FH KREMS
UNIVERSITY OF APPLIED SCIENCES / AUSTRIA

Final Report

Big Data for Bioanalytics and Medicine

RNA Sequencing Analysis

Melanoma - GSE158403

Becerra Larcher, Juan Cruz
Orschanski, Daniela

05/05/2022

Introduction

Gene expression and its regulation are very important to understand the behavior of cells under different conditions.

Nowadays, various techniques are used to study gene expression. One of those is RNA sequencing, which can examine the quantity and sequences of RNA in a sample using next-generation sequencing (NGS). It analyzes the transcriptome, indicating which of the genes encoded in our DNA are turned on or off and to what extent. This technique is the one we have used to develop this analysis.

In this project we have developed an RNA sequencing analysis for a study carried out by the University of California, department of Medicine in the United States of America, which has melanoma patients as the main subjects of the research.

The accession number of the publication we have based our report on is GSE158403 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE158403) and the publication link is https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7721806/ .

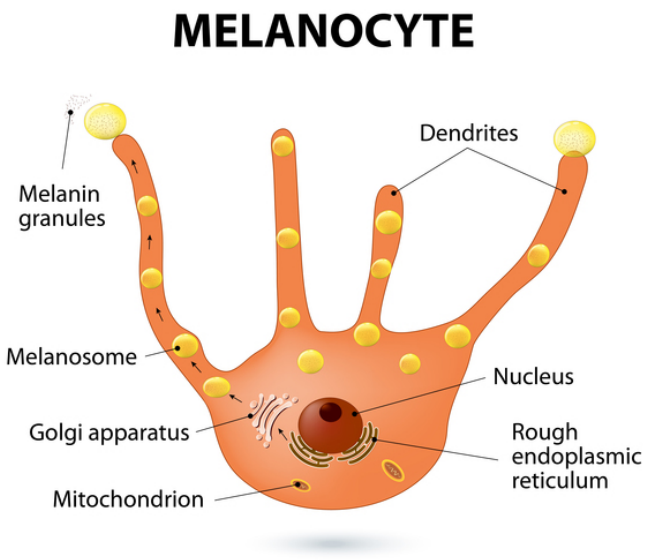
We have splitted this report into two main sections: the first one will describe the biological context, which means we will explain the analysis we are performing, a general idea about the disease in question, some details about the research done by the university in order to understand the information we are dealing with, among other topics.

The second section is purely based on the coding session we have done in order to read the files, such as the count matrix and the sample annotation file, from which we could extract all the information needed to perform the analysis of the differentially expressed profiles.

I. Biological Context

1. *Skin cancer - Melanoma*

Melanoma is a type of skin cancer that begins in pigment-producing cells called melanocytes.



Source: https://medlineplus.gov/genetics/condition/melanoma/#causes

This cancer typically occurs in areas that are only occasionally sun-exposed; tumors are most commonly found on the back in men and on the legs in women.

There are different kinds of this disease which can be classified in regard of the type of tissue in which the melanoma appears:

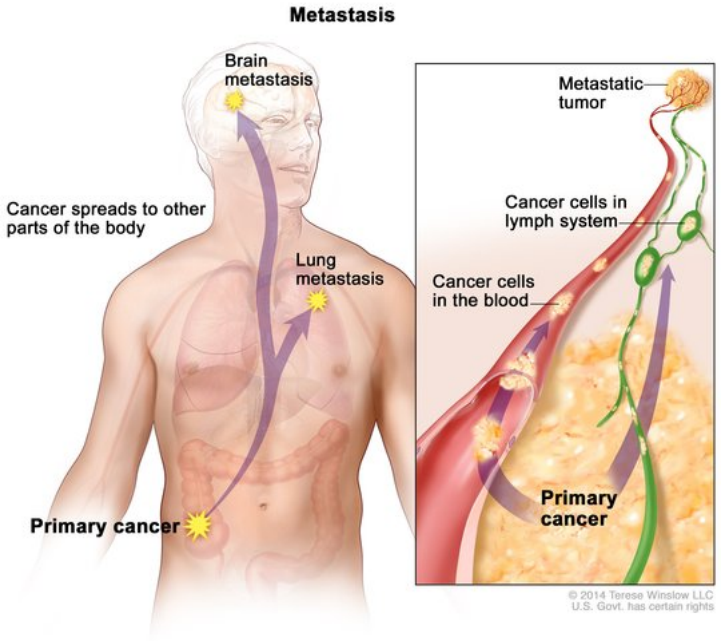
Melanoma is usually present on the skin (cutaneous melanoma), but in about 5% of cases it develops in melanocytes living in other tissues, including the eyes (uveal melanoma) or mucous membranes that line the body's cavities, such as the moist lining of the mouth (mucosal melanoma).

