also arise in the genes coding for tyrosine kinase receptors (EGFR, c-MET, and c-KIT). [1]

Level	Gene	Mutations	Breast cancer (%)
Receptor	ALK	F1174L, R1275Q, R1245V	7.3
Protein adaptor	SHC1	E343D	0.6
	GRB2		2.2
	PTPN11	Mutations in SH2 and PTP domains	1.3
SOS	SOS1	N993Sfs*5	3.3
RAS	N-RAS	Q61K/E/L, G13R, A59T	0.5
	K-RAS	G12V	1.4
	H-RAS	Q61K	0.6
RAF	BRAF	V600E. F595L. R719P	2.1
	RAF1	L397V	1.6
	ARAF		0.6
MEK1/2	MAP2K1	K57N	1.8
	MAP2K2	c.920-66G > T	0.8
ERK1/2	MAPK3	E367D	0.4
	MAPK1		1.6
NF	NF1	Inactivating nonsense mutations	6.1

Table 3.1: Frequency of mutations in the RAS-MAPK genes of breast cancer (source: COSMIC database, April 6th, 2020). [13]

JNK and p38

The activation of the JNK and p38 MAPK pathways in normal tissue is mostly caused by metabolic stress, DNA damage, cytokines, and growth factors, which govern cell viability. Activation of p38 MAPK signalling has been found to induce the expression of pro-inflammatory mediators such as cyclooxygenase-2 (COX-2) and tumor necrosis factor-α(TNF-α), while activation of JNK causes apoptosis in response to stress, inflammatory, or oncogenic signals.

Both the JNK and p38 MAPK pathways are frequently dysregulated in cancerous cells. Upregulation of JNK and p38 MAPK signalling has been shown in several studies to promote tumor development and cancer cell invasion. However, p38 MAPK signalling is also downregulated in tumor cells, resulting in anoikis resistance and promoting the survival of circulating cancer cells, according to a number of studies. As a result, the role of JNK and p38 MAPK signalling in cancer is still controversial. JNK and p38 MAPK signaling have been hypothesized to have an oncosuppressive or oncogenic effect depending on the cell environment. [21]

Pathway	**Status**	**Clinical implications**
JNK	JNK1,2 activities downregulated	Decreased p-JNK1/2 expression was observed in breast infiltrating ductal carcinoma (IDC) cases and was correlated significantly with increased tumor grade and decreased age at diagnosis.
p38 MAPK	p38α, δactivities upregulated	High levels of active p38αwere correlated with poor prognosis, lymph node metastasis, and tamoxifen resistance in breast cancer patients. High p38δlevels were associated with poor prognosis in breast cancer patients of all tumor subtypes, especially estrogen receptor (ER)-positive/human epidermal growth factor receptor 2 (HER2)-negative types.

Table 3.2: Status of the JNK and p38 MAPKs in breast cancer and their clinical implications. [21]

Chapter 4

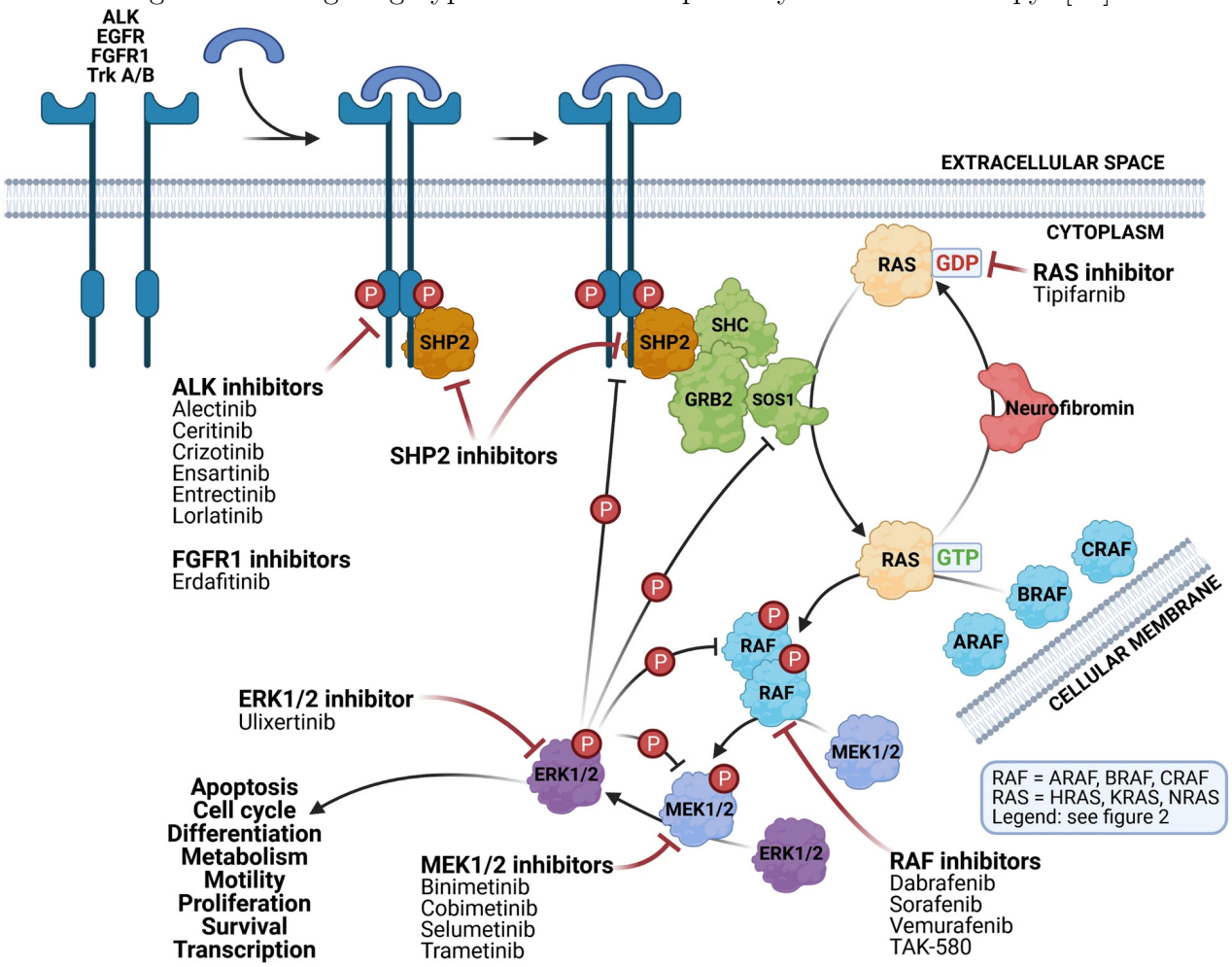
Methodology

4.1 Current targets and therapies

Several molecules targeting the MAPK signalling pathway are being researched and some more have already been approved for cancer therapy.

The figure below shows molecules targeting the RAS-MAPK pathway, which has been more extensively researched than the JNK and p38 pathways. Likely, more public datasets will be available.

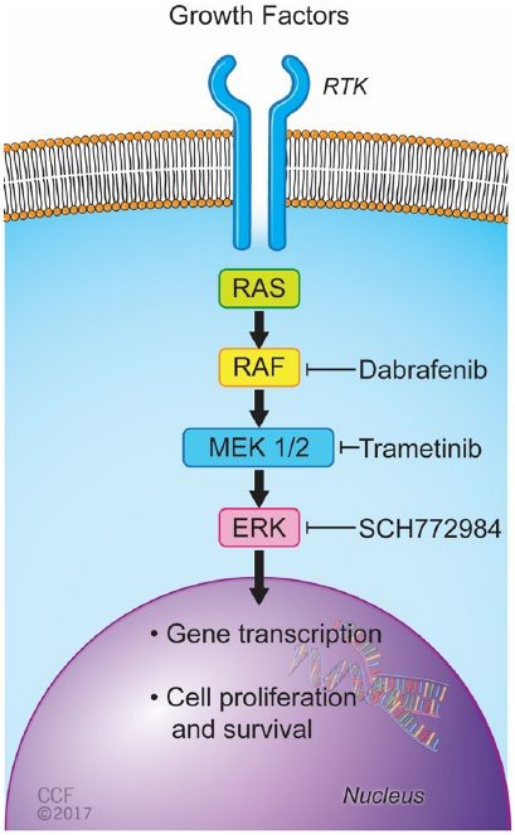
Figure 4.1: Targeting hyperactive MAPK pathway for cancer therapy. [13]



Among those, one specific drug needs to be chosen in order to analyze its effect on breast cancer. Finally, it is decided to continue this work with Trametinib.

Trametinib, the generic name for the chemotherapeutic medication Mekinist sold by Novartis, is a targeted therapy against cancer cells that targets the MEK 1 and MEK 2 protein (kinases). It is commonly used in conjunction with a BRAF kinase inhibitor (i.e. dabrafenib) and it is a once-daily tablet that is taken by mouth. [4]

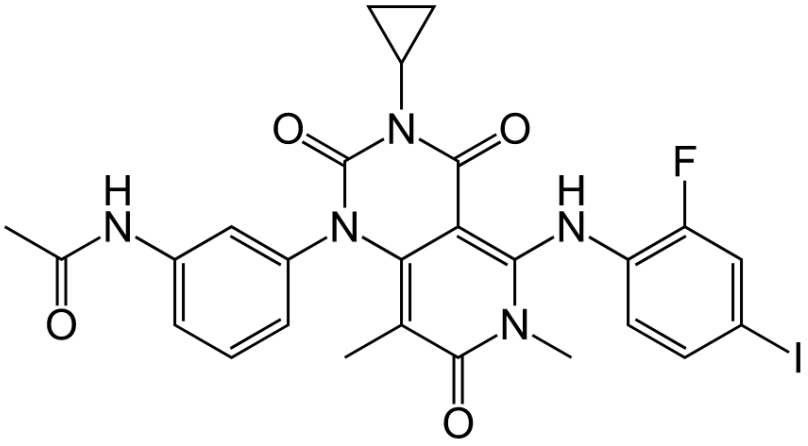
Figure 4.2: Binding of BRAF and MEK inhibitors forms a blockage point in the MAPK pathway at two separate levels, blocking oncogenic downstream signalling and causes cell cycle arrest, which is the mechanism of action of dabrafenib and trametinib. [8]



In both normal and malignant cells, the BRAF gene plays a vital function. BRAF protein is produced as a result of this gene. A mutation in the BRAF gene can change how the BRAF protein functions. The BRAF protein is out of control and signals all of the time, rather than waiting for its turn to signal a cell to divide or develop. This out of control BRAF signalling may fuel the uncontrolled growth of cancer cells. MEK1-2 are proteins that are further down the molecular chain in this pathway.

Trametinib interferes with the signal at the MEK 1-2 section of the chain, which may result in a slowing of tumour growth.

Figure 4.3: Chemical formula of trametinib. [31]



4.2 Selection of datasets in GEO

It is decided to obtain data from Gene expression Omnibus (GEO).

GEO is a public genomic data repository, powered by the National Center for Biotechnology Information, where gene expression studies based on microarray or RNA-seq, among other techniques, can be found. GEO series with the terms "breast cancer trametinib" and "mek inhibitor" in the description are searched. [6]

Microarray or RNA-Seq data are preferred over other types, because of my experience analysing them.

From the first search we obtain only one result, accession GSE82032. The second one provides us 38 results. Of those, we chose GSE138780, which is the only one that corresponds to breast cancer in RNA-Seq format.

Accession	Title	Series Type	Taxonomy	Sample count	Release Date
GSE82032	Human basal-like breast cancer cell line HCC1143 treated with BET inhibitor JQ1 in combination with MEK inhibitor Trametinib or PI3K/mTOR inhibitor BEZ235	Expression profiling by high throughput sequencing	Homo sapiens	21	May 30, 2017
GSE138780	Discrete Adaptive Responses to MEK Inhibitor In Subpopulations of Triple-Negative Breast Cancer [RNA-seq]	Expression profiling by high throughput sequencing	Homo sapiens	62	Nov 10, 2021

Table 4.1: GEO datasets selected for the analysis [18] [17]

For the set of selected series, a preprocessing and subsequent analysis of differential expression will be carried out with different packages in RStudio.

Lastly, the biological analysis and interpretation of the results obtained will be carried out in order to identify the differences between samples treated with Trametinib and control ones.

The R 4.1.2 environment and the 2021.09.1-372 Rstudio interface were used in this study. The majority of the R packages employed belong to the Bioconductor project, which provides a large number of packages for genomic data analysis, graphic representation, and gene annotation.

4.3 Pipeline development in R

4.3.1 Data for analysis and experimental design

The cell lines to be analyzed have the following characteristics:

1. HCC1143 TNBC Stage IIA, grade 3 basal-like.
2. HCC1806 TNBC Stage IIB, grade 2 basal-like.
3. SUM-159 TNBC, mesenchymal.

The objective of the analysis is to identify which genes are differentially expressed between treated and untreated samples with Trametinib, in the three different cell lines and in the two TNBC subtypes.

Cell line	TNBC subtype	No. of samples	Dataset
HCC1143	Basal-like	3	GSE82032
HCC1806	Basal-like	3	GSE138780
SUM-159	Mesenchymal	3	GSE138780

Table 4.2: Control samples for the analysis

Drug	Dose	No. of samples	Dataset
Trametinib	1µM for 3 days	3	GSE82032
Trametinib	0.03µM for 1 day	3	GSE138780

Table 4.3: Treatment samples for the analysis

The groups compared will be as follows:

- Trametinib 1µM for 3 days vs HCC11443 basal-like control
- Trametinib 0.03µM for 1 day vs HCC1806 basal-like control
- Trametinib 0.03µM for 1 day vs SUM-159 mesenchymal control

The data provided is not obtained directly from the sequencing process, images or FastQ files, but the result of pre-processing them, that is, a table of counts, stored in Rawcounts.csv files, which contain the number of "reads" for each sample in each transcript. We created three counts files: one for each comparison.

Reading the counts matrices

The data from the Rawcounts.csv files must be read. It is handy to convert such an object to an array, which is the format used by most parsing packages.

With the information about the groups or other covariates or auxiliary variables, an object will be created (the usual "targets") that must be synchronized with the previous one.

Figure 4.4: Targets file used for the analysis. The targets file will be the same in all comparisons.

```
   sample      group cols
C1     C1    Control  red
C2     C2    Control  red
C3     C3    Control  red
T1     T1 Trametinib blue
T2     T2 Trametinib blue
T3     T3 Trametinib blue
```

The row names of the "targets" object must match - or at least identify - the column names of the counts matrix.

4.3.2 Data preprocessing

Standardization of counts

In addition to filtering, it is good to express the counts in "CPMs" that is 'counts per million', which will not modify the filtering results, but will standardize the values, which is useful and necessary for subsequent analyses.

For standardization, the cpm() function of the edgeR package will be used.

As can be seen, the cpm function only affects the highest counts.

Once the data is as CPMs, we proceed to filter them.

Filtering of poorly expressed genes

Genes with very low counts in all libraries provide little evidence of differential expression, so it is usual to eliminate those genes that are either not very variable or have little or no expression in most samples.

In this case, following the indications provided, it is decided to conserve only those genes that present some value in at least three samples of each group.

We thus obtain filtered matrices of genes.

Using specific classes to handle data

When working with different objects referring to the same data, such as the count matrix and the "targets" object, it is useful to have container classes that allow working with all of them at the same time, which not only facilitates the work but also helps to avoid 'desynchronizations'.

For counting data it is usual to use a class called 'DGEList' designed to handle counting data, defined in the edgeR package. This class, simpler than the previous ones, uses lists to store read counts and associated information from sequencing technologies or digital gene expression.

Although we could have created the object from all the samples, and performed the gene and sample extraction afterwards, we have chosen not to do so to make it easier to track the process.

Normalization

In addition to standardizing the counts, it is important to eliminate other composition biases between libraries. This can be done by applying normalization by the TMM method which generates a set of normalization factors, where the product of these factors and the library sizes define the effective size of the library.

The calcNormFactors function, from the edgeR library, calculates the normalization factors between libraries.

This will not modify the matrix of counts, but will update the normalization factors in the DGEList object (they default to 1).

That is, even if no changes are observed in the count matrix, when these normalization factors are used in some calculation, the importance of the different columns will be taken into account.

To summarize: the analyzes carried out below will be based on the matrix of counts, filtered, standardized and normalized, on which logarithm base two is also taken.

This will be our starting matrix for the following analyses.

4.3.3 Data exploration

Once the low expressed genes have been discarded and with the counts stored in a DGEList object, we can proceed to make some exploratory plots to determine if the data appear to be of good quality and/or if they present any problems.

Counts distribution

A boxplot with the data, normalized or not, shows that the distribution of the counts is very asymmetric, which justifies the decision to work with the logarithms of the data.

Figure 4.5: Boxplot of normalized counts distribution for the first comparison.

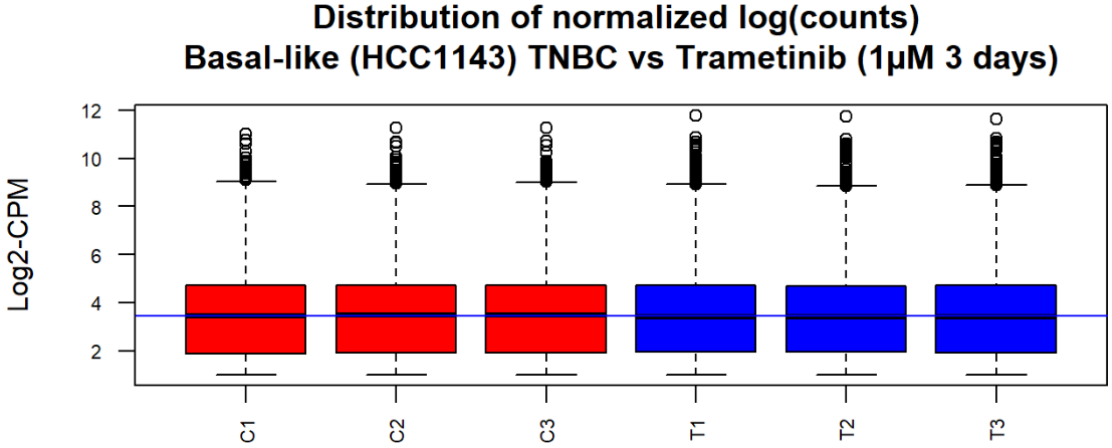


Figure 4.6: Boxplot of normalized counts distribution for the second comparison.

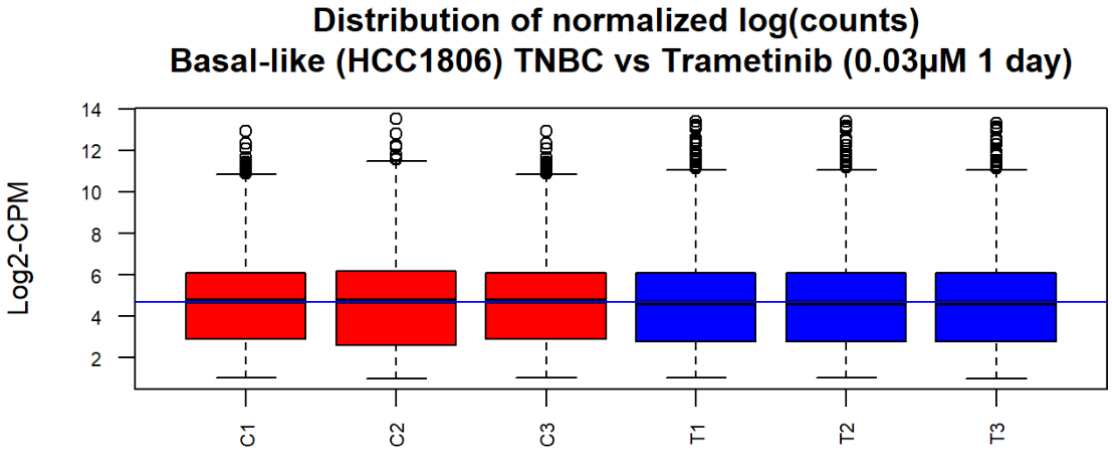
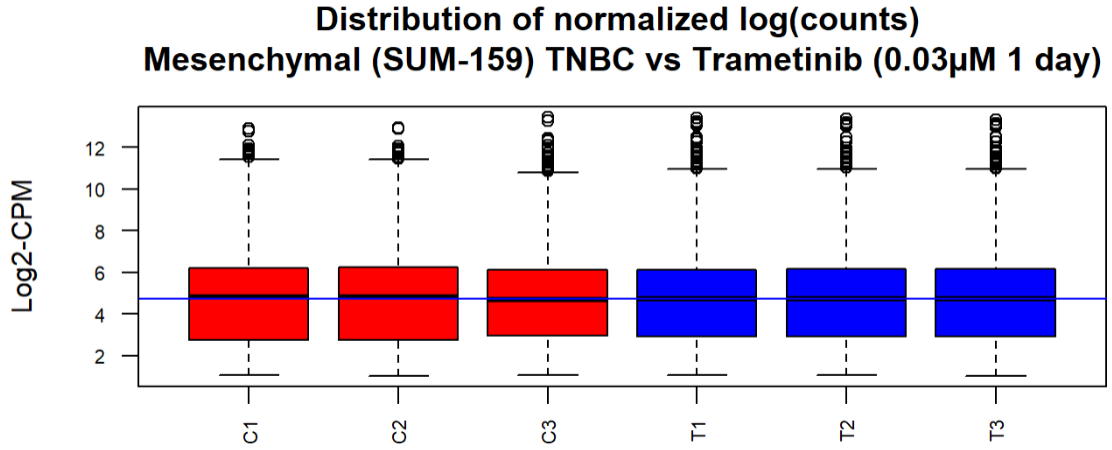


Figure 4.7: Boxplot of normalized counts distribution for the third comparison.



The logarithmic transformation can be done directly but it is better to use the cpm function, as has been done, which adds a small amount to avoid taking logarithms of zero.