Melanoma can develop at any age, but it most frequently occurs in people in their fifties to seventies and is becoming more common in teenagers and young adults.

The origin of this skin cancer may come from an existing mole or other normal skin growth that becomes cancerous (malignant); however, many melanomas are new growths. These skin marks are often identified because they have ragged edges and an irregular shape. They can range from a few millimeters to several centimeters across. They can also be a variety of colors: brown, black, red, pink, blue, or white.

Source: https://medlineplus.gov/genetics/condition/melanoma/#causes

Most melanomas affect only the outermost layer of skin (the epidermis). If a melanoma becomes thicker and involves multiple layers of skin, it can spread to other parts of the body (metastasize).



Source: https://medlineplus.gov/genetics/condition/melanoma/#causes

A large number of moles or other pigmented skin growths on the body, generally more than 25, is associated with an increased risk of developing melanoma. Melanoma is also a common feature of genetic syndromes affecting the skin such as xeroderma pigmentosum (inherited condition characterized by an extreme sensitivity to ultraviolet (UV) rays from sunlight).

Additionally, individuals who have previously had melanoma are nearly nine times more likely than the general population to develop melanoma again. It is estimated that about 90% of individuals with melanoma survive at least 5 years after being diagnosed.

Melanoma is a disease that affects the genetic information because the UV radiation damages DNA. Most of the time, this kind of DNA damage causes cells to self-destruct (undergo apoptosis) or permanently stop cell division (undergo senescence).

However, melanocytes are more resistant than other kinds of cells to the effects of UV radiation. Even if their DNA is damaged, melanocytes are unlikely to undergo apoptosis.

As abnormal melanocytes continue to grow, they accumulate genetic mutations, particularly in genes that control cell growth and division (proliferation), senescence, and apoptosis. Ultimately, the cells become able to proliferate without control or limit and can resist cell death, leading to the formation and growth of a melanoma.

In regards to the source of this skin cancer, most cases of melanoma are sporadic, which means that the genetic changes are acquired during a person's lifetime and are present only in the melanocytes that give rise to the melanoma. These changes, which are called somatic mutations, are not inherited.

Studies suggest that in some cases somatic variations in multiple genes, each with a small effect, combine to increase the risk of developing the condition. Many of these gene variations are associated with a light complexion and increased freckles and moles. However, it is unclear what contribution each of these genetic changes makes to disease risk.

Other somatic gene mutations have large effects on melanoma risk and a mutation in one gene is enough to significantly increase the risk of developing cancer. Somatic mutations in the BRAF and CDKN2A genes are some of the most common in sporadic melanoma.

On the other hand, in about 10% of cases, Melanoma occurs in multiple members of the same family. Unlike sporadic melanoma, these familial cases are typically caused by inherited genetic changes that increase the risk of developing this type of cancer. These genetic changes, which are classified as germline mutations, are present in essentially all of the body's cells.

The primary genes involved in familial melanoma are CDKN2A and MC1R. The CDKN2A gene plays a role in regulating cell senescence and the MC1R gene influences skin pigmentation. Shared nongenetic factors can also influence the development of melanoma in family members; for example, family members may live in the same sun-exposed environment. In people with germline mutations, changes in other genes, together with environmental and lifestyle factors, also contribute to the possible development of melanoma.

2. *RNA Sequencing*

RNA-seq (RNA-sequencing) is a technique that can examine the quantity and sequences of RNA in a sample using next-generation sequencing (NGS). It analyzes the transcriptome, indicating which of the genes encoded in our DNA are turned on or off and to what extent.

3. *GSE file*

GSE is a list of GSM files that together form a single experiment. While GSM are files that contain all the data from the use of a single chip. For each gene there will be multiple scores including the main one, held in the value column.

GSE is a system to represent microarray data and metadata in a relational database.

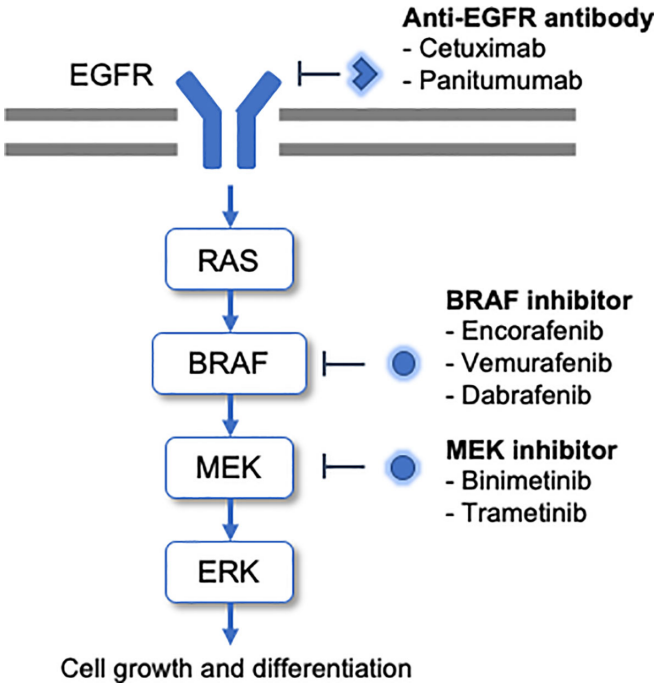
4. *BRAF*

BRAF is a gene, encoded on chromosome 7q34, that provides instructions for making the B-RAF protein which is composed of 766 amino acids and is a member of the RAF kinase family. This protein helps transmit chemical signals from outside the cell to the cell's nucleus. It is made of 21 exons.

This protein is part of a signaling pathway known as the RAS/MAPK pathway, which controls several important cell functions. Specifically, the RAS/MAPK pathway regulates:

- Proliferation: the growth and division (proliferation) of cells
- Differentiation: the process by which cells mature to carry out specific functions
- Migration: Cell movement
- Apoptosis: Self-destruction of cells. Chemical signaling through this pathway is essential for normal development before birth.

Basically, BRAF is a gene that tells the cells how to grow.

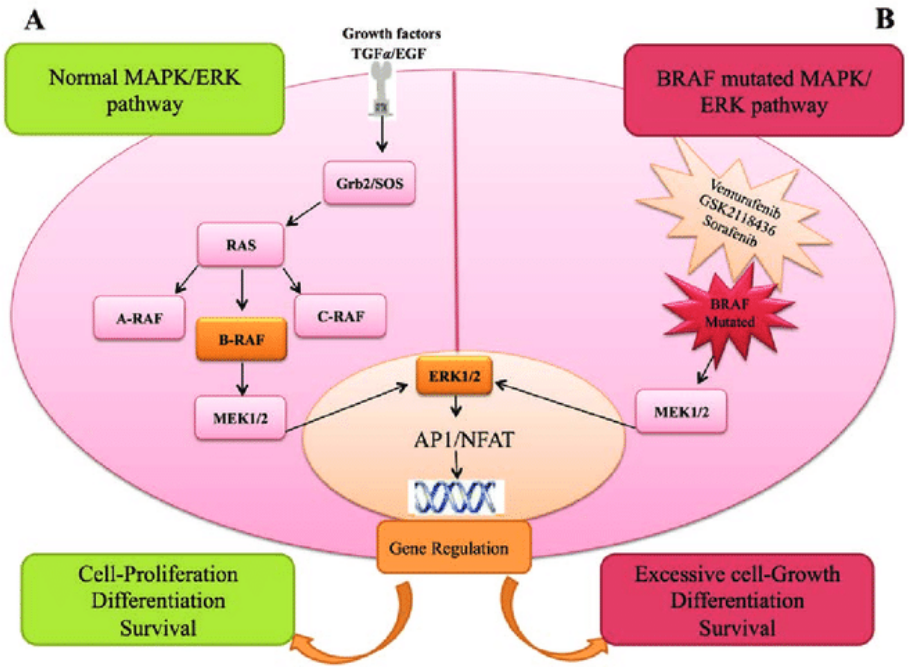


Source: https://www.frontiersin.org/articles/10.3389/fonc.2021.602194/full

The BRAF gene belongs to a class of genes known as oncogenes. When mutated, oncogenes have the potential to cause normal cells to become cancerous. So BRAF mutation is a change in a BRAF gene which can lead to an alteration in a protein that regulates cell growth that could allow the melanoma to grow more aggressively.

About 50% of melanomas