Similarity analysis between samples

In general, in an experimental study where we seek to compare different conditions or treatments, we will expect the samples belonging to the same group to be more similar to each other than to those of the other groups.

This intuitive idea can be realized through calculating and visualizing in some way the similarity between the samples.

This can be done in different ways, but some of the most common are hierarchical clustering and dimension reduction methods such as principal component analysis (PCA) or multidimensional

scaling (MDS). The latter has the advantage that it allows the similarities between samples to be visualized in a reduced dimension, rather than direct data, which is what PCA does.

Distance between samples The dist function allows us to compute a distance matrix containing the two-by-two comparisons between all samples. By default an Euclidean distance is used.

Distance matrices can be directly visualized with a heatmap.

Figure 4.8: Heatmap of the basal-like (HCC1143) TNBC vs Trametinib (1μM 3 days) samples.

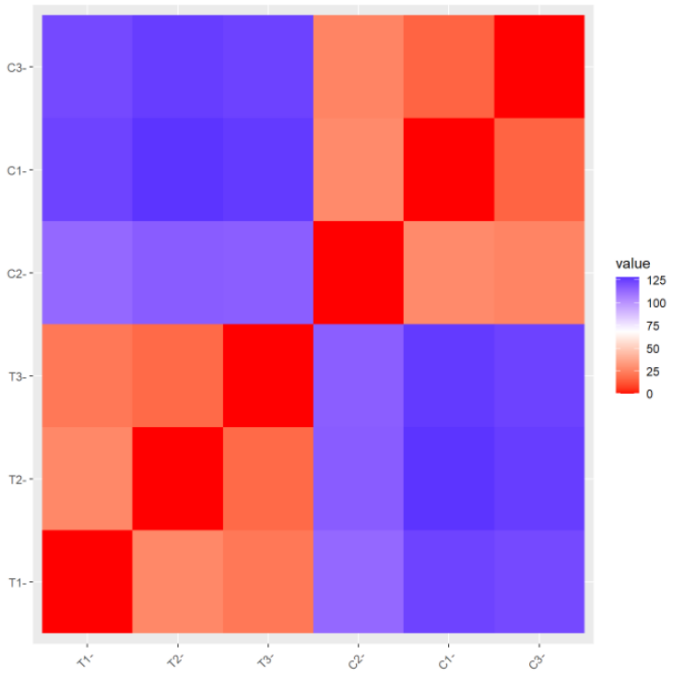


Figure 4.9: Heatmap of the basal-like (HCC1806) TNBC vs Trametinib (0.03μM 1 day) samples.

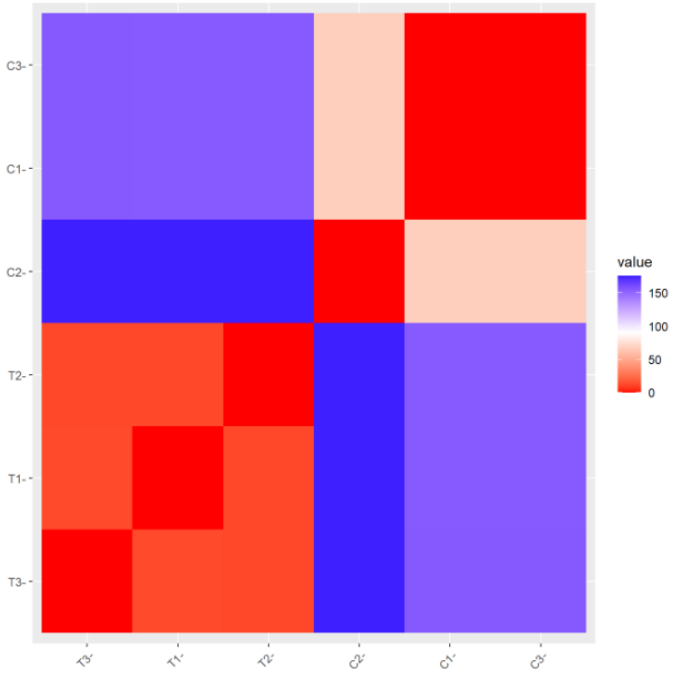
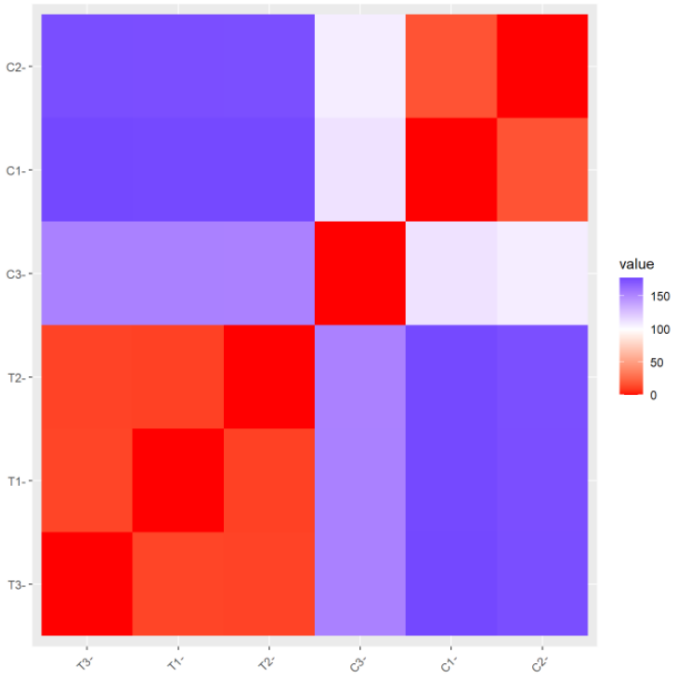


Figure 4.10: Heatmap of the mesenchymal (SUM-159) TNBC vs Trametinib (0.03μM 1 day) samples.



As it can be seen the samples tend to be grouped by Control vs Trametinib factor. This

behaviour is stronger in the first comparison.

Hierarchical clustering A hierarchical clustering provides an alternative representation, also based on the distance matrix.

Figure 4.11: Hierarchical clustering of the first comparison.

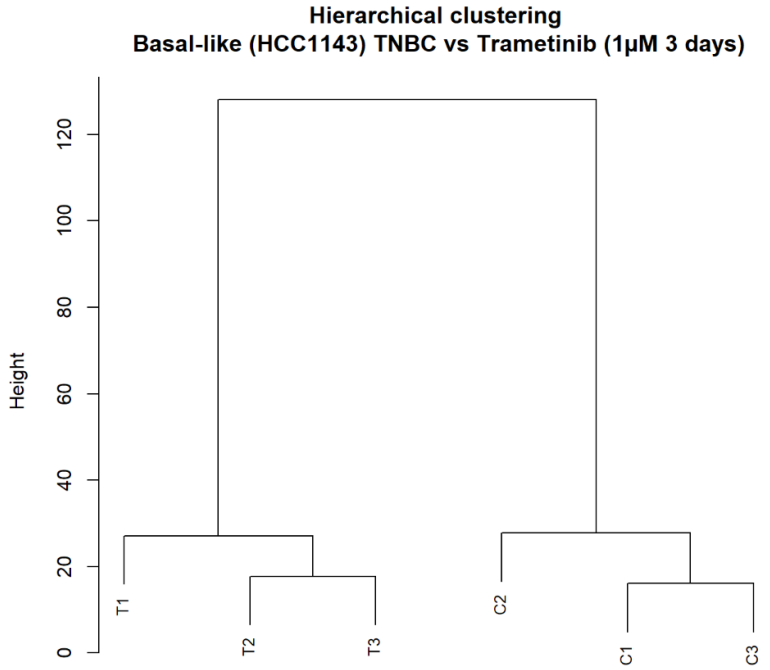


Figure 4.12: Hierarchical clustering of the second comparison.

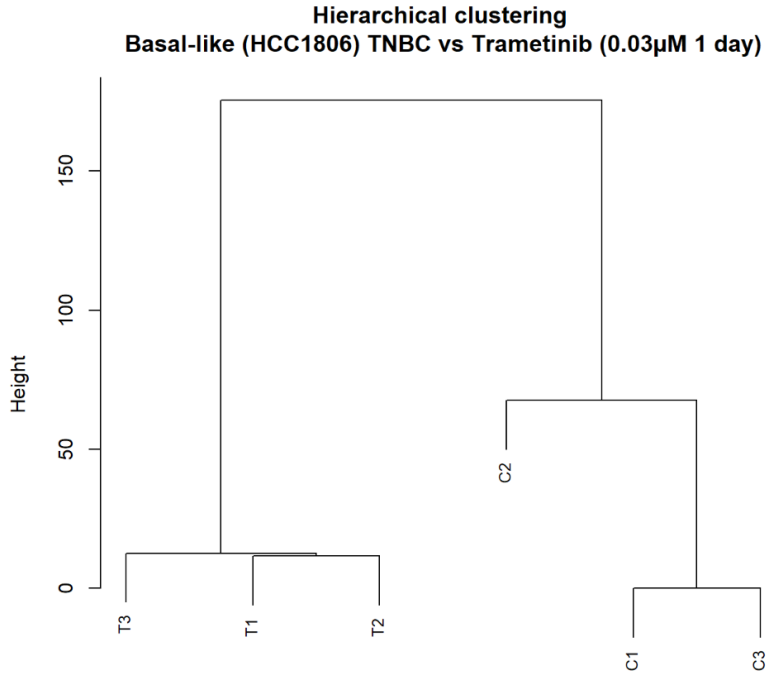
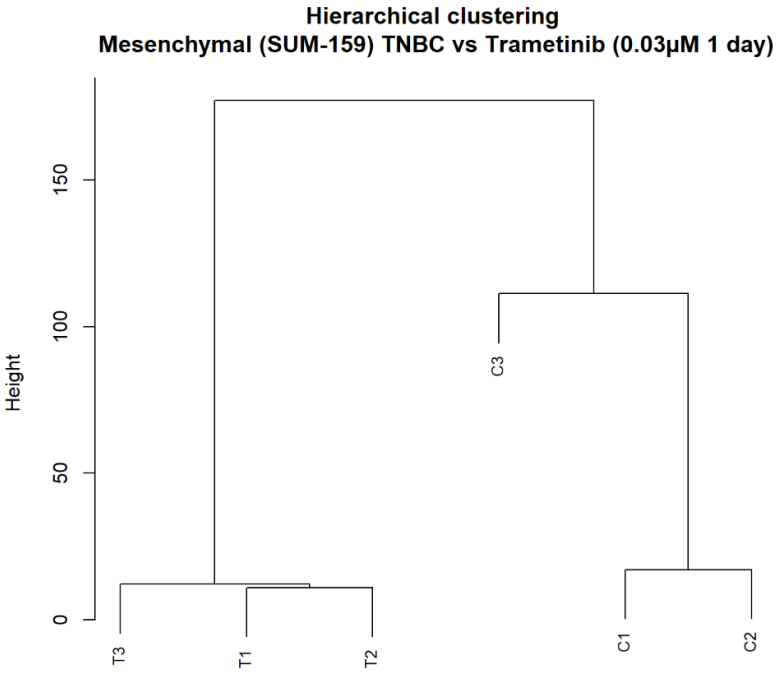


Figure 4.13: Hierarchical clustering of the third comparison.



The dendrograms show the same grouping in all three comparisons, Control on one hand and Trametinib on the other, as expected. Therefore, no further action is needed.

Reduced dimension visualization A complementary approach to determine the main sources of variability in the data is visualization in reduced dimension, either of the data or of the similarity matrix.

For the first representation, it is usual to rely on the result of a principal component analysis (PCA) that represents the directions along which the variation in the data matrix is maximum, with the advantage that said directions are orthogonal (i.e. say independent) and that each one explains more information than the next, so that with a few dimensions it is usually possible to explain a high percentage of the variability.

Similarly, multidimensional scaling allows carrying out a transformation similar to that of PCA, but with the distance matrix, which provides a reduced dimension representation that describes with relative fidelity the differences and similarities between samples.

For this second representation we will use the plotMDS function: a multidimensional scaling plot of distances between gene expression profiles. This function allows us to plot samples on a two-dimensional scatterplot so that distances on the plot approximate the typical log2 fold changes between the samples.

To make it more informative, we can color the samples according to the grouping information. These graphs are available in the Appendix.

The plots show the same natural grouping as the previous visualizations, suggesting that the groups being compared appear to be well defined.

Batch effect detection

Batch effect detection can be done either by viewing the samples or by analyzing the covariates. Since there are no covariates, it only remains to observe the visualizations to determine if they suggest some "non-natural" grouping, that is, not linked to the variable "Treatment".

Since this is not the case, i.e. the data is grouped by Control/Trametinib, we conclude that there does not appear to be any overlapping source of variation, which could be attributed to a possible batch effect.

Chapter 5

Results

5.1 Differential expression analysis

The goal of differential expression analysis is to select genes whose expression differs between groups.

As these are counts, which are not continuous variables, the comparison can be carried out using generalized linear models or extensions of these, created specifically for sequencing data.

This is specifically the case for edgeR or DESEQ2.

An alternative to these packages is to use the limma package, which offers the voom function, which transforms the "reads" counts into logCMM, taking into account the mean-variance relationship in the data and allowing them to be analyzed using the usual approach based on linear models. .

5.1.1 Analysis using Limma-Voom

The main advantage of this approach is that it allows working with all the flexibility of linear models to represent experimental designs, and, in many cases, taking advantage of the user's previous experience in handling limma.

Design and contrast matrix

Using the group variable we can define a design matrix and, on top of it, the contrasts that interest us.

Figure 5.1: Design matrix of the first comparison, which is identical to the other two matrices.

```
   Control Trametinib
C1       1          0
C2       1          0
C3       1          0
T1       0          1
T2       0          1
T3       0          1
attr(,"assign")
[1] 1 1
attr(,"contrasts")
attr(,"contrasts")$group_GSE82032
[1] "contr.treatment"
```

Since we are interested in the differences between the groups, we need to specify which comparisons we want to make. The comparisons of interest can be specified using the makeContrasts function. The contrast matrix indicates which columns of the design matrix we are going to compare. In this case only one comparison will be carried out: control vs Trametinib.

Counts transformation

As indicated, it is not possible to apply a normal linear model with counting data. the voom transformation will create a new object with fields equivalent to those of the DGELIST, in which the counts have been transformed so that they can be analyzed using linear models. Or, better said, so that the inferences made using a normal linear model are valid.

Selection of differentially expressed genes

As in the case of microarrays, the voomObj object and the design and contrast matrices are used to fit a model and then perform the specified comparisons on the fitted model. The process ends with the regularization of the error estimator using the eBayes function.

Top tables

The results of a differential expression analysis can be extracted with the topTable function. This function generates a table of results whose columns contain information about the genes and the difference between the compared groups. Specifically:

Finally we merge our limma toptable with the gene information obtained, adding three columns: ACCNUM/ENTREZID, SYMBOL and GENENAME.

Figure 5.2: Toptable obtained with limma for the basal-like (HCC1143) TNBC vs Trametinib (1μM 3 days) comparison.

	ACCNUM	logFC	AveExpr	t	P.Value	adj.P.Val	B	SYMBOL	GENENAME
1	NM_000728	-4.661891	2.2247233	-79.54742	2.485848e-15	8.071524e-11	25.04444	CALCB	calcitonin related polypeptide beta
2	NM_005547	-5.839554	3.1203254	-76.16768	3.832906e-15	8.071524e-11	24.70675	IVL	involucrin
3	NM_005554	-5.058731	7.5579641	-69.99643	8.900784e-15	8.727363e-11	24.42866	KRT6A	keratin 6A
4	NM_001301875	-4.061446	5.2087189	-65.51400	1.721650e-14	8.727363e-11	23.72934	CCL28	C-C motif chemokine ligand 28
5	NM_148672	-4.065752	5.2736275	-65.47057	1.733068e-14	8.727363e-11	23.73234	CCL28	C-C motif chemokine ligand 28
6	NM_001135091	-3.718688	1.2380001	-65.33459	1.769356e-14	8.727363e-11	23.45293	MUC15	mucin 15, cell surface associated

Figure 5.3: Toptable obtained with limma for the basal-like (HCC1806) TNBC vs Trametinib (0.03μM 1 day) comparison.

	ENTREZID	logFC	AveExpr	t	P.Value	adj.P.Val	B	SYMBOL	GENENAME
1	3295	-8.316299	3.476463	-150.60531	2.282836e-11	3.192546e-07	15.12739	HSD17B4	hydroxysteroid 17-beta dehydrogenase 4
2	11259	-7.239602	2.790934	-126.36530	6.129843e-11	4.286293e-07	14.74015	FILIP1L	filamin A interacting protein 1 like
3	79614	-7.629317	3.214267	-113.98709	1.095057e-10	5.104788e-07	14.45175	*NA*	*NA*
4	119	6.726498	3.385508	98.21779	2.531347e-10	6.114682e-07	13.93046	ADD2	adducin 2
5	8000	-8.268313	3.700388	-94.76279	3.096472e-10	6.114682e-07	13.87285	PSCA	prostate stem cell antigen
6	83987	6.172927	2.571210	93.42002	3.355388e-10	6.114682e-07	13.79881	CCDC8	coiled-coil domain containing 8

Figure 5.4: Toptable obtained with limma for the mesenchymal (SUM-159) TNBC vs Trametinib (0.03μM 1 day) comparison.

	ENTREZID	logFC	AveExpr	t	P.Value	adj.P.Val	B	SYMBOL	GENENAME
1	3860	-10.997729	4.5576700	-149.93186	2.616388e-11	3.431264e-07	13.737584	KRT13	keratin 13
2	84985	-10.254139	4.5373487	-133.40268	5.031917e-11	3.431264e-07	13.428352	FAM83A	family with sequence similarity 83 member A
3	4070	-11.486800	5.0928647	-117.95309	1.002321e-10	4.477144e-07	13.309954	TACSTD2	tumor associated calcium signal transducer 2
4	2878	-10.096106	4.7983541	-112.39727	1.313138e-10	4.477144e-07	13.052002	GPX3	glutathione peroxidase 3
5	79977	-7.808241	2.9946500	-100.21404	2.495961e-10	5.721668e-07	12.927885	GRHL2	grainyhead like transcription factor 2
6	2877	-8.330937	3.2613995	-98.25961	2.786870e-10	5.721668e-07	12.915028	GPX2	glutathione peroxidase 2

Visualization of the results

Volcano plot To visualize the results we can use a volcanoPlot.

Figure 5.5: Volcano plot of the differentially expressed genes found by LimmaVoom, first comparison.

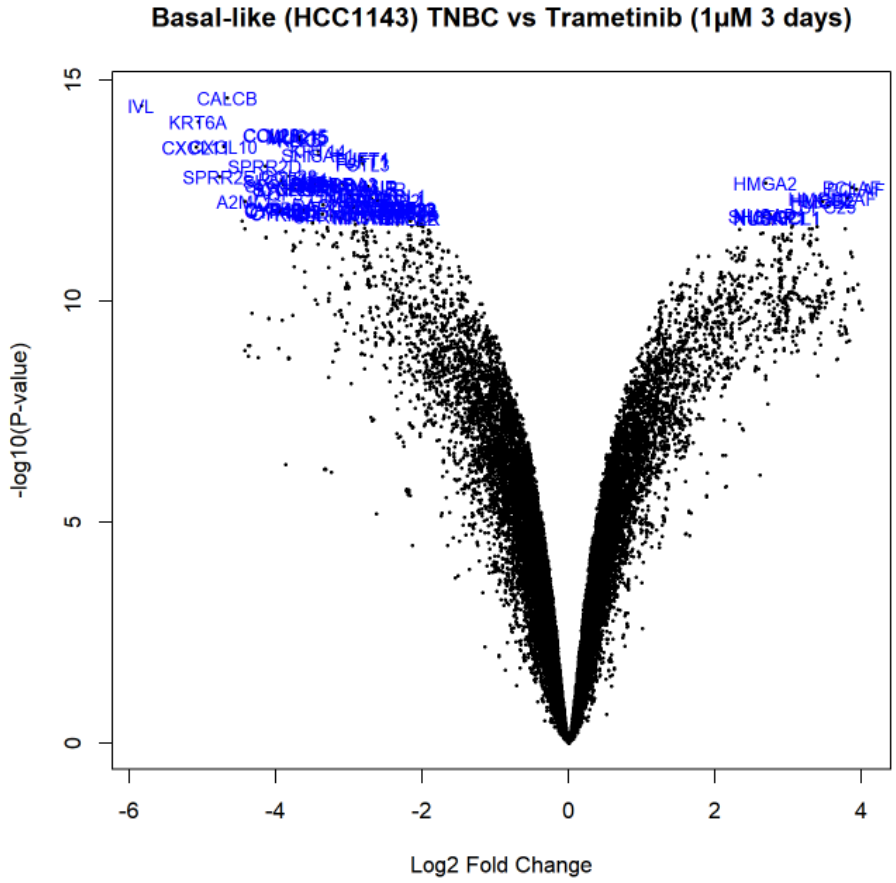


Figure 5.6: Volcano plot of the differentially expressed genes found by LimmaVoom, second comparison.

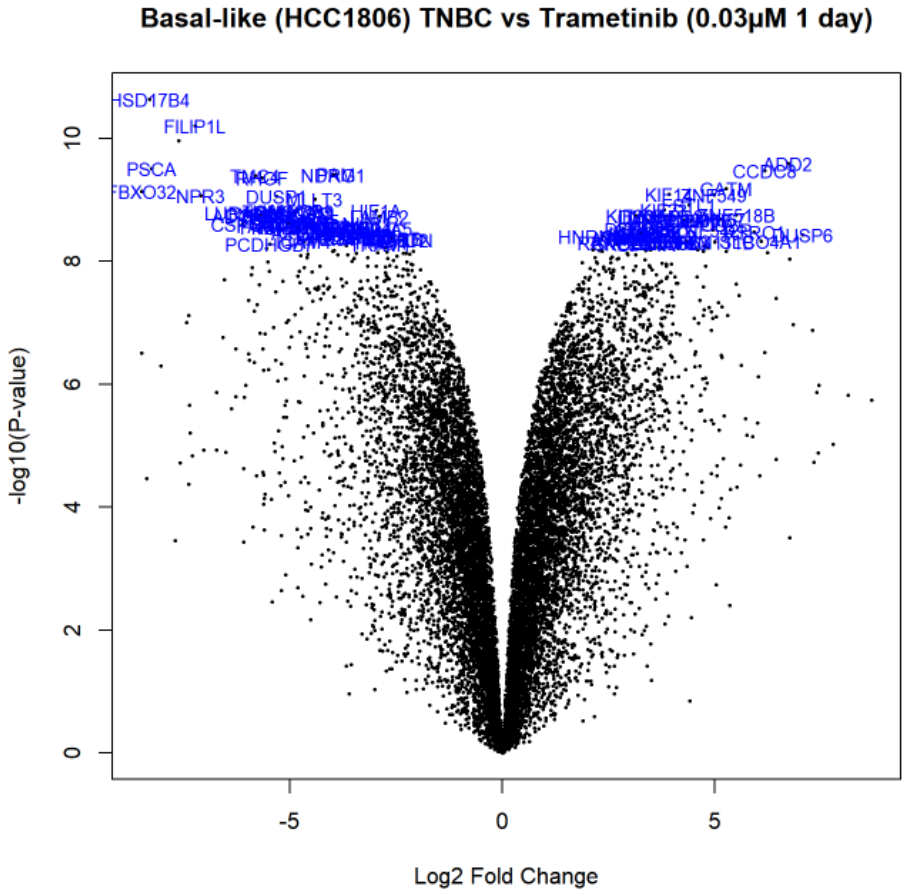
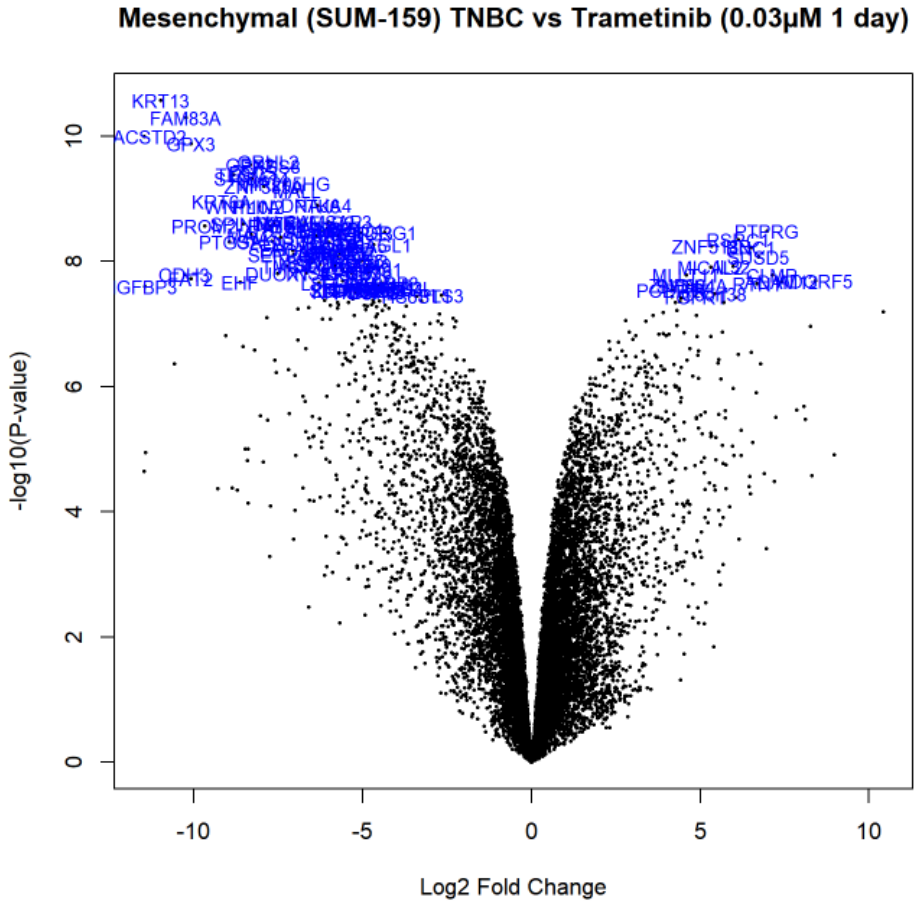


Figure 5.7: Volcano plot of the differentially expressed genes found by LimmaVoom, third comparison.



Heatmap In order to see if there are differentiated expression profiles, we can make a color map with the most differentially expressed genes.

That is, we set a gene selection criterion and retain those components of the results table that meet it. For example: Genes with an adjusted p-value less than 0.01 and a fold-change greater than 2.

With the expression matrix of the genes that verify this condition, a heatmap can be constructed.

Figure 5.8: Clustered heatmap of the differentially expressed genes found by LimmaVoom, first comparison.

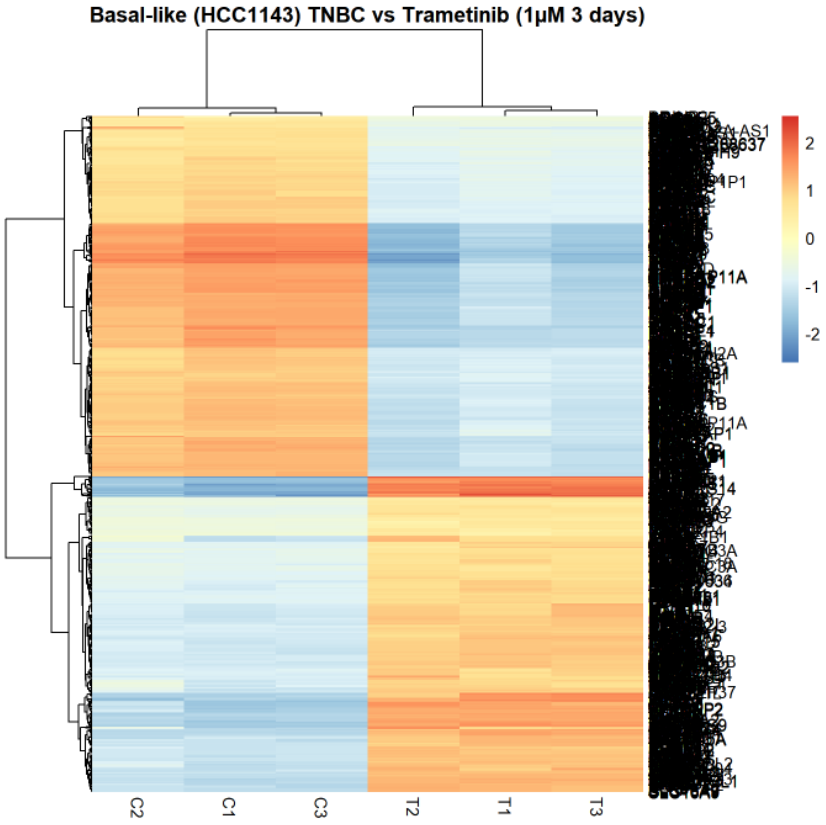


Figure 5.9: Clustered heatmap of the differentially expressed genes found by LimmaVoom, second comparison.

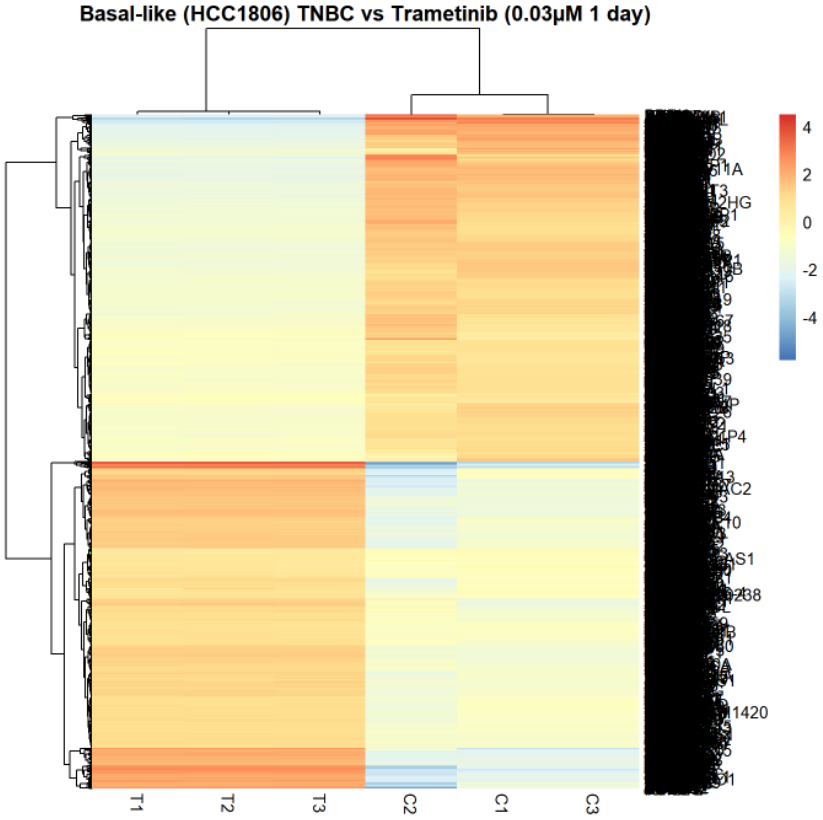
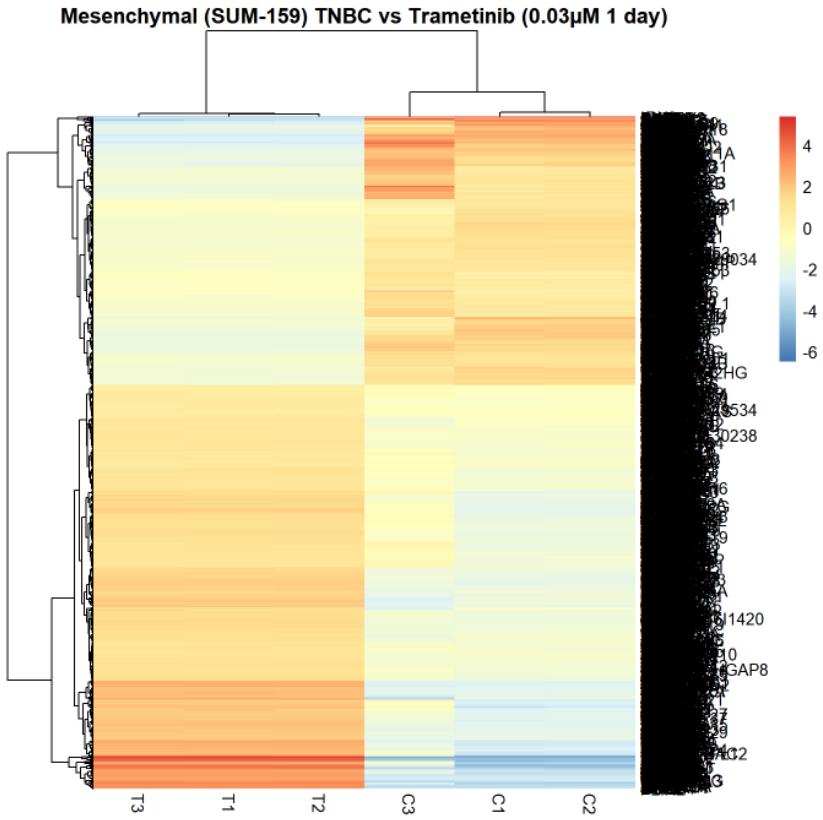


Figure 5.10: Clustered heatmap of the differentially expressed genes found by LimmaVoom, third comparison.



We can clearly see a group of differentially expressed genes that are over-expressed in the Trametinib groups (T1, T2, T3) and barely expressed in the control groups (C1, C2, C3). This behaviour is present in all three comparisons, as expected.

The relationship, however, seems stronger in the first comparison, that is, in the basal-like HCC1143 cell line treated with 1μM of Trametinib for 3 days.

5.1.2 Analysis using edgeR

The analysis with edgeR is similar to the previous one (originating from the same research team) but the modeling is different.

The analysis uses a GLM but, in a way reminiscent of limma, performs an additional step in which an improved estimate of the spread (variability) of the samples is computed. This integrates the individual and global estimates using empirical Bayes estimation.

For this purpose, which adds the improved dispersion estimators to the normalized counts, a generalized linear model with binomial distribution for the errors is fitted.

Once the model is adjusted, the contrast is constructed and the test is performed.

We can use the contrast matrix we built for limma-voom, in the same way we reused the design one.

The results are stored in a "topTable" similar to the one in limma.

We can select for the most differentially expressed genes in the same way that we did with limma-voom: genes with an adjusted p-value less than 0.01 and a fold-change greater than 2.

Figure 5.11: Toptable obtained with edgeR for the basal-like (HCC1143) TNBC vs Trametinib (1μM 3 days) comparison.

	ACCNUM	logFC	logCPM	F	PValue	FDR	SYMBOL	GENENAME
1	NM_005547	-6.366660	5.091765	3925.116	1.597356e-16	6.727582e-12	IVL	involucrin
2	NM_000728	-5.346822	3.745896	3244.823	5.011172e-16	7.256231e-12	CALCB	calcitonin related polypeptide beta
3	NM_005554	-5.069329	9.099168	3228.150	5.168624e-16	7.256231e-12	KRT6A	keratin 6A
4	NM_148672	-4.116374	6.381837	2911.167	9.613149e-16	7.716461e-12	CCL28	C-C motif chemokine ligand 28
5	NM_001282424	-4.510522	5.799973	2892.699	9.987484e-16	7.716461e-12	A2ML1	alpha-2-macroglobulin like 1
6	NM_001301875	-4.114819	6.316316	2826.214	1.148334e-15	7.716461e-12	CCL28	C-C motif chemokine ligand 28

Figure 5.12: Toptable obtained with edgeR for the basal-like (HCC1806) TNBC vs Trametinib (0.03μM 1 day) comparison.

	ENTREZID	logFC	logCPM	F	PValue	FDR	SYMBOL	GENENAME
1	3295	-9.520675	6.622712	9071.925	2.443370e-10	1.080543e-06	HSD17B4	hydroxysteroid 17-beta dehydrogenase 4
2	114907	-8.842718	8.158920	8578.786	2.864539e-10	1.080543e-06	FBXO32	F-box protein 32
3	8000	-9.178084	6.821702	7826.911	3.718721e-10	1.080543e-06	PSCA	prostate stem cell antigen
4	5066	-3.923695	7.961598	6275.037	6.973138e-10	1.080543e-06	PAM	peptidylglycine alpha-amidating monooxygenase
5	119	7.445952	5.846459	6158.891	7.353749e-10	1.080543e-06	ADD2	adducin 2
6	9928	3.967948	6.673136	5854.287	8.495130e-10	1.080543e-06	KIF14	kinesin family member 14

Figure 5.13: Toptable obtained with edgeR for the mesenchymal (SUM-159) TNBC vs Trametinib (0.03μM 1 day) comparison.

	ENTREZID	logFC	logCPM	F	PValue	FDR	SYMBOL	GENENAME
1	3860	-12.643350	9.058893	6394.7304	7.859976e-10	3.634585e-06	KRT13	keratin 13
2	84985	-11.351663	8.667669	5989.8252	9.450108e-10	3.634585e-06	FAM83A	family with sequence similarity 83 member A
3	4070	-12.645970	9.837603	5904.2854	9.840819e-10	3.634585e-06	TACSTD2	tumor associated calcium signal transducer 2
4	2878	-10.880399	8.849650	5738.9897	1.066017e-09	3.634585e-06	GPX3	glutathione peroxidase 3
5	3853	-9.236828	10.968594	4047.1692	2.850404e-09	5.193027e-06	KRT6A	keratin 6A
6	25946	-8.380808	7.309198	4025.2837	2.894253e-09	5.193027e-06	ZNF385A	zinc finger protein 385A

We observe that, on the first comparison, the CALCB, CCL28, IVL and KRT6A genes were also detected by limma.

On the second comparison, the ADD2, HSD17B4 and PSCA genes are also present in the limma toptable.

The same happens for the FAM83A, GPX3, KRT13 and TACSTD2 genes.

Finally, we observe that the KRT6A gen is present in both the first and third comparisons, in both limma and edgeR, which tells us that this is a very relevant gene.

5.1.3 Limma-voom and edgeR results comparison

We note that the list of selected genes is very similar in the three comparisons for the two methods used.

Figure 5.14: Comparison of genes found by LimmaVoom and EdgeR. Basal-like (HCC1143) TNBC vs Trametinib (1μM 3 days).

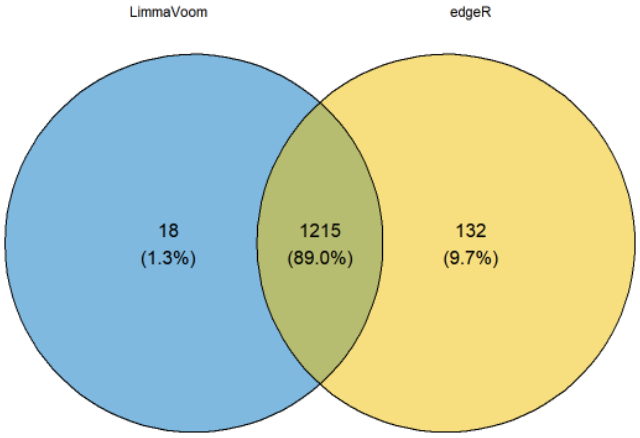


Figure 5.15: Comparison of genes found by LimmaVoom and EdgeR. Basal-like (HCC1806) TNBC vs Trametinib (0.03μM 1 day).

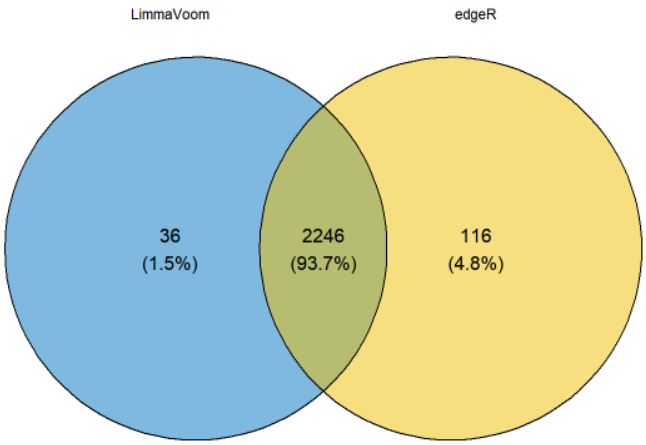
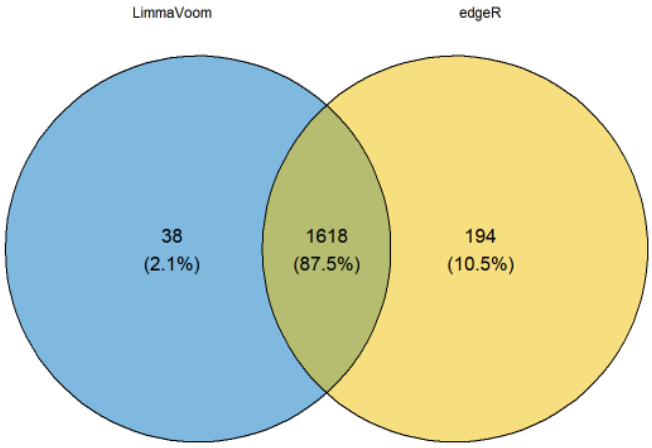


Figure 5.16: Comparison of genes found by LimmaVoom and EdgeR. Mesenchymal (SUM-159) TNBC vs Trametinib (0.03μM 1 day).



5.2 Results annotation

Two lists of transcripts are used for significance analysis:

1. The list of differentially expressed transcripts
2. The list of all transcripts or "Universe"

An obvious alternative to define the list of differentially expressed transcripts would be to take the genes that have been selected by both methods, ie by edgeR and limma-Voom, which is a restrictive but reliable selection.

As this type of analysis is very exploratory, a less restrictive option will be used, the union of both lists, that is, the genes that have been selected by one or the other of the methods.

An important detail in RNA-seq studies is that the expression units are usually transcripts, not genes. In practice, this determines that most enrichment analysis programs can lose detail, because their use requires having the identifiers in "gene" format, usually ACCNUM/ENTREZ or SYMBOL.

This is possible, and in fact easy to do, using the annotate package.

5.3 Enrichment analysis

Enrichment analysis is a technique used to identify gene or protein classes that are over-represented in a large group of genes or proteins that may be linked to disease phenotypes. The method uses statistical techniques to discover gene groupings that are significantly enriched or deficient.

The clusterProfiler package supports ENSEMBL-type identifiers and allows a wide variety of complementary enrichment analyses, making it one of the best options for biological significance analysis.

In the first place, an enrichment analysis is carried out with categories of the "Biological Process" ontology that allows selecting those categories that are more enriched in the list of differentially expressed genes.

The resulting object of the analysis contains information about the enriched categories, the degree of enrichment, or the genes annotated in them.

With the results of the enrichment analysis, different visualizations can be carried out, the exact interpretation of which can be seen in the clusterProfiler manual.

5.3.1 Dotplot

A dotplot displays some categories using a size code for the number of genes in the category and a color code for significance.

Figure 5.17: Dotplot of the first comparison.

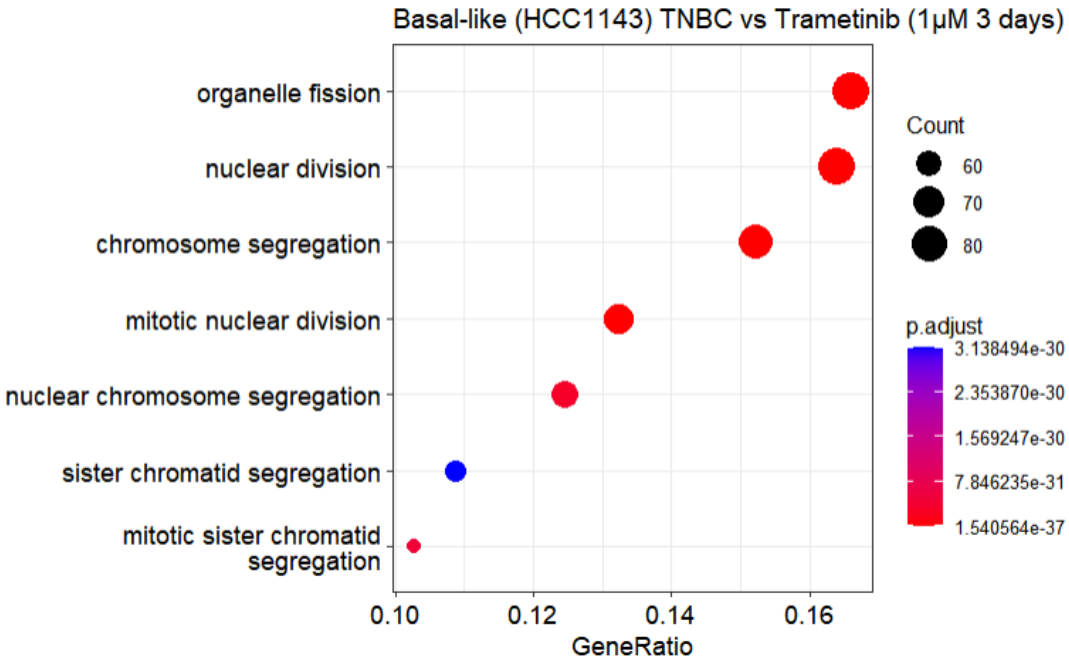


Figure 5.18: Dotplot of the second comparison.

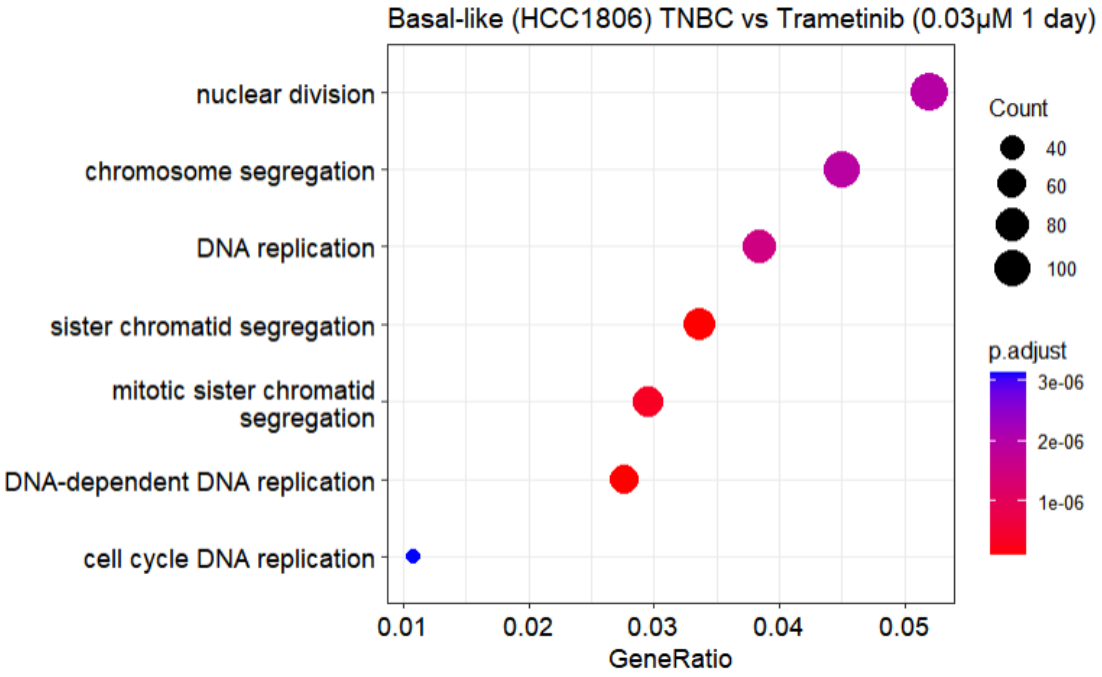
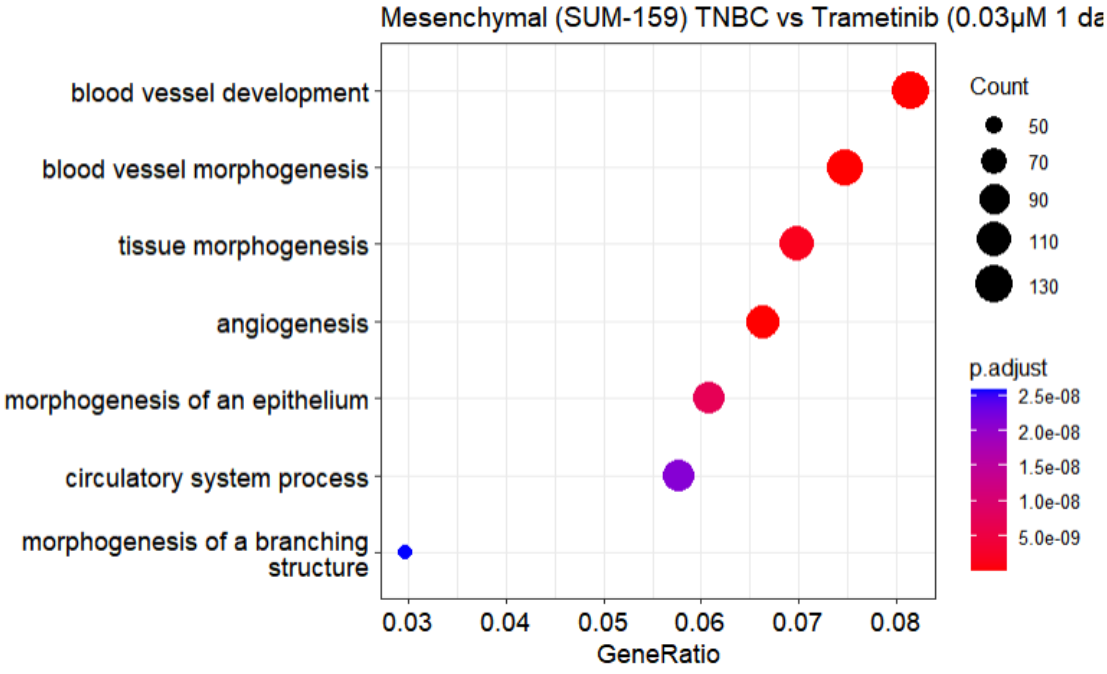


Figure 5.19: Dotplot of the third comparison.



According to these plots, the first and second comparisons are the ones for which the enrichment results are the most similar. The differentially expressed genes in these two groups are linked

to nuclear division, chromosome segregation, sister chromatid segregation and mitotic sister chromatid segregation.

5.3.2 Cnetplot

A cnetplot shows the relationship between genes and enriched categories.

Figure 5.20: Cnetplot of the basal-like (HCC1143) TNBC vs Trametinib (1μM 3 days) comparison.

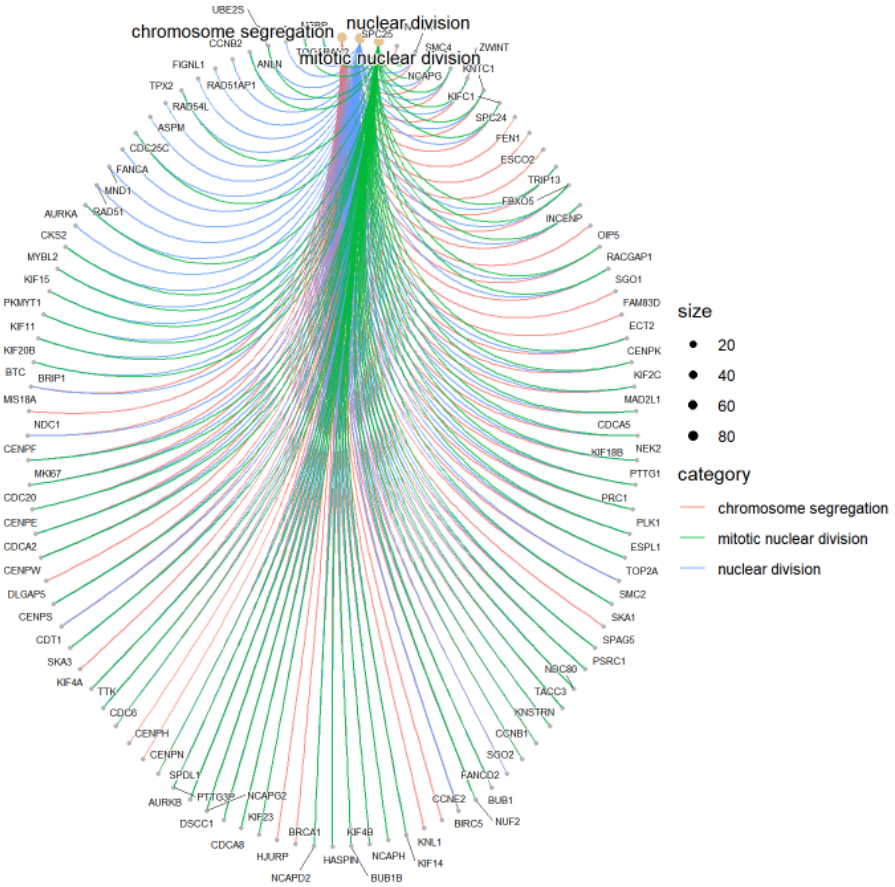


Figure 5.21: Cnetplot of the basal-like (HCC1806) TNBC vs Trametinib (0.03μM 1 day) comparison.

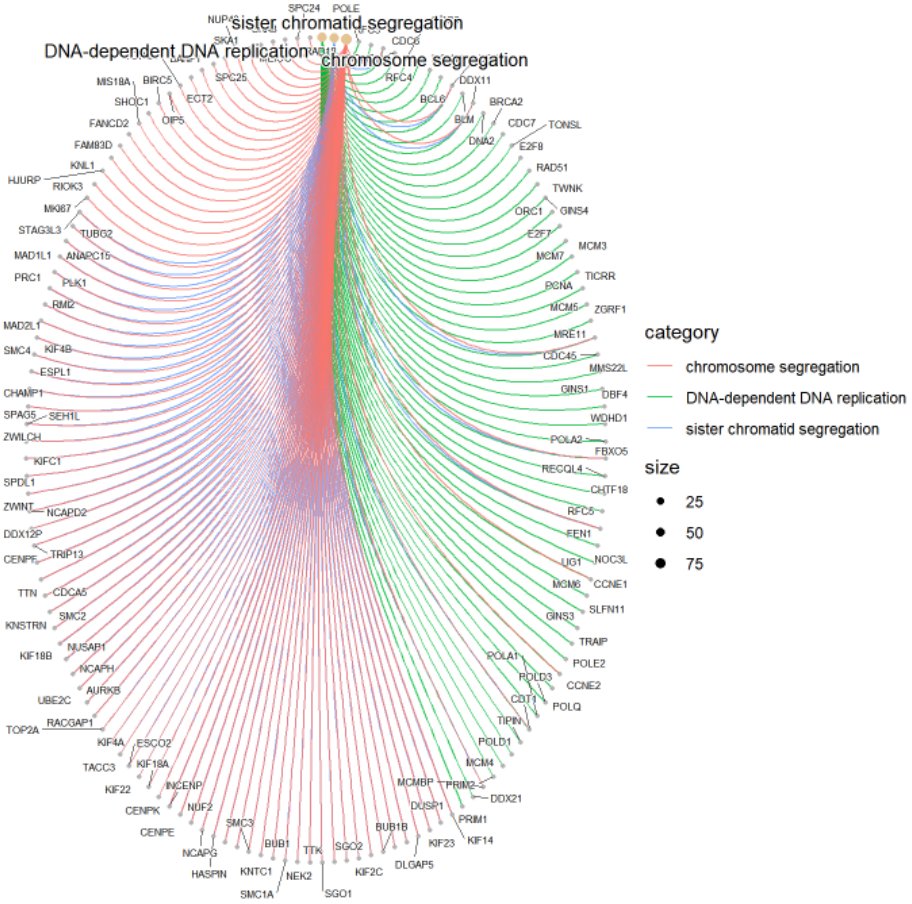
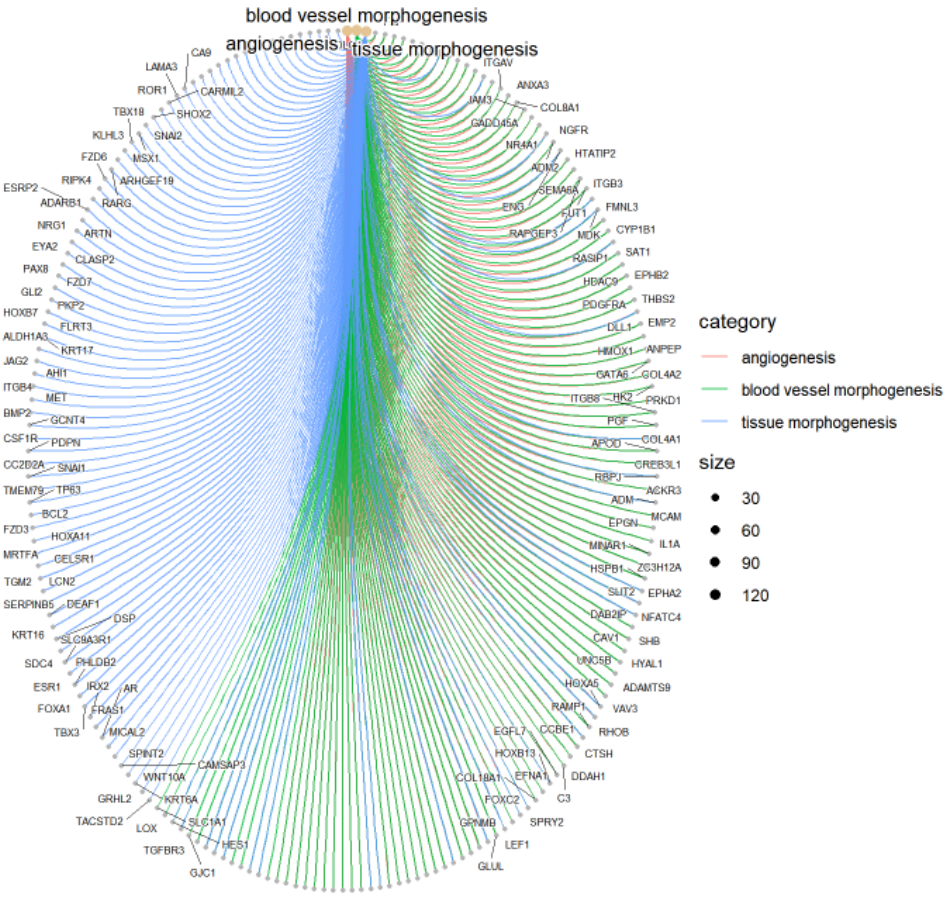


Figure 5.22: Cnetplot of the mesenchymal (SUM-159) TNBC vs Trametinib (0.03μM 1 day) comparison.



Genes related to chromosome segregation can be seen in both the first and second graph (red lines).

5.3.3 Goplot

A goplot displays, as a Gene Ontology subgraph, the categories that are related, as parents, to the enriched categories. It helps to see these in context.

The goplots obtained can be found in the Appendix.

5.3.4 Emapplot

Finally, given the abundance of components that appear in the analyses, we can try to group them by their similarity (closeness within the graph) and thus obtain a more compact visualization.

Figure 5.23: Emapplot of the of the basal-like (HCC1143) TNBC vs Trametinib (1μM 3 days) comparison.

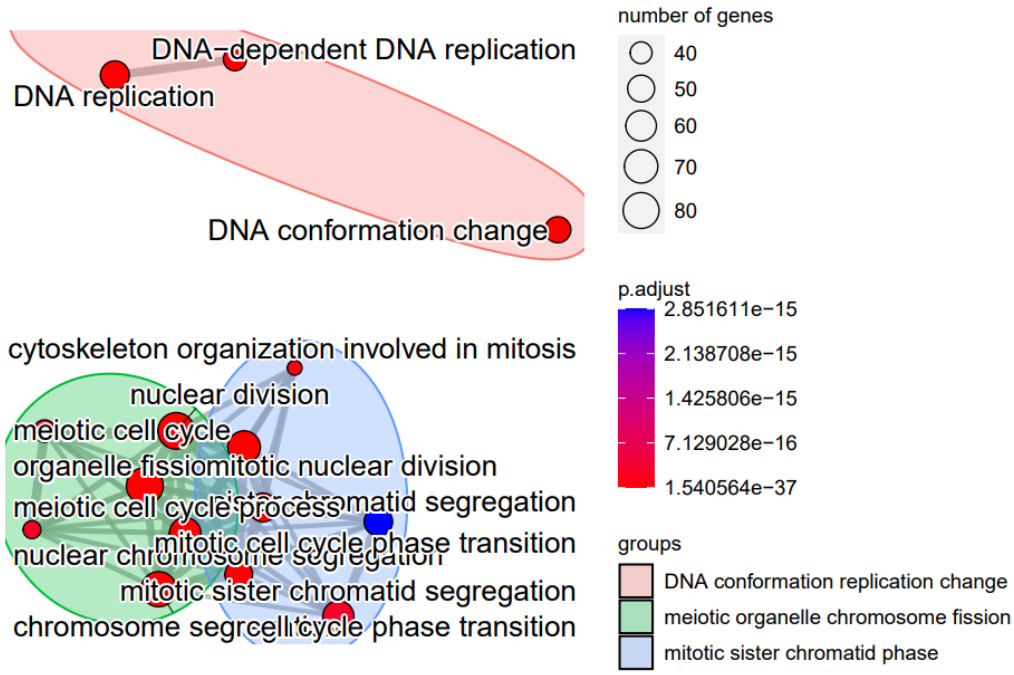


Figure 5.24: Emapplot of the basal-like (HCC1806) TNBC vs Trametinib (0.03μM 1 day) comparison.

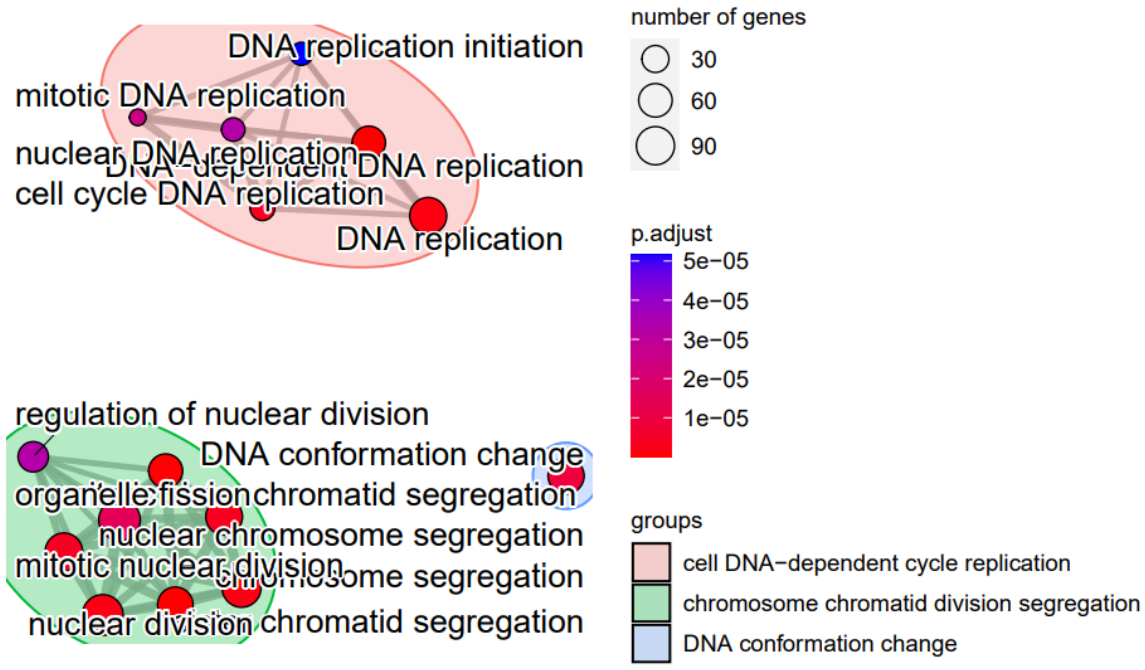
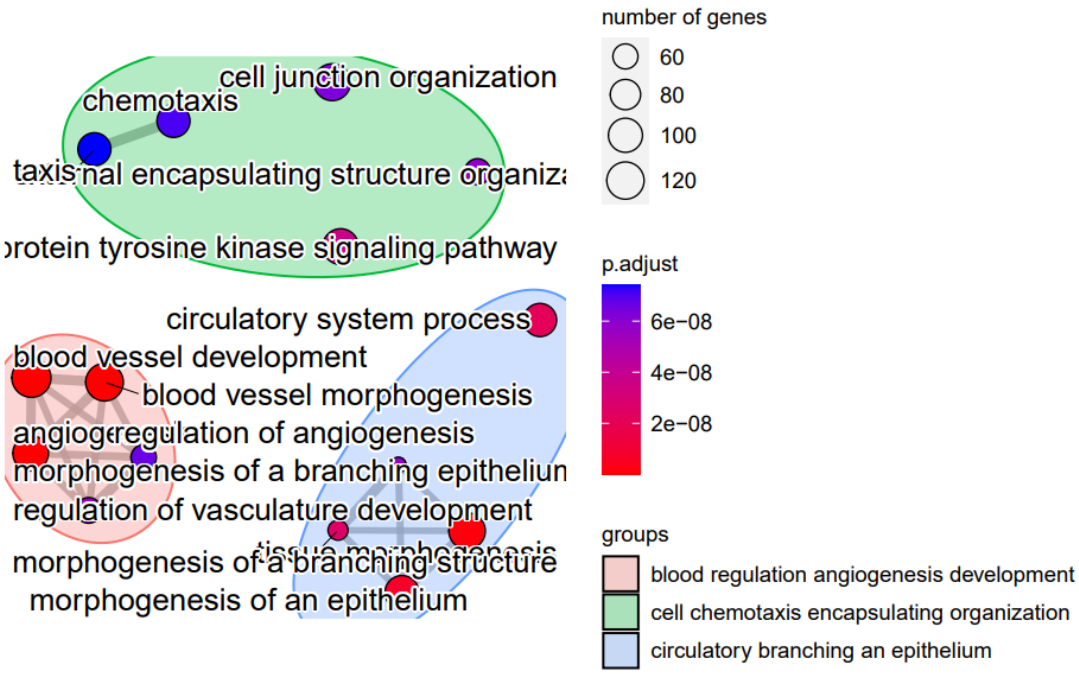


Figure 5.25: Emapplot of the mesenchymal (SUM-159) TNBC vs Trametinib (0.03μM 1 day) comparison.



Finally, in the first and second comparisons, we observe clusters of genes related to DNA replication, among others.

Chapter 6

Discussion

The main goals of this work were the study of breast cancer, of the MAPK pathway, of their relationship and to see the effects of breast cancer treatment with a MAPK inhibitor.

Through the development of an R pipeline for RNA-Seq data analysis, the study of two subtypes of triple negative breast cancer after Trametinib treatment has been possible.

Two GEO datasets have been analysed: GSE82032 and GSE138780. Both of them contained control and treatment samples, for basal-like and mesenchymal TNBC subtypes.

The differential expression analysis showed us that the response to Trametinib is different in the two different TNBC subtypes, as well as in the different cell lines.

Basal-like breast cancer showed different behaviours on the two cell lines analysed. In the HCC1143 cell line, the differentially expressed genes presented by both limma-voom and edgeR were CALCB, CCL28, IVL and KRT6A. However, in the HCC1806 cell line, the over or under expressed genes were ADD2, HSD17B4 and PSCA, among others.

These differences in expression might be due to different treatment regimes, as HCC1143 was treated with 1 µM for 3 days and HCC1806 was treated with 0.03 µM for 1 day. In addition, it is possible that the discrepancy is due to the fact that these are two different cell lines.

Mesenchymal breast cancer, cell line SUM-159, showed that the FAM83A, GPX3, KRT13, KRT6A and TACSTD2 genes were differentially expressed according to both methods, among others.

Moreover, it is worth noting the expressed genes are quite different among the three TNBC subtypes. The only gene that is differentially expressed in both basal-like and mesenchymal breast cancer after Trametinib treatment is KRT6A, keratin 6A.

The clustered heatmap shows us that the control and treatment samples are clearly defined, more so in the HCC1143 cell line treated with 1µM of Trametinib for 3 days.

The results obtained with limma-voom and edgeR are very similar, with the genes found in by

both methods being 89%, 93.7% and 87.5% similar for the first, second and third comparison, respectively.

The enrichment analysis results show that the differentially expressed genes in basal-like triple negative breast cancer are responsible for more similar processes. In both cases we find that these genes are linked to nuclear division, chromosome segregation, sister chromatid segregation and mitotic sister chromatid segregation. However, in the mesenchymal subtype, the most enriched categories are blood vessel development, blood vessel morphogenesis, tissue morphogenesis and angiogenesis, among others.

Finally, the emapplot of the biological significance analysis confirms that the differentially expressed genes in the HCC1143 and HCC1806 basal-like breat cancer treated or not with Trametinib are related to 1) DNA conformation replication changes and DNA conformation changes, 2) meiotic organelle chromosome fission and chromosome chromatid division and 3) mitotic sister chromatid phase and cell DNA-dependent cycle replication, respectively.

On the other hand, mesenchymal breast cancer shows a different behaviour to Trametinib treatment, with genes linked to blood regulation angiogenesis development, cell chemotaxis encapsulation organization and circulatory branching an epithelium.

These results suggest that the treatment of TNBC with an MAPK inhibitor drug may influence the proliferation rate of the cancer cells, among others, particularly in the basal-like subtype.

Therefore, we conclude that our hypothesis was correct: inhibiting the MAPK signalling pathway is indeed a promising strategy to treat breast cancer. Consequently, more in-depth studies are warranted.

Limitations The main limitation of the study is the small and diverse sample size, only 9 control samples and 9 treatment samples. Another limitation, linked to the previous one, is that the three comparisons are between two different treatment regimes, two different triple negative breast cancer subtypes and three different cell lines.

Because of all of this, it is not clear if the differences observed in the expressed genes among the three comparisons are due to the TNBC subtype, basal-like or mesenchymal, to the cell lines, HCC1143, HCC1806 and SUM-159, or to the treatment regime, 1µM of Trametinib for 3 days or 0.03 µM for 1 day.

Chapter 7

Conclusions

7.1 Conclusions

In this work, the conducted bibliographic research has provided an introduction to breast cancer, the MAPK signalling pathway and the implication of the later with carcinogenesis.

Afterwards, an overview of therapeutic targets within the MAPK pathway and current therapies have been researched. Subsequently, datasets of drugs targeting this pathway have been searched in GEO and selected.

Finally, an R pipeline for RNA-Seq data analysis has been developed to analyse two different regimes of Trametinib treatment on three different triple negative breast cancer cell lines.

The differential expression analysis using limma-voom and edgeR and the enrichment analyses comprise the results of this work.

Differentially expressed genes have been identified in both basal-like and mesenchymal subtypes, with KRT6A being the only gene differentially expressed in both subtypes.

The two basal-like cell lines analysed present differentially expressed genes linked to similar processes: nuclear division, chromosome segregation, sister chromatid segregation and mitotic sister chromatid segregation.

However, in the mesenchymal subtype, the most enriched categories are blood vessel development, blood vessel morphogenesis, tissue morphogenesis and angiogenesis, among others.

These findings may imply that treating TNBC with an MAPK inhibitor may affect the cancer cells' proliferation rate, among other things, notably in the basal-like subtype.

As a result, it is concluded that most likely the hypothesis is correct: blocking the MAPK signalling pathway is a promising technique for treating breast cancer. Therefore, more in-depth research is required.

7.2 Future lines

Future lines of this study would be to perform a cross-analysis of the TNBC breast cancer cell lines examined in this work with different Trametinib regimes.

Thus, analyzing the differentially expressed genes in basal-like (HCC1143) samples treated with 0.03 µM of Trametinib for 1 day, basal-like (HCC1806) samples treated with 1µM for 3 days and mesenchymal (SUM-159) samples treated with 1µM for 3 days.

Thanks to that, a limitation of this study would be overcome: knowing if the changes observed in the three groups are due to the treatment regime, the cell line or the TNBC breast cancer subtype.

Moreover, more samples need to be analysed in order to obtain meaningful results and to draw definitive conclusions.

Afterwards, it would be interesting to examine the effects of Trametinib in other breast cancer subtypes, such as luminal A, luminal B and HER+, to see if the results obtained are consistent and if Trametinib is indeed a good treatment option for those subgroups.

7.3 Planning follow-up

The work planning has been followed throughout the thesis and the used methodology has proven adequate. No major changes have been introduced to achieve success, as the tasks were planned appropriately and the objectives were realistic.

Chapter 8

Bibliography

[1] Cornelia Braicu, Mihail Buse, Constantin Busuioc, Rares Drula, Diana Gulei, Lajos Raduly, Alexandru Rusu, Alexandru Irimie, Atanas G. Atanasov, Ondrej Slaby, Calin Ionescu, and Ioana Berindan-Neagoe. A comprehensive review on mapk: A promising therapeutic target in cancer. *Cancers*, 11:1618, 10 2019.

[2] Scitable by Nature Education. *Cell Signaling — Learn Science at Scitable.* https://www.nature.com/scitable/topicpage/cell-signaling-14047077/, 2014.

[3] Cancer.net. *Breast Cancer - Metastatic: Introduction — Cancer.Net.* https://www.cancer.net/cancer-types/breast-cancer-metastatic/introduction, 2021.

[4] Chemocare. *Trametinib — Chemotherapy Drug Information — Chemocare.com.* https://chemocare.com/chemotherapy/drug-info/trametinib.aspx, 2022.

[5] Mayo Clinic. *Breast cancer - Diagnosis and treatment.* https://www.mayoclinic.org/diseases-conditions/breast-cancer/diagnosis-treatment/drc-20352475, 2022.

[6] National Center for Biotechnology Information. *GEO Browser - GEO - NCBI.* 2022.

[7] Yan-Jun Guo, Wei-Wei Pan, Sheng-Bing Liu, Zhong-Fei Shen, Ying Xu, and Ling-Ling Hu. Erk/mapk signalling pathway and tumorigenesis. *Experimental and Therapeutic Medicine*, 19:1997, 1 2020.

[8] Arjun Khunger, Monica Khunger, and Vamsidhar Velcheti. Dabrafenib in combination with trametinib in the treatment of patients with braf v600-positive advanced or metastatic non-small cell lung cancer: clinical evidence and experience. *Therapeutic Advances in Respiratory Disease*, 12, 3 2018.

[9] Wenken Liang, Chune Mo, Jianfen Wei, Wei Chen, Weiwei Gong, Jianling Shi, Xianliang Hou, Chunhong Li, Yecheng Deng, and Minglin Ou. Fam65a as a novel prognostic biomarker in human tumors reveal by a pan-cancer analysis. *Discover. Oncology*, 12:60, 12 2021.

[10] Jack McCain. The mapk (erk) pathway: Investigational combinations for the treatment of braf-mutated metastatic melanoma. *Pharmacy and Therapeutics*, 38:96, 2013.

[11] Johns Hopkins Medicine. *Breast Biopsy*. https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/breast-biopsy, 2022.

[12] Lorenzo Melchor and Javier Benítez. The complex genetic landscape of familial breast cancer. *Human Genetics*, 132:845–863, 8 2013.

[13] Vid Mlakar, Edouard Morel, Simona Jurkovic Mlakar, Marc Ansari, and Fabienne Gumy Pause. *A review of the biological and clinical implications of RAS/MAPK pathway alterations in neuroblastoma*, volume 40. https://jeccr.biomedcentral.com/articles/10.1186/s13046-021-01967-x, 12 2021.

[14] Renan Gomes Do Nascimento and Kaléu Mormino Otoni. Histological and molecular classification of breast cancer: what do we know? *Mastology*, 30:20200024, 2020.

[15] MD Douglas A. Nelson. *Breast Biopsy Procedure: Uses, Side Effects, Results*. https://www.verywellhealth.com/open-surgical-breast-biopsy-429949, 2022.

[16] Massachusetts Institute of Technology. *Using AI to predict breast cancer and personalize care*. https://news.mit.edu/2019/using-ai-predict-breast-cancer-and-personalize-care-0507, 2019.

[17] Gene Expression Omnibus. *GSE138780*. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13
2022.

[18] Gene Expression Omnibus. *GSE82032*. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE820
2022.

[19] World Health Organization. *Breast cancer now most common form of cancer: WHO taking action*. https://www.who.int/news/item/03-02-2021-breast-cancer-now-most-common-form-of-cancer-who-taking-action, 2021.

[20] World Health Organization. *Latest global data on cancer burden and alcohol consumption – IARC*. https://www.iarc.who.int/infographics/latest-global-data-on-cancer-burden-and-alcohol-consumption/, 2021.

[21] Lesley Jia Wei Pua, Chun Wai Mai, Felicia Fei Lei Chung, Alan Soo Beng Khoo, Chee Onn Leong, Wei Meng Lim, and Ling Wei Hii. Functional roles of jnk and p38 mapk signaling in nasopharyngeal carcinoma. *International Journal of Molecular Sciences 2022, Vol. 23, Page 1108*, 23:1108, 1 2022.

[22] Emad A. Rakha, Jorge S. Reis-Filho, Frederick Baehner, David J. Dabbs, Thomas Decker, Vincenzo Eusebi, Stephen B. Fox, Shu Ichihara, Jocelyne Jacquemier, Sunil R. Lakhani, José Palacios, Andrea L. Richardson, Stuart J. Schnitt, Fernando C. Schmitt, Puay Hoon Tan, Gary M. Tse, Sunil Badve, and Ian O. Ellis. Breast cancer prognostic classification in the molecular era: The role of histological grade. *Breast Cancer Research*, 12, 8 2010.

[23] National Health Service. *Breast cancer in women - Symptoms*. https://www.nhs.uk/conditions/breast-cancer/symptoms/, 2019.

[24] The American Cancer Society. *Breast Cancer Risk Factors You Can't Change*. https://www.cancer.org/cancer/breast-cancer/risk-and-prevention/breast-cancer-risk-factors-you-cannot-change.html, 2021.

[25] The American Cancer Society. *Can I Lower My Risk of Breast Cancer?* https://www.cancer.org/cancer/breast-cancer/risk-and-prevention/can-i-lower-my-risk.html, 2021.

[26] The American Cancer Society. *Lifestyle-related Breast Cancer Risk Factors*. https://www.cancer.org/cancer/breast-cancer/risk-and-prevention/lifestyle-related-breast-cancer-risk-factors.html, 2021.

[27] The American Cancer Society. *Stages of Breast Cancer — Understand Breast Cancer Staging*. https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/stages-of-breast-cancer.html, 2021.

[28] The American Cancer Society. *What Is a Breast Cancer's Grade? — Grading Breast Cancer*. https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/breast-cancer-grades.html, 2021.

[29] Hang Sun, Hong Li, Shuang Si, Shouliang Qi, Wei Zhang, He Ma, Siqi Liu, Li Yingxue, and Wei Qian. Performance evaluation of breast cancer diagnosis with mammography, ultrasonography and magnetic resonance imaging. *Journal of X-ray science and technology*, 26:805–813, 2018.

[30] Anne M. Wallace, Christopher Comstock, Carl K. Hoh, and David R. Vera. Breast imaging: A surgeon's prospective. *Nuclear Medicine and Biology*, 32:781–792, 2005.

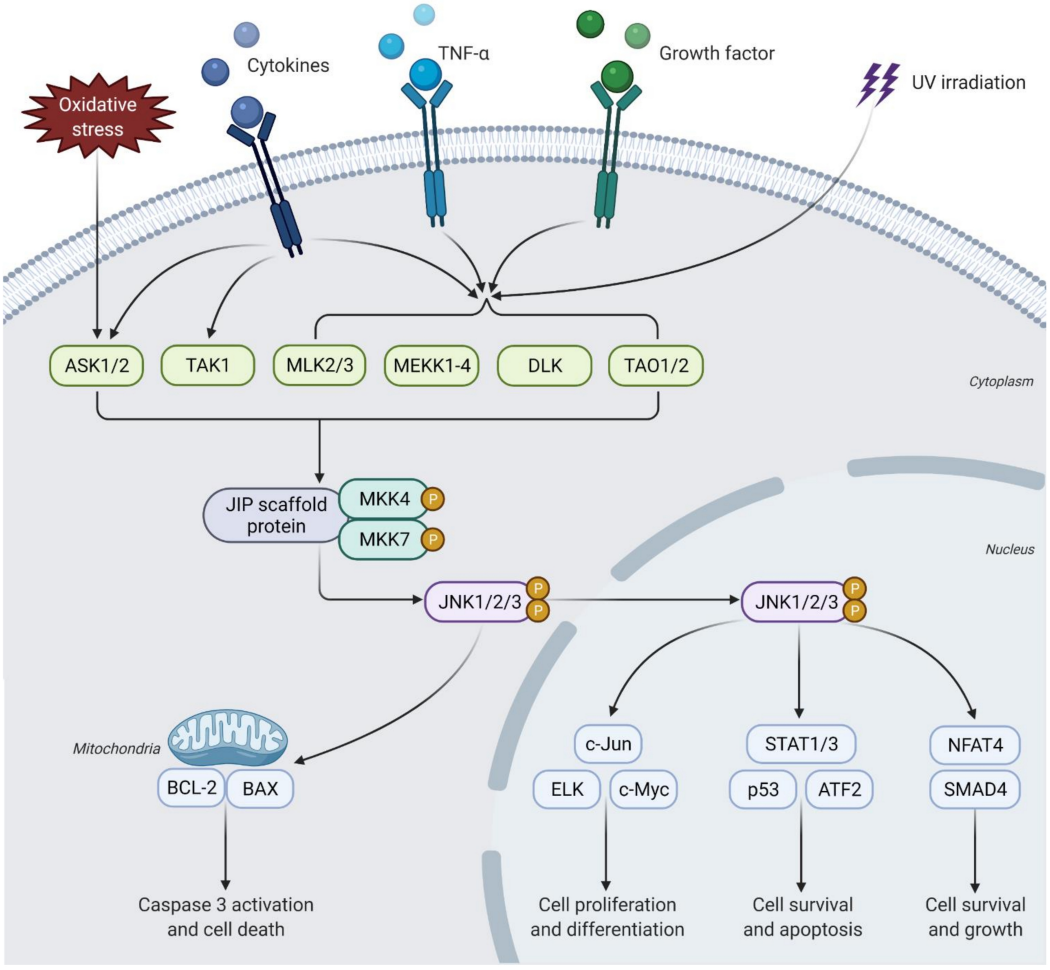
[31] Wikipedia. *Skeletal formula of Trametinib*. https://commons.wikimedia.org/wiki/File:Trametinib.svg, 2012.

Appendix A

MAPK signalling pathways

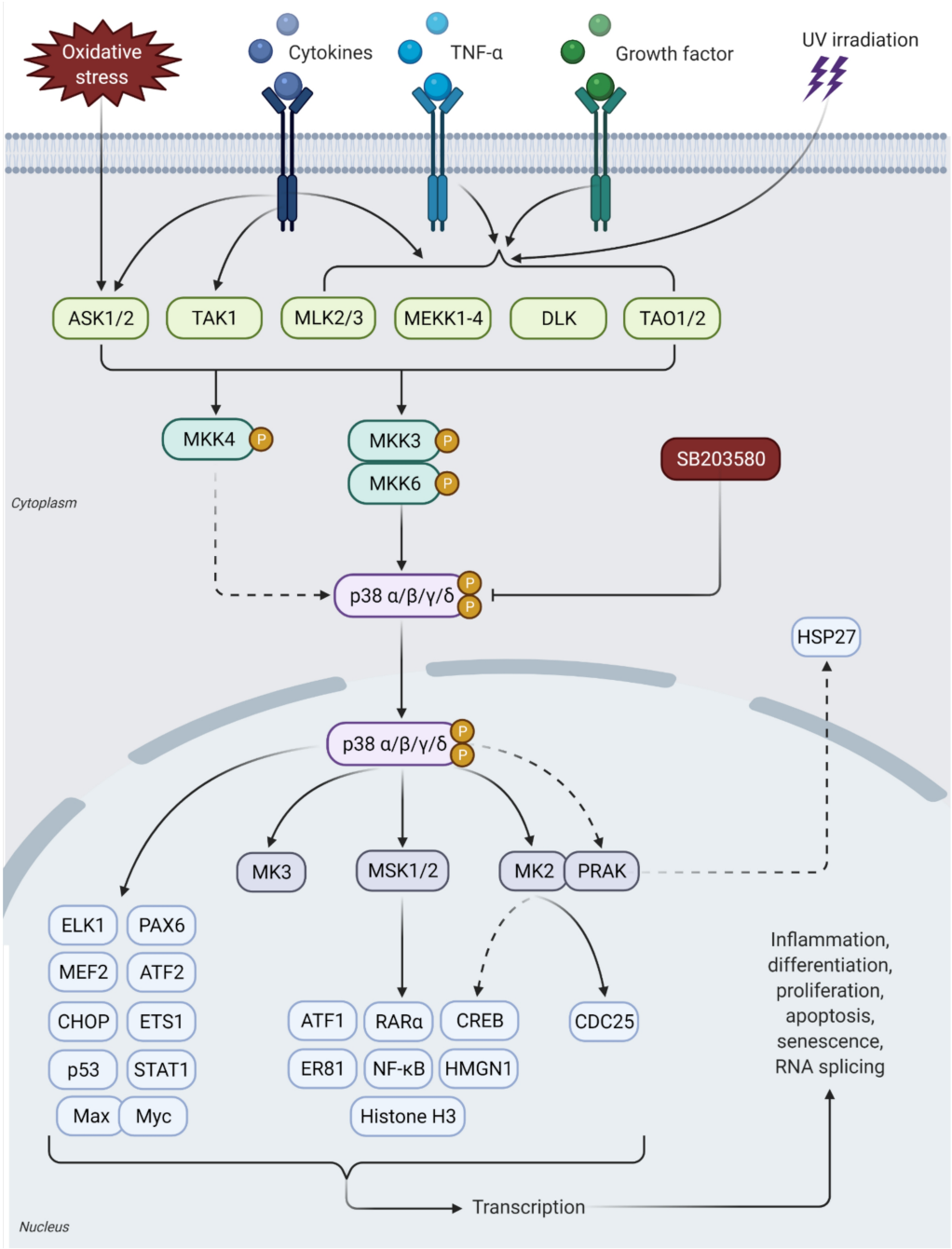
A.1 JNK pathway

Figure A.1: The upstream activators and downstream targets of the JNK pathway. [21]



A.2 p38 pathway

Figure A.2: p38 MAPKs pathway and its upstream and downstream activation. [21]



Appendix B

Data exploration

B.1 PlotMDS

Figure B.1: PlotMDS of the first comparison.

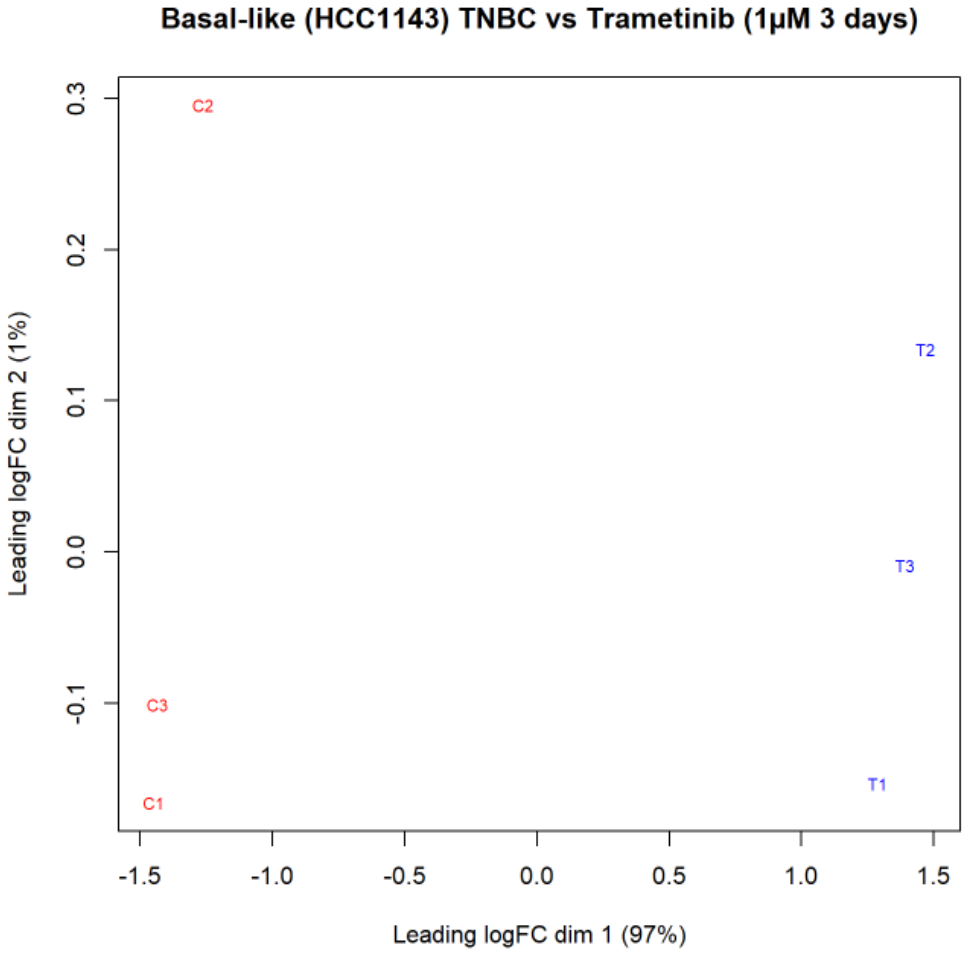


Figure B.2: PlotMDS of the second comparison.

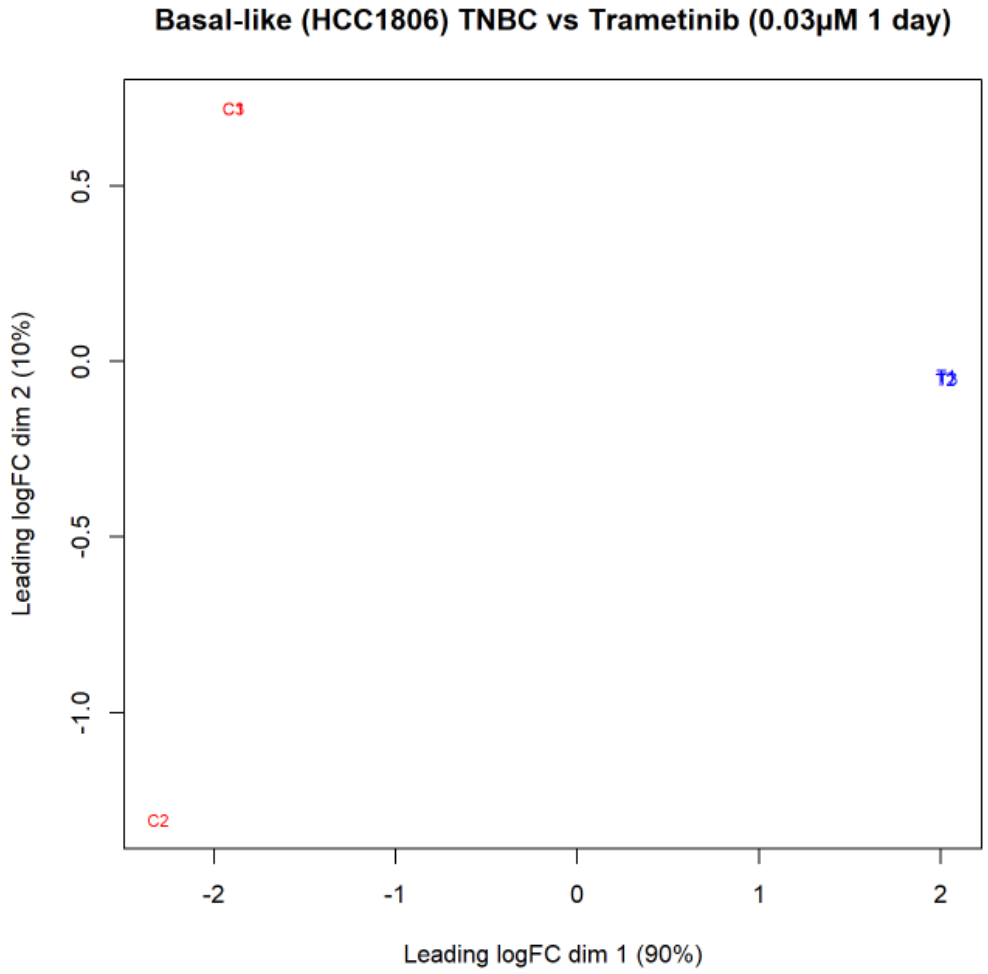
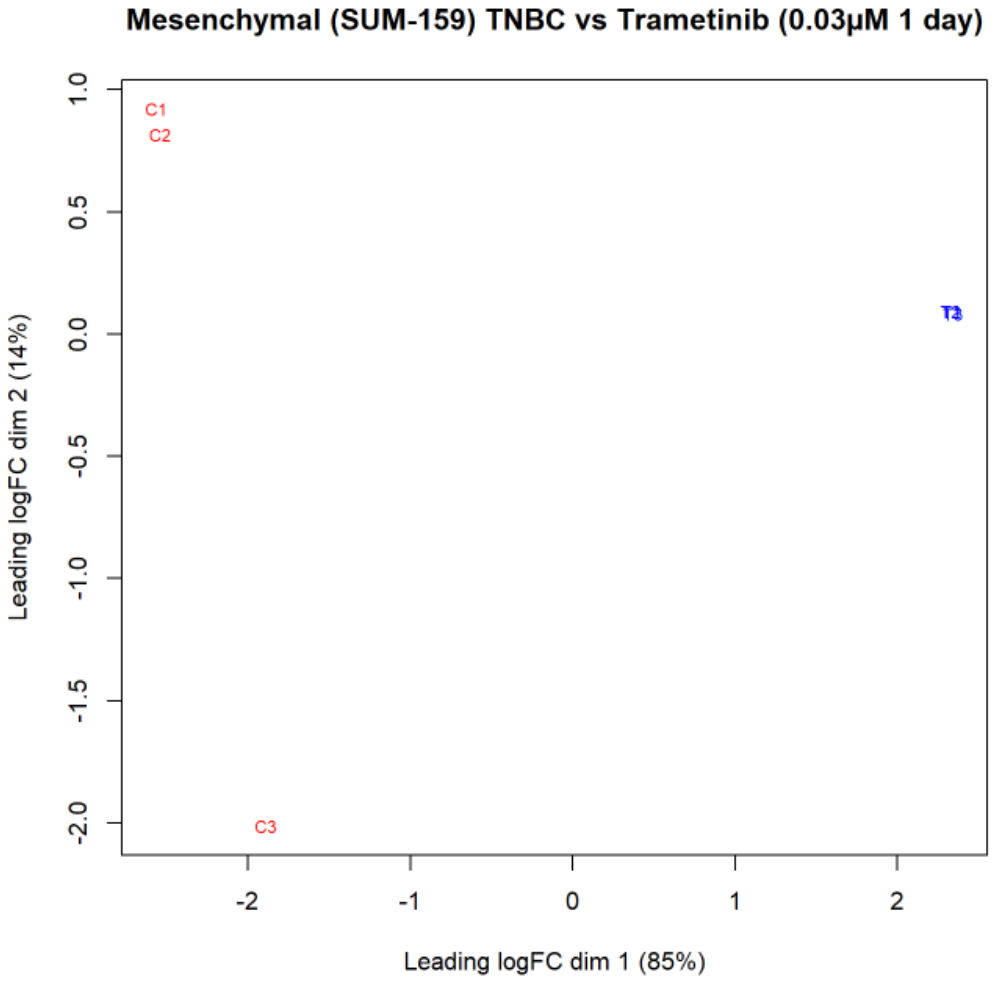


Figure B.3: PlotMDS of the third comparison.



Appendix C

Biological significance analysis

C.1 Goplot

Figure C.1: Goplot of the basal-like (HCC1143) TNBC vs Trametinib (1μM 3 days) comparison.

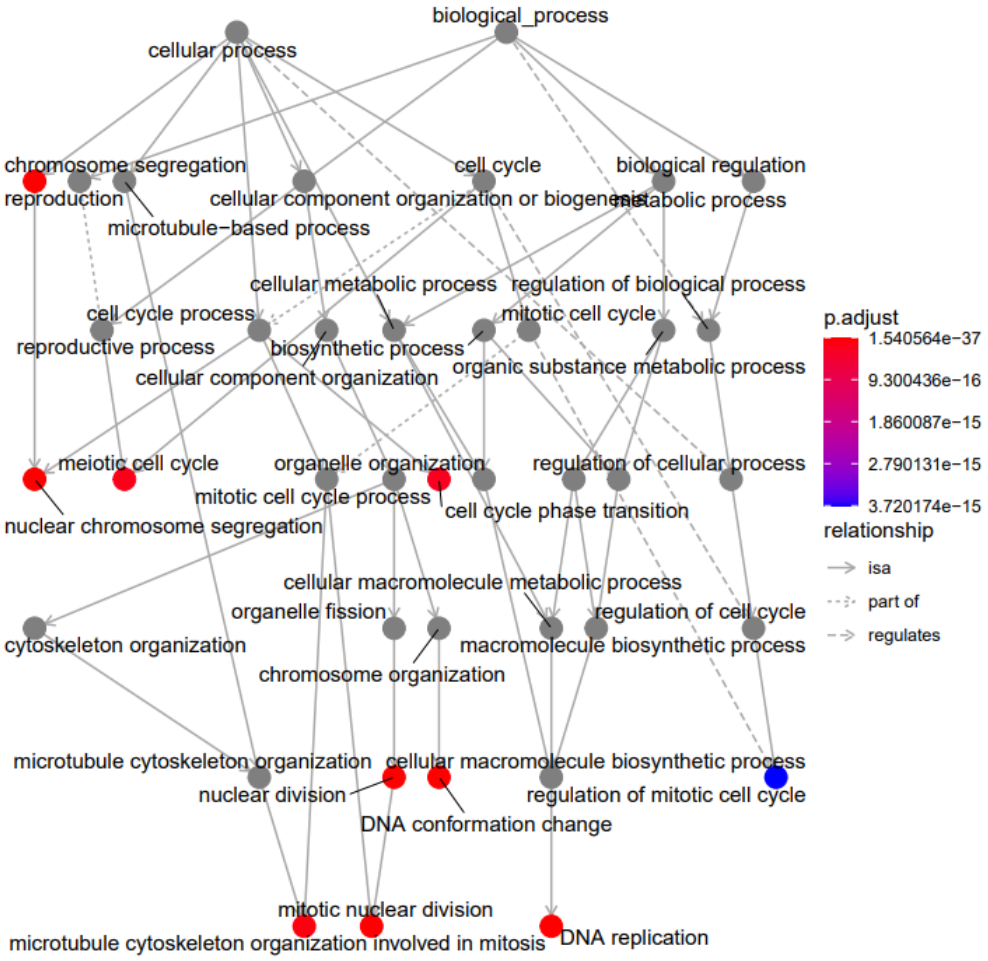


Figure C.2: Goplot of the basal-like (HCC1806) TNBC vs Trametinib (0.03μM 1 day) comparison.

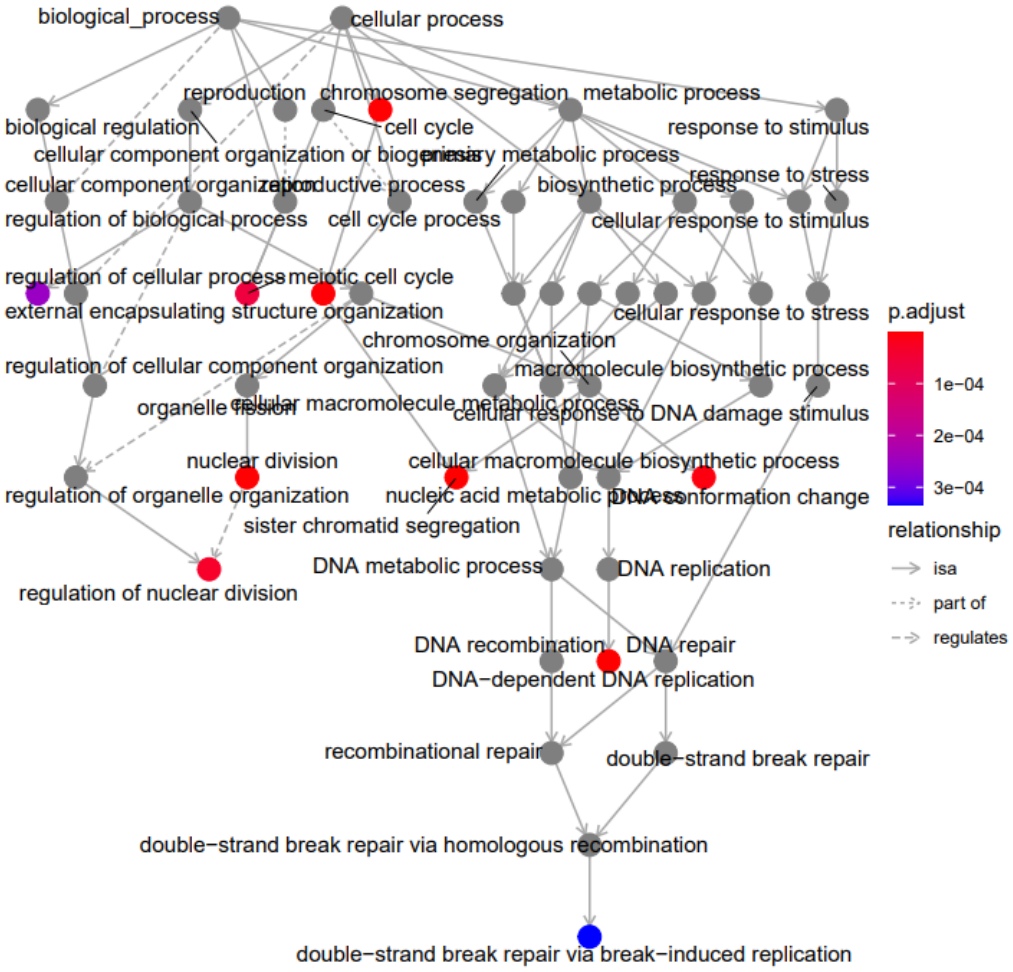
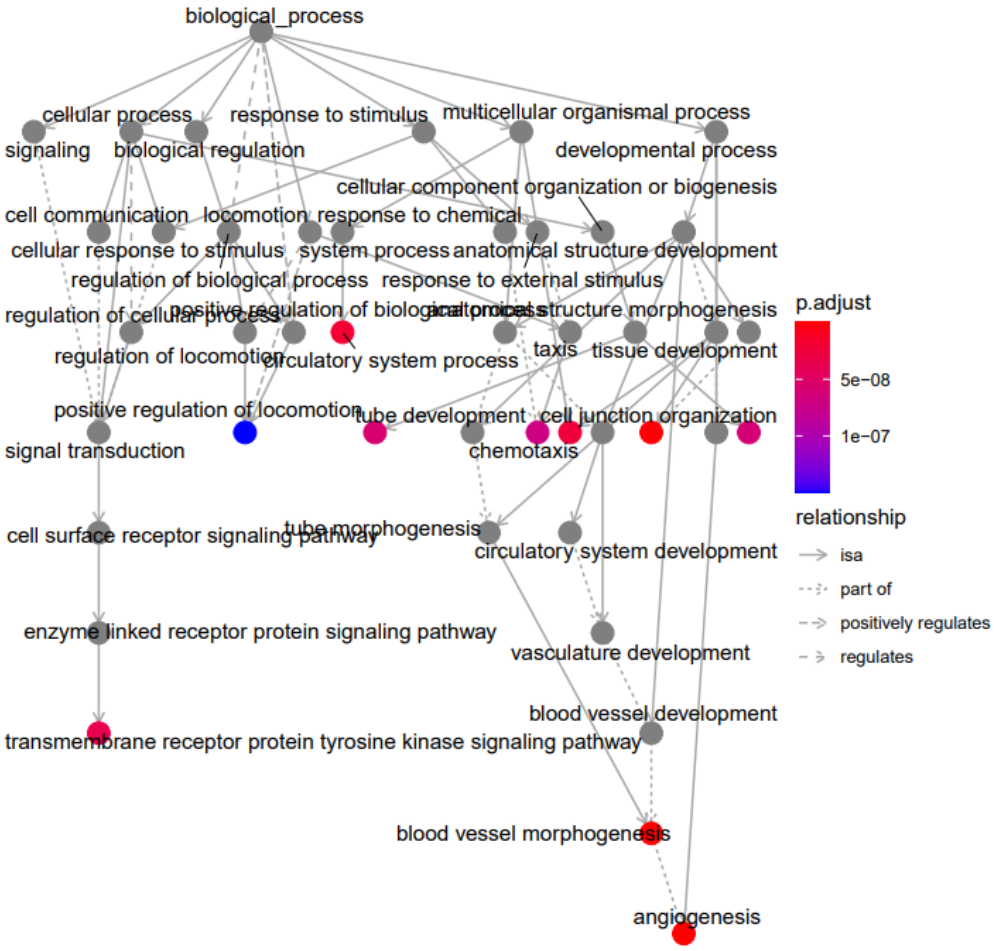


Figure C.3: Goplot of the mesenchymal (SUM-159) TNBC vs Trametinib (0.03μM 1 day) comparison.



Appendix D

R code

D.1 R code used for the RNA-Seq analysis

```
## ----setup, include=FALSE-----------------------------------------------------
library(knitr)
knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE,
                      comment = NA, prompt = TRUE, tidy = FALSE,
                      fig.width = 7, fig.height = 7, fig_caption = TRUE,
                      cache=FALSE)


## ----echo=TRUE, eval=FALSE, cache=TRUE----------------------------------------

if(!require(BiocManager)){
      install.packages("BiocManager", dep=TRUE)
}

installifnot <- function (pckgName, BioC=TRUE){
  if(BioC){
    if(!require(pckgName, character.only=TRUE)){
      BiocManager::install(pckgName)
    }
  }else{
    if(!require(pckgName, character.only=TRUE)){
      install.packages(pckgName, dep=TRUE)
    }
  }
}
installifnot("limma")
installifnot("edgeR")
# installifnot("DESeq2")
installifnot("org.Hs.eg.db")
installifnot("clusterProfiler")
installifnot("dplyr", BioC=FALSE)
installifnot("gplots", BioC=FALSE)
installifnot("ggvenn", BioC=FALSE)
installifnot("pheatmap")
installifnot("stringi", BioC=FALSE)
```

```
installifnot("prettydoc", BioC=FALSE)
installifnot("ggnewscale", BioC=FALSE)


## ------------------------------------------------------------------------------------------
memory.limit(size=999999999)
gc()


## ------------------------------------------------------------------------------------------
if(!dir.exists("dades")) dir.create("dades")
if(!dir.exists("resultats")) dir.create("resultats")


## ------------------------------------------------------------------------------------------
counts_GSE82032 <- read.delim("dades/raw_counts_GSE82032.txt", header=TRUE)
rownames(counts_GSE82032) <- counts_GSE82032[,1]
counts_GSE82032 <- counts_GSE82032[2:7]
head(counts_GSE82032)


## ------------------------------------------------------------------------------------------
counts_GSE138780 <- read.delim("dades/raw_counts_GSE138780.txt", header=TRUE)
rownames(counts_GSE138780) <- counts_GSE138780[,1]
counts_GSE138780 <- counts_GSE138780[2:7]
head(counts_GSE138780)


## ------------------------------------------------------------------------------------------
counts_GSE138780_SUM159 <- read.delim("dades/raw_counts_GSE138780_SUM159.txt",
                                      header=TRUE)
rownames(counts_GSE138780_SUM159) <- counts_GSE138780_SUM159[,1]
counts_GSE138780_SUM159 <- counts_GSE138780_SUM159[2:7]
head(counts_GSE138780_SUM159)


## ------------------------------------------------------------------------------------------
mostres_GSE82032<- colnames(counts_GSE82032)
grups_GSE82032 <- c(rep("Control", 3), rep("Trametinib", 3))
colors_GSE82032=c(rep("red", 3), rep("blue", 3))
targets_GSE82032 <- data.frame(sample=mostres_GSE82032, group=grups_GSE82032,
                               cols=colors_GSE82032)
rownames(targets_GSE82032) <- targets_GSE82032[,1]
head(targets_GSE82032)


## ----creaTargets-------------------------------------------------------------------------
mostres_GSE138780<- colnames(counts_GSE138780)
grups_GSE138780 <- c(rep("Control", 3), rep("Trametinib", 3))
colors_GSE138780=c(rep("red", 3), rep("blue", 3))
targets_GSE138780 <- data.frame(sample=mostres_GSE138780, group=grups_GSE138780,
                                cols=colors_GSE138780)
rownames(targets_GSE138780) <- targets_GSE138780[,1]
```

```
head(targets_GSE138780)


## ------------------------------------------------------------------------------------------
mostres_GSE138780_SUM159<- colnames(counts_GSE138780_SUM159)
grups_GSE138780_SUM159 <- c(rep("Control", 3), rep("Trametinib", 3))
colors_GSE138780_SUM159=c(rep("red", 3), rep("blue", 3))
targets_GSE138780_SUM159 <- data.frame(sample=mostres_GSE138780_SUM159,
                                       group=grups_GSE138780_SUM159,
                                       cols=colors_GSE138780_SUM159)
rownames(targets_GSE138780_SUM159) <- targets_GSE138780_SUM159[,1]
head(targets_GSE138780_SUM159)


## ------------------------------------------------------------------------------------------
if (sum(colnames(counts_GSE82032)!=rownames(targets_GSE82032))>0)
  cat("Verifique que las filas del objeto targets coinciden con las columnas
       de la matriz de datos")


## ----sincronizacion------------------------------------------------------------------------
if (sum(colnames(counts_GSE138780)!=rownames(targets_GSE138780))>0)
  cat("Verifique que las filas del objeto targets coinciden con las columnas
      de la matriz de datos")


## ------------------------------------------------------------------------------------------
if (sum(colnames(counts_GSE138780_SUM159)!=
        rownames(targets_GSE138780_SUM159))>0)
  cat("Verifique que las filas del objeto targets coinciden con las columnas
      de la matriz de datos")


## ------------------------------------------------------------------------------------------
library(edgeR)
counts.CPM_GSE82032 <- cpm(counts_GSE82032)
head(counts.CPM_GSE82032)


## ----getCPM--------------------------------------------------------------------------------
library(edgeR)
counts.CPM_GSE138780 <- cpm(counts_GSE138780)
head(counts.CPM_GSE138780)


## ------------------------------------------------------------------------------------------
library(edgeR)
counts.CPM_GSE138780_SUM159 <- cpm(counts_GSE138780_SUM159)
head(counts.CPM_GSE138780_SUM159)


## ------------------------------------------------------------------------------------------
thresh_GSE82032 <- counts.CPM_GSE82032 > 0
```

```
keep_GSE82032 <- (rowSums(thresh_GSE82032[,1:3]) >= 3) &
        (rowSums(thresh_GSE82032[,4:6]) >= 3)
counts.keep_GSE82032 <- counts.CPM_GSE82032[keep_GSE82032,]
dim(counts.CPM_GSE82032)
dim(counts.keep_GSE82032)


## ----subsetThreshold--------------------------------------------------------------------------
thresh_GSE138780 <- counts.CPM_GSE138780 > 0
keep_GSE138780 <- (rowSums(thresh_GSE138780[,1:3]) >= 3) &
        (rowSums(thresh_GSE138780[,4:6]) >= 3)
counts.keep_GSE138780 <- counts.CPM_GSE138780[keep_GSE138780,]
dim(counts.CPM_GSE138780)
dim(counts.keep_GSE138780)


## ---------------------------------------------------------------------------------------------
thresh_GSE138780_SUM159 <- counts.CPM_GSE138780_SUM159 > 0
keep_GSE138780_SUM159 <- (rowSums(thresh_GSE138780_SUM159[,1:3]) >= 3) &
        (rowSums(thresh_GSE138780_SUM159[,4:6]) >= 3)
counts.keep_GSE138780_SUM159 <-
  counts.CPM_GSE138780_SUM159[keep_GSE138780_SUM159,]
dim(counts.CPM_GSE138780_SUM159)
dim(counts.keep_GSE138780_SUM159)


## ----comparaMatrices--------------------------------------------------------------------------
head(counts.CPM_GSE82032)
head(counts.keep_GSE82032)


## ---------------------------------------------------------------------------------------------
head(counts.CPM_GSE138780)
head(counts.keep_GSE138780)


## ---------------------------------------------------------------------------------------------
head(counts.CPM_GSE138780_SUM159)
head(counts.keep_GSE138780_SUM159)


## ----makeDGEObj0------------------------------------------------------------------------------
dgeObj_GSE82032 <- DGEList(counts = counts.keep_GSE82032,
                  lib.size = colSums(counts.keep_GSE82032),
                  norm.factors = rep(1,ncol(counts.keep_GSE82032)),
                  samples = targets_GSE82032,
                  group = targets_GSE82032$group,
                  # genes = rownames(counts.keep_GSE82032),
                  remove.zeros = FALSE)
dgeObj_GSE82032


## ---------------------------------------------------------------------------------------------
```

```
dgeObj_GSE138780 <- DGEList(counts = counts.keep_GSE138780,
                  lib.size = colSums(counts.keep_GSE138780),
                  norm.factors = rep(1,ncol(counts.keep_GSE138780)),
                  samples = targets_GSE138780,
                  group = targets_GSE138780$group,
                  # genes = rownames(counts.keep_GSE138780),
                  remove.zeros = FALSE)
dgeObj_GSE138780


## ------------------------------------------------------------------------------------
dgeObj_GSE138780_SUM159 <- DGEList(counts = counts.keep_GSE138780_SUM159,
                  lib.size = colSums(counts.keep_GSE138780_SUM159),
                  norm.factors = rep(1,ncol(counts.keep_GSE138780_SUM159)),
                  samples = targets_GSE138780_SUM159,
                  group = targets_GSE138780_SUM159$group,
                  # genes = rownames(counts.keep_GSE138780),
                  remove.zeros = FALSE)
dgeObj_GSE138780_SUM159


## ------------------------------------------------------------------------------------
library(edgeR)
dgeObj_norm_GSE82032 <- calcNormFactors(dgeObj_GSE82032)


## ------------------------------------------------------------------------------------
library(edgeR)
dgeObj_norm_GSE138780 <- calcNormFactors(dgeObj_GSE138780)


## ------------------------------------------------------------------------------------
library(edgeR)
dgeObj_norm_GSE138780_SUM159 <- calcNormFactors(dgeObj_GSE138780_SUM159)


## ------------------------------------------------------------------------------------
dgeObj_GSE82032$counts[1:3, 1:6]
dgeObj_norm_GSE82032$counts[1:3, 1:6]
dgeObj_GSE82032$samples$norm.factors
dgeObj_norm_GSE82032$samples$norm.factors


## ------------------------------------------------------------------------------------
dgeObj_GSE138780$counts[1:3, 1:6]
dgeObj_norm_GSE138780$counts[1:3, 1:6]
dgeObj_GSE138780$samples$norm.factors
dgeObj_norm_GSE138780$samples$norm.factors


## ------------------------------------------------------------------------------------
dgeObj_GSE138780_SUM159$counts[1:3, 1:6]
dgeObj_norm_GSE138780_SUM159$counts[1:3, 1:6]
```

```
dgeObj_GSE138780_SUM159$samples$norm.factors
dgeObj_norm_GSE138780_SUM159$samples$norm.factors


## ------------------------------------------------------------------------
log2count_norm_GSE82032 <- cpm(dgeObj_norm_GSE82032, log=TRUE)


## ------------------------------------------------------------------------
log2count_norm_GSE138780 <- cpm(dgeObj_norm_GSE138780, log=TRUE)


## ------------------------------------------------------------------------
log2count_norm_GSE138780_SUM159 <- cpm(dgeObj_norm_GSE138780_SUM159, log=TRUE)


## ------------------------------------------------------------------------
log2count_GSE82032 <- cpm(dgeObj_GSE82032, log=TRUE)

par(mfrow=c(2,1))
rawCounts_GSE82032 <- dgeObj_norm_GSE82032$counts
# boxplot(rawCounts_GSE82032, ylab="CPM",las=2, xlab="",
# col = dgeObj_GSE82032$samples$cols, cex.axis=0.7,
# main="DistribuciÃ³n de contajes GSE82032, cell line HCC1143")
# boxplot(log2count_norm_GSE82032, ylab="Log2-CPM",las=2, xlab="",
# col=dgeObj_GSE82032$samples$cols, cex.axis=0.7,
# main="DistribuciÃ³n de log(contajes) GSE82032, cell line HCC1143")
# abline(h=median(log2count_norm_GSE82032), col="blue")
# par(mfrow=c(1,1))


## ------------------------------------------------------------------------
par(mfrow=c(2,1))
rawCounts_GSE82032 <- dgeObj_norm_GSE82032$counts
boxplot(rawCounts_GSE82032, ylab="CPM",las=2, xlab="",
        col = dgeObj_GSE82032$samples$cols, cex.axis=0.7,
        main="Distribution of normalized counts \nBasal-like (HCC1143) TNBC vs
        Trametinib (1ÃµM 3 days)")
boxplot(log2count_norm_GSE82032, ylab="Log2-CPM",las=2, xlab="",
        col=dgeObj_GSE82032$samples$cols, cex.axis=0.7, main="Distribution of
        normalized log(counts) \nBasal-like (HCC1143) TNBC vs Trametinib
        (1ÃµM 3 days)")
abline(h=median(log2count_norm_GSE82032), col="blue")
par(mfrow=c(1,1))


## ------------------------------------------------------------------------
log2count_GSE138780 <- cpm(dgeObj_GSE138780, log=TRUE)

par(mfrow=c(2,1))
rawCounts_GSE138780 <- dgeObj_GSE138780$counts
# boxplot(rawCounts_GSE138780, ylab="CPM",las=2, xlab="",
# col = dgeObj_GSE138780$samples$cols, cex.axis=0.7,
```

```
# main="DistribuciÃ³n de contajes GSE138780, cell line HCC1806")
# boxplot(log2count_GSE138780, ylab="Log2-CPM",las=2, xlab="",
# col=dgeObj_GSE138780$samples$cols, cex.axis=0.7,
# main="DistribuciÃ³n de log(contajes) GSE138780, cell line HCC1806")
# abline(h=median(log2count_GSE138780), col="blue")
# par(mfrow=c(1,1))


## ------------------------------------------------------------------------------------------
par(mfrow=c(2,1))
rawCounts_GSE138780 <- dgeObj_norm_GSE138780$counts
boxplot(rawCounts_GSE138780, ylab="CPM",las=2, xlab="",
        col = dgeObj_GSE138780$samples$cols, cex.axis=0.7,
        main="Distribution of normalized counts \nBasal-like (HCC1806) TNBC vs
        Trametinib (0.03ÃµM 1 day)")
boxplot(log2count_norm_GSE138780, ylab="Log2-CPM",las=2, xlab="",
        col=dgeObj_GSE138780$samples$cols, cex.axis=0.7, main="Distribution of
        normalized log(counts) \nBasal-like (HCC1806) TNBC vs Trametinib
        (0.03ÃµM 1 day)")
abline(h=median(log2count_norm_GSE138780), col="blue")
par(mfrow=c(1,1))


## ------------------------------------------------------------------------------------------
log2count_GSE138780_SUM159 <- cpm(dgeObj_GSE138780_SUM159, log=TRUE)

par(mfrow=c(2,1))
rawCounts_GSE138780_SUM159 <- dgeObj_GSE138780_SUM159$counts
# boxplot(rawCounts_GSE138780_SUM159, ylab="CPM",las=2, xlab="",
# col = dgeObj_GSE138780_SUM159$samples$cols, cex.axis=0.7,
# main="DistribuciÃ³n de contajes GSE138780, cell line SUM-159")
# boxplot(log2count_GSE138780_SUM159, ylab="Log2-CPM",las=2, xlab="",
# col=dgeObj_GSE138780_SUM159$samples$cols, cex.axis=0.7,
# main="DistribuciÃ³n de log(contajes) GSE138780, cell line SUM-159")
# abline(h=median(log2count_GSE138780_SUM159), col="blue")
# par(mfrow=c(1,1))


## ------------------------------------------------------------------------------------------
par(mfrow=c(2,1))
rawCounts_GSE138780_SUM159 <- dgeObj_norm_GSE138780_SUM159$counts
boxplot(rawCounts_GSE138780_SUM159, ylab="CPM",las=2, xlab="",
        col = dgeObj_GSE138780_SUM159$samples$cols, cex.axis=0.7,
        main="Distribution of normalized counts \nMesenchymal (SUM-159) TNBC vs
        Trametinib (0.03ÃµM 1 day)")
boxplot(log2count_norm_GSE138780_SUM159, ylab="Log2-CPM",las=2, xlab="",
        col=dgeObj_GSE138780_SUM159$samples$cols, cex.axis=0.7,
        main="Distribution of normalized log(counts) \nMesenchymal (SUM-159)
        TNBC vs Trametinib (0.03ÃµM 1 day)")
abline(h=median(log2count_norm_GSE138780_SUM159), col="blue")
par(mfrow=c(1,1))
```

```
## ------------------------------------------------------------------------------------------
sampleDists_GSE82032 <- dist(t(log2count_norm_GSE82032))
round(sampleDists_GSE82032,1)


## ------------------------------------------------------------------------------------------
sampleDists_GSE138780 <- dist(t(log2count_norm_GSE138780))
round(sampleDists_GSE138780,1)


## ------------------------------------------------------------------------------------------
sampleDists_GSE138780_SUM159 <- dist(t(log2count_norm_GSE138780_SUM159))
round(sampleDists_GSE138780_SUM159,1)


## ------------------------------------------------------------------------------------------
library(factoextra)
fviz_dist(sampleDists_GSE82032)


## ------------------------------------------------------------------------------------------
library(factoextra)
fviz_dist(sampleDists_GSE138780)


## ------------------------------------------------------------------------------------------
library(factoextra)
fviz_dist(sampleDists_GSE138780_SUM159)


## ------------------------------------------------------------------------------------------
hc <-hclust(sampleDists_GSE82032)
plot(hc,labels = colnames(log2count_norm_GSE82032),
     main = "Hierarchical clustering \nBasal-like (HCC1143) TNBC vs Trametinib
     (1ÃµM 3 days)", cex=0.8)


## ------------------------------------------------------------------------------------------
hc <-hclust(sampleDists_GSE138780)
plot(hc,labels = colnames(log2count_norm_GSE138780),main = "Hierarchical
     clustering \nBasal-like (HCC1806) TNBC vs Trametinib (0.03ÃµM 1 day)",
     cex=0.8)


## ------------------------------------------------------------------------------------------
hc <-hclust(sampleDists_GSE138780_SUM159)
plot(hc,labels = colnames(log2count_norm_GSE138780_SUM159),main = "Hierarchical
     clustering \nMesenchymal (SUM-159) TNBC vs Trametinib (0.03ÃµM 1 day)",
     cex=0.8)


## ------------------------------------------------------------------------------------------
normFactors_GSE82032 <- dgeObj_norm_GSE82032$samples$norm.factors
```

```
names(normFactors_GSE82032)<- sampleNames_GSE82032 <-
  rownames(dgeObj_norm_GSE82032$samples)
plot(normFactors_GSE82032, main= "Factors de normalitzaciÃ³ GSE82032,
     cell line HCC1143")
text(normFactors_GSE82032, sampleNames_GSE82032, pos=2, cex=0.7)


## ------------------------------------------------------------------------------------------
normFactors_GSE138780 <- dgeObj_norm_GSE138780$samples$norm.factors
names(normFactors_GSE138780)<- sampleNames_GSE138780 <-
  rownames(dgeObj_norm_GSE138780$samples)
plot(normFactors_GSE138780, main= "Factors de normalitzaciÃ³ GSE138780,
     cell line HCC1806")
text(normFactors_GSE138780, sampleNames_GSE138780, pos=2, cex=0.7)


## ------------------------------------------------------------------------------------------
normFactors_GSE138780_SUM159 <-
  dgeObj_norm_GSE138780_SUM159$samples$norm.factors
names(normFactors_GSE138780_SUM159)<- sampleNames_GSE138780_SUM159 <-
  rownames(dgeObj_norm_GSE138780_SUM159$samples)
plot(normFactors_GSE138780_SUM159, main= "Factors de normalitzaciÃ³ GSE138780,
     cell line SUM-159")
text(normFactors_GSE138780_SUM159, sampleNames_GSE138780_SUM159, pos=2,
     cex=0.7)


## ------------------------------------------------------------------------------------------
#sampleinfo$Status <- factor (sampleinfo$group)
col.status_GSE82032 <- dgeObj_norm_GSE82032$samples$cols
plotMDS(log2count_norm_GSE82032,col=col.status_GSE82032, main="Basal-like
        (HCC1143) TNBC vs Trametinib (1ÃµM 3 days)", cex=0.7)


## ------------------------------------------------------------------------------------------
#sampleinfo$Status <- factor (sampleinfo$group)
col.status_GSE138780 <- dgeObj_norm_GSE138780$samples$cols
plotMDS(log2count_norm_GSE138780,col=col.status_GSE138780, main="Basal-like
        (HCC1806) TNBC vs Trametinib (0.03ÃµM 1 day)", cex=0.7)


## ------------------------------------------------------------------------------------------
#sampleinfo$Status <- factor (sampleinfo$group)
col.status_GSE138780_SUM159 <- dgeObj_norm_GSE138780_SUM159$samples$cols
plotMDS(log2count_norm_GSE138780_SUM159,col=col.status_GSE138780_SUM159,
        main="Mesenchymal (SUM-159) TNBC vs Trametinib (0.03ÃµM 1 day)",
        cex=0.7)


## ------------------------------------------------------------------------------------------
group_GSE82032 = as.factor(dgeObj_norm_GSE82032$samples$group)
design_GSE82032 = model.matrix(~ 0 + group_GSE82032)
colnames(design_GSE82032) = gsub("group_GSE82032", "",
```

```
                                  colnames(design_GSE82032))
row.names(design_GSE82032) = sampleNames_GSE82032
design_GSE82032


## ------------------------------------------------------------------------------------------
group_GSE138780 = as.factor(dgeObj_norm_GSE138780$samples$group)
design_GSE138780 = model.matrix(~ 0 + group_GSE138780)
colnames(design_GSE138780) = gsub("group_GSE138780", "",
                                  colnames(design_GSE138780))
row.names(design_GSE138780) = sampleNames_GSE138780
design_GSE138780


## ------------------------------------------------------------------------------------------
group_GSE138780_SUM159 = as.factor(dgeObj_norm_GSE138780_SUM159$samples$group)
design_GSE138780_SUM159 = model.matrix(~ 0 + group_GSE138780_SUM159)
colnames(design_GSE138780_SUM159) = gsub("group_GSE138780_SUM159", "",
                                         colnames(design_GSE138780_SUM159))
row.names(design_GSE138780_SUM159) = sampleNames_GSE138780_SUM159
design_GSE138780_SUM159


## ------------------------------------------------------------------------------------------
cont.matrix_GSE82032 = makeContrasts(
  ControlvsTrametinib_GSE82032 = Control - Trametinib,
levels=colnames(design_GSE82032))
cont.matrix_GSE82032


## ------------------------------------------------------------------------------------------
cont.matrix_GSE138780 = makeContrasts(
  ControlvsTrametinib_GSE138780 = Control - Trametinib,
levels=colnames(design_GSE138780))
cont.matrix_GSE138780


## ------------------------------------------------------------------------------------------
cont.matrix_GSE138780_SUM159 = makeContrasts(
  ControlvsTrametinib_GSE138780_SUM159 = Control - Trametinib,
levels=colnames(design_GSE138780_SUM159))
cont.matrix_GSE138780_SUM159


## ------------------------------------------------------------------------------------------
voomObj_GSE82032 <- voom(dgeObj_norm_GSE82032, design_GSE82032)
voomObj_GSE82032


## ------------------------------------------------------------------------------------------
voomObj_GSE138780 <- voom(dgeObj_norm_GSE138780, design_GSE138780)
voomObj_GSE138780
```

```
## ------------------------------------------------------------------------------------------------
voomObj_GSE138780_SUM159 <- voom(dgeObj_norm_GSE138780_SUM159,
                                 design_GSE138780_SUM159)
voomObj_GSE138780_SUM159


## ------------------------------------------------------------------------------------------------
fit_GSE82032 <- lmFit(voomObj_GSE82032)
fit.cont_GSE82032 <- contrasts.fit(fit_GSE82032, cont.matrix_GSE82032)
fit.cont_GSE82032 <- eBayes(fit.cont_GSE82032)


## ------------------------------------------------------------------------------------------------
fit_GSE138780 <- lmFit(voomObj_GSE138780)
fit.cont_GSE138780 <- contrasts.fit(fit_GSE138780, cont.matrix_GSE138780)
fit.cont_GSE138780 <- eBayes(fit.cont_GSE138780)


## ------------------------------------------------------------------------------------------------
fit_GSE138780_SUM159 <- lmFit(voomObj_GSE138780_SUM159)
fit.cont_GSE138780_SUM159 <- contrasts.fit(fit_GSE138780_SUM159,
                                           cont.matrix_GSE138780_SUM159)
fit.cont_GSE138780_SUM159 <- eBayes(fit.cont_GSE138780_SUM159)


## ------------------------------------------------------------------------------------------------
library(org.Hs.eg.db)
library(dplyr)

toptab_GSE82032 <- topTable(fit.cont_GSE82032,coef=1,sort.by="p",
                            number=nrow(fit.cont_GSE82032))
geneIds <- AnnotationDbi::select(org.Hs.eg.db, keys = rownames(toptab_GSE82032),
                                 keytype = "ACCNUM", columns =
                                   c("SYMBOL", "GENENAME"))

anotaTopTab <- function(toptab) {
    topTab_Anot <- toptab %>%
        mutate(ACCNUM = rownames(toptab)) %>%
        merge(geneIds, "ACCNUM") %>%
        arrange(P.Value)
    return(topTab_Anot)
}
toptab_anot_GSE82032 <- anotaTopTab(toptab_GSE82032)
head(toptab_anot_GSE82032)


## ------------------------------------------------------------------------------------------------
toptab_GSE138780 <- topTable(fit.cont_GSE138780,coef=1,sort.by="p",
                             number=nrow(fit.cont_GSE138780))

anotaTopTab <- function(toptab) {
    topTab_Anot <- toptab %>%
        mutate(ENTREZID = rownames(toptab)) %>%
```

```
        merge(geneIds, "ENTREZID") %>%
        arrange(P.Value)
    return(topTab_Anot)
}

geneIds <- AnnotationDbi::select(org.Hs.eg.db,
                                 keys = rownames(toptab_GSE138780),
                                 keytype = "ENTREZID", columns =
                                   c("SYMBOL", "GENENAME"))
toptab_anot_GSE138780 <- anotaTopTab(toptab_GSE138780)
head(toptab_anot_GSE138780)


## ----------------------------------------------------------------------------
toptab_GSE138780_SUM159 <- topTable(fit.cont_GSE138780_SUM159,coef=1,
                                    sort.by="p",
                                    number=nrow(fit.cont_GSE138780_SUM159))

geneIds <- AnnotationDbi::select(org.Hs.eg.db,
                                 keys = rownames(toptab_GSE138780_SUM159),
                                 keytype = "ENTREZID", columns =
                                   c("SYMBOL", "GENENAME"))
toptab_anot_GSE138780_SUM159 <- anotaTopTab(toptab_GSE138780_SUM159)
head(toptab_anot_GSE138780_SUM159)


## ----------------------------------------------------------------------------
write.csv(toptab_anot_GSE82032,
          file="resultats/topTable_limmaVoom_GSE82032.csv")


## ----------------------------------------------------------------------------
write.csv(toptab_anot_GSE138780,
          file="resultats/topTable_limmaVoom_GSE138780.csv")


## ----------------------------------------------------------------------------
write.csv(toptab_anot_GSE138780_SUM159,
          file="resultats/topTable_limmaVoom_GSE138780_SUM159.csv")


## ----------------------------------------------------------------------------
library(org.Hs.eg.db)
genenames_GSE82032 <- AnnotationDbi::select(org.Hs.eg.db,
                                            rownames(fit.cont_GSE82032),
                                            keytype = "ACCNUM",
                                            c("SYMBOL"))$SYMBOL

volcanoplot(fit.cont_GSE82032,coef=1,highlight=100, main="Basal-like (HCC1143)
            TNBC vs Trametinib (1ÃµM 3 days)", names = genenames_GSE82032)


## ----------------------------------------------------------------------------
```

```
genenames_GSE138780 <- AnnotationDbi::select(org.Hs.eg.db,
                                             rownames(fit.cont_GSE138780),
                                             keytype = "ENTREZID",
                                             c("SYMBOL"))$SYMBOL

volcanoplot(fit.cont_GSE138780,coef=1,highlight=100,
            main="Basal-like (HCC1806) TNBC vs Trametinib (0.03ÃµM 1 day)",
            names = genenames_GSE138780)


## ------------------------------------------------------------------------------------
genenames_GSE138780_SUM159 <- AnnotationDbi::select(org.Hs.eg.db,
                                    rownames(fit.cont_GSE138780_SUM159),
                                    keytype = "ENTREZID", c("SYMBOL"))$SYMBOL

volcanoplot(fit.cont_GSE138780_SUM159,coef=1,highlight=100, main="Mesenchymal
            (SUM-159) TNBC vs Trametinib (0.03ÃµM 1 day)",
            names = genenames_GSE138780_SUM159)


## ------------------------------------------------------------------------------------
topGenesBas_GSE82032 <- rownames(subset(toptab_GSE82032, (abs(logFC)> 2) &
                                          (adj.P.Val < 0.01)))
length(topGenesBas_GSE82032)


## ------------------------------------------------------------------------------------
topGenesBas_GSE138780 <- rownames(subset(toptab_GSE138780, (abs(logFC)> 2) &
                                           (adj.P.Val < 0.01)))
length(topGenesBas_GSE138780)


## ------------------------------------------------------------------------------------
topGenesBas_GSE138780_SUM159 <- rownames(subset(toptab_GSE138780_SUM159,
                                        (abs(logFC)> 2) & (adj.P.Val < 0.01)))
length(topGenesBas_GSE138780_SUM159)


## ------------------------------------------------------------------------------------
library(pheatmap)
mat_GSE82032  <- log2count_norm_GSE82032[topGenesBas_GSE82032, ]
mat_GSE82032  <- mat_GSE82032 - rowMeans(mat_GSE82032)
rownames(mat_GSE82032) <- AnnotationDbi::select(org.Hs.eg.db,
                                                rownames(mat_GSE82032),
                                                keytype = "ACCNUM",
                                                c("SYMBOL"))$SYMBOL
pheatmap(mat_GSE82032, main="Basal-like (HCC1143) TNBC vs Trametinib
         (1ÃµM 3 days)")


## ------------------------------------------------------------------------------------
library(pheatmap)
mat_GSE138780  <- log2count_norm_GSE138780[topGenesBas_GSE138780, ]
```

```
mat_GSE138780  <- mat_GSE138780 - rowMeans(mat_GSE138780)
rownames(mat_GSE138780) <- AnnotationDbi::select(org.Hs.eg.db,
                                                 rownames(mat_GSE138780),
                                                 keytype = "ENTREZID",
                                                 c("SYMBOL"))$SYMBOL
pheatmap(mat_GSE138780, main="Basal-like (HCC1806) TNBC vs Trametinib
         (0.03ÃµM 1 day)")


## ------------------------------------------------------------------------------------------
library(pheatmap)
mat_GSE138780_SUM159  <-
  log2count_norm_GSE138780_SUM159[topGenesBas_GSE138780_SUM159, ]
mat_GSE138780_SUM159  <- mat_GSE138780_SUM159 - rowMeans(mat_GSE138780_SUM159)
rownames(mat_GSE138780_SUM159) <- AnnotationDbi::select(org.Hs.eg.db,
            rownames(mat_GSE138780_SUM159),
            keytype = "ENTREZID", c("SYMBOL"))$SYMBOL
pheatmap(mat_GSE138780_SUM159, main="Mesenchymal (SUM-159) TNBC vs Trametinib
         (0.03ÃµM 1 day)")


## ------------------------------------------------------------------------------------------
y_GSE82032 = estimateDisp(dgeObj_norm_GSE82032, design_GSE82032, robust=TRUE)
plotBCV(y_GSE82032)


## ------------------------------------------------------------------------------------------
y_GSE138780 = estimateDisp(dgeObj_norm_GSE138780, design_GSE138780, robust=TRUE)
plotBCV(y_GSE138780)


## ------------------------------------------------------------------------------------------
y_GSE138780_SUM159 = estimateDisp(dgeObj_norm_GSE138780_SUM159,
                                  design_GSE138780_SUM159, robust=TRUE)
plotBCV(y_GSE138780_SUM159)


## ------------------------------------------------------------------------------------------
fit_GSE82032 <- glmQLFit(y_GSE82032, design_GSE82032, robust = TRUE)


## ------------------------------------------------------------------------------------------
fit_GSE138780 <- glmQLFit(y_GSE138780, design_GSE138780, robust = TRUE)


## ------------------------------------------------------------------------------------------
fit_GSE138780_SUM159 <- glmQLFit(y_GSE138780_SUM159, design_GSE138780_SUM159,
                                 robust = TRUE)


## ------------------------------------------------------------------------------------------
res_GSE82032 <- glmQLFTest(fit_GSE82032, contrast = cont.matrix_GSE82032)
head(res_GSE82032)
```

```
## ------------------------------------------------------------------------------------------
res_GSE138780 <- glmQLFTest(fit_GSE138780, contrast = cont.matrix_GSE138780)
head(res_GSE138780)


## ------------------------------------------------------------------------------------------
res_GSE138780_SUM159 <- glmQLFTest(fit_GSE138780_SUM159,
                                   contrast = cont.matrix_GSE138780_SUM159)
head(res_GSE138780_SUM159)


## ------------------------------------------------------------------------------------------
topTags_edge_GSE82032 <- topTags(res_GSE82032,
                                 n=dim(log2count_norm_GSE82032)[1])
# todos los genes

geneIds <- AnnotationDbi::select(org.Hs.eg.db,
                                 keys = rownames(topTags_edge_GSE82032),
                                 keytype = "ACCNUM",
                                 columns =  c("SYMBOL", "GENENAME"))

anotaTopTab <- function(toptab) {
   topTab_Anot <- toptab %>%
       mutate(ACCNUM = rownames(toptab)) %>%
       merge(geneIds, "ACCNUM") %>%
       arrange(PValue)
   return(topTab_Anot)
}
toptab_anot_GSE82032_edge <- anotaTopTab(topTags_edge_GSE82032$table)

head(toptab_anot_GSE82032_edge)


## ------------------------------------------------------------------------------------------
topTags_edge_GSE138780 <- topTags(res_GSE138780,
                                  n=dim(log2count_norm_GSE138780)[1])
# todos los genes

geneIds <- AnnotationDbi::select(org.Hs.eg.db,
                                 keys = rownames(topTags_edge_GSE138780),
                                 keytype = "ENTREZID",
                                 columns =  c("SYMBOL", "GENENAME"))

anotaTopTab <- function(toptab) {
   topTab_Anot <- toptab %>%
       mutate(ENTREZID = rownames(toptab)) %>%
       merge(geneIds, "ENTREZID") %>%
       arrange(PValue)
   return(topTab_Anot)
}
toptab_anot_GSE138780_edge <- anotaTopTab(topTags_edge_GSE138780$table)
```

```
head(toptab_anot_GSE138780_edge)


## ------------------------------------------------------------------------------------------------
topTags_edge_GSE138780_SUM159 <- topTags(res_GSE138780_SUM159,
                                         n=dim(log2count_norm_GSE138780_SUM159)[1])
# todos los genes

geneIds <- AnnotationDbi::select(org.Hs.eg.db,
                                 keys = rownames(topTags_edge_GSE138780_SUM159),
                                 keytype = "ENTREZID",
                                 columns =  c("SYMBOL", "GENENAME"))

toptab_anot_GSE138780_edge_SUM159 <-
  anotaTopTab(topTags_edge_GSE138780_SUM159$table)

head(toptab_anot_GSE138780_edge_SUM159)


## ------------------------------------------------------------------------------------------------
write.csv(topTags_edge_GSE82032, file="resultats/toptab_anot_GSE82032_edge.csv")


## ------------------------------------------------------------------------------------------------
write.csv(topTags_edge_GSE138780,
          file="resultats/toptab_anot_GSE138780_edge.csv")


## ------------------------------------------------------------------------------------------------
write.csv(topTags_edge_GSE138780_SUM159,
          file="resultats/toptab_anot_GSE138780_edge_SUM159.csv")


## ------------------------------------------------------------------------------------------------
topGenes_edge_GSE82032 <- rownames(subset(topTags_edge_GSE82032$table,
                                          (abs(logFC)> 2) & (PValue < 0.01)))

length(topGenes_edge_GSE82032)


## ------------------------------------------------------------------------------------------------
topGenes_edge_GSE138780 <- rownames(subset(topTags_edge_GSE138780$table,
                                           (abs(logFC)> 2) & (PValue < 0.01)))
length(topGenes_edge_GSE138780)


## ------------------------------------------------------------------------------------------------
topGenes_edge_GSE138780_SUM159 <-
  rownames(subset(topTags_edge_GSE138780_SUM159$table, (abs(logFC)> 2) &
                    (PValue < 0.01)))
length(topGenes_edge_GSE138780_SUM159)
```

```
## ------------------------------------------------------------------------------------
library(ggvenn)
x = list(LimmaVoom = topGenesBas_GSE82032, edgeR = topGenes_edge_GSE82032)
ggvenn(x, fill_color = c("#0073C2FF", "#EFC000FF"), stroke_size = 0.5,
       set_name_size = 3)


## ------------------------------------------------------------------------------------
library(ggvenn)
x = list(LimmaVoom = topGenesBas_GSE138780, edgeR = topGenes_edge_GSE138780)
ggvenn(x, fill_color = c("#0073C2FF", "#EFC000FF"), stroke_size = 0.5,
       set_name_size = 3)


## ------------------------------------------------------------------------------------
library(ggvenn)
x = list(LimmaVoom = topGenesBas_GSE138780_SUM159,
         edgeR = topGenes_edge_GSE138780_SUM159)
ggvenn(x, fill_color = c("#0073C2FF", "#EFC000FF"), stroke_size = 0.5,
       set_name_size = 3)


## ------------------------------------------------------------------------------------
topGenes_GSE82032 <- union(topGenesBas_GSE82032, topGenes_edge_GSE82032)
length(topGenes_GSE82032)
universe_GSE82032 <- rownames(toptab_GSE82032)
length(universe_GSE82032)


## ------------------------------------------------------------------------------------
topGenes_GSE138780 <- union(topGenesBas_GSE138780, topGenes_edge_GSE138780)
length(topGenes_GSE138780)
universe_GSE138780 <- rownames(toptab_GSE138780)
length(universe_GSE138780)


## ------------------------------------------------------------------------------------
topGenes_GSE138780_SUM159 <- union(topGenesBas_GSE138780_SUM159,
                                   topGenes_edge_GSE138780_SUM159)
length(topGenes_GSE138780_SUM159)
universe_GSE138780_SUM159 <- rownames(toptab_GSE138780_SUM159)
length(universe_GSE138780_SUM159)


## ------------------------------------------------------------------------------------
library(org.Hs.eg.db)
AnnotationDbi::keytypes(org.Hs.eg.db)
topAnots_GSE82032 = AnnotationDbi::select(org.Hs.eg.db, topGenes_GSE82032,
                                          c("SYMBOL", "ENTREZID", "GENENAME"),
keytype = "ACCNUM")
head(topAnots_GSE82032)
dim(topAnots_GSE82032)
```

```
## ------------------------------------------------------------------------------------------
topAnots_GSE138780 = AnnotationDbi::select(org.Hs.eg.db, topGenes_GSE138780,
                                           c("SYMBOL", "ENTREZID", "GENENAME"),
keytype = "ENTREZID")
head(topAnots_GSE138780)
dim(topAnots_GSE138780)


## ------------------------------------------------------------------------------------------
topAnots_GSE138780_SUM159 = AnnotationDbi::select(org.Hs.eg.db,
                                           topGenes_GSE138780_SUM159,
                                           c("SYMBOL", "ENTREZID", "GENENAME"),
keytype = "ENTREZID")
head(topAnots_GSE138780_SUM159)
dim(topAnots_GSE138780_SUM159)


## ------------------------------------------------------------------------------------------
univAnots_GSE82032 = AnnotationDbi::select(org.Hs.eg.db, universe_GSE82032,
                                           c("SYMBOL", "ENTREZID", "GENENAME"),
                                           keytype = "ACCNUM")
head(univAnots_GSE82032)
dim(univAnots_GSE82032)


## ------------------------------------------------------------------------------------------
univAnots_GSE138780 = AnnotationDbi::select(org.Hs.eg.db, universe_GSE138780,
                                           c("SYMBOL", "ENTREZID", "GENENAME"),
                                           keytype = "ENTREZID")
head(univAnots_GSE138780)
dim(univAnots_GSE138780)


## ------------------------------------------------------------------------------------------
univAnots_GSE138780_SUM159 = AnnotationDbi::select(org.Hs.eg.db,
                                          universe_GSE138780_SUM159,
                                           c("SYMBOL", "ENTREZID", "GENENAME"),
                                          keytype = "ENTREZID")
head(univAnots_GSE138780_SUM159)
dim(univAnots_GSE138780_SUM159)


## ------------------------------------------------------------------------------------------
library(clusterProfiler)
ego_GSE82032 = enrichGO(gene = topAnots_GSE82032$ENTREZID,
             universe=univAnots_GSE82032$ENTREZID,
             keyType = "ENTREZID",
             OrgDb = org.Hs.eg.db,
             ont="BP",
             pAdjustMethod = "BH",
             pvalueCutoff = 0.05,
             qvalueCutoff = 0.05,
             readable = TRUE)
```

```
## ----------------------------------------------------------------------------------------------
ego_GSE138780 = enrichGO(gene = topAnots_GSE138780$ENTREZID,
              universe=univAnots_GSE138780$ENTREZID,
              keyType = "ENTREZID",
              OrgDb = org.Hs.eg.db,
              ont="BP",
              pAdjustMethod = "BH",
              pvalueCutoff = 0.05,
              qvalueCutoff = 0.05,
              readable = TRUE)


## ----------------------------------------------------------------------------------------------
ego_GSE138780_SUM159 = enrichGO(gene = topAnots_GSE138780_SUM159$ENTREZID,
              universe=univAnots_GSE138780_SUM159$ENTREZID,
              keyType = "ENTREZID",
              OrgDb = org.Hs.eg.db,
              ont="BP",
              pAdjustMethod = "BH",
              pvalueCutoff = 0.05,
              qvalueCutoff = 0.05,
              readable = TRUE)


## ----------------------------------------------------------------------------------------------
head(ego_GSE82032[,-c(2,8)], n=5)
head(ego_GSE82032[,2:3], n=5)
head(ego_GSE82032[,c(2,8)], n=5)
write.table(ego_GSE82032, file="resultats/EnrichmentResults_GSE82032.csv",
            dec=".", sep=";")


## ----------------------------------------------------------------------------------------------
head(ego_GSE138780[,-c(2,8)], n=5)
head(ego_GSE138780[,2:3], n=5)
head(ego_GSE138780[,c(2,8)], n=5)
write.table(ego_GSE138780, file="resultats/EnrichmentResults_GSE138780.csv",
            dec=".", sep=";")


## ----------------------------------------------------------------------------------------------
head(ego_GSE138780_SUM159[,-c(2,8)], n=5)
head(ego_GSE138780_SUM159[,2:3], n=5)
head(ego_GSE138780_SUM159[,c(2,8)], n=5)
write.table(ego_GSE138780_SUM159,
            file="resultats/EnrichmentResults_GSE138780_SUM159.csv", dec=".",
            sep=";")


## ----viewEnrichment1----------------------------------------------------------------------------
dotplot(ego_GSE82032, showCategory=7, title="Basal-like (HCC1143) TNBC vs
        Trametinib (1ÃµM 3 days)")
```

```
## ------------------------------------------------------------------------------------------
dotplot(ego_GSE138780, showCategory=7, title="Basal-like (HCC1806) TNBC vs
        Trametinib (0.03ÃµM 1 day)")


## ------------------------------------------------------------------------------------------
dotplot(ego_GSE138780_SUM159, showCategory=7, title="Mesenchymal (SUM-159) TNBC
        vs Trametinib (0.03ÃµM 1 day)")


## ----viewEnrichment2-----------------------------------------------------------------------
library(ggplot2)
ego2_GSE82032 = simplify(ego_GSE82032)
cnetplot(ego2_GSE82032, showCategory = 3, cex_category =0.3,
       cex_label_category =0.7, cex_gene=0.2, cex_label_gene=0.4,
       circular=TRUE, colorEdge=TRUE, title="Basal-like (HCC1143) TNBC vs
       Trametinib (1ÃµM 3 days)")


## ------------------------------------------------------------------------------------------
ego2_GSE138780 = simplify(ego_GSE138780)
cnetplot(ego2_GSE138780, showCategory = 3, cex_category =0.3,
       cex_label_category =0.7, cex_gene=0.2, cex_label_gene=0.4,
       circular=TRUE, colorEdge=TRUE, title="Basal-like (HCC1806) TNBC vs
       Trametinib (0.03ÃµM 1 day)")


## ------------------------------------------------------------------------------------------
ego2_GSE138780_SUM159 = simplify(ego_GSE138780_SUM159)
cnetplot(ego2_GSE138780_SUM159, showCategory = 3, cex_category =0.3,
       cex_label_category =0.7, cex_gene=0.2, cex_label_gene=0.4,
       circular=TRUE, colorEdge=TRUE, title="Mesenchymal (SUM-159) TNBC vs
       Trametinib (0.03ÃµM 1 day)")


## ----viewEnrichment3-----------------------------------------------------------------------
library(enrichplot)
goplot(ego2_GSE82032, showCategory=10, cex=0.1, title="Basal-like (HCC1143)
       TNBC vs Trametinib (1ÃµM 3 days)")


## ------------------------------------------------------------------------------------------
goplot(ego2_GSE138780, showCategory=10, cex=0.1, title="Basal-like (HCC1806)
       TNBC vs Trametinib (0.03ÃµM 1 day)")


## ------------------------------------------------------------------------------------------
goplot(ego2_GSE138780_SUM159, showCategory=10, cex=0.1, title="Mesenchymal
       (SUM-159) TNBC vs Trametinib (0.03ÃµM 1 day)")


## ------------------------------------------------------------------------------------------
```

```
term_similarity_matrix_GSE82032 = pairwise_termsim(ego_GSE82032)
emapplot(term_similarity_matrix_GSE82032, showCategory = 15,
       group_category=TRUE, group_legend=TRUE, title="Basal-like (HCC1143)
       TNBC vs Trametinib (1ÃµM 3 days)")


## ------------------------------------------------------------------------------------------
term_similarity_matrix_GSE138780 = pairwise_termsim(ego_GSE138780)
emapplot(term_similarity_matrix_GSE138780, showCategory = 15,
       group_category=TRUE, group_legend=TRUE, title="Basal-like (HCC1806)
       TNBC vs Trametinib (0.03ÃµM 1 day)")


## ------------------------------------------------------------------------------------------
term_similarity_matrix_GSE138780_SUM159 = pairwise_termsim(ego_GSE138780_SUM159)
emapplot(term_similarity_matrix_GSE138780_SUM159, showCategory = 15,
       group_category=TRUE, group_legend=TRUE, title="Mesenchymal (SUM-159)
       TNBC vs Trametinib (0.03ÃµM 1 day)")


## ---- eval = FALSE, code=readLines("TFM.R")-----------------------------------------------
  code=readLines("TFM.R")
 # knitr::purl("TFM.Rmd")
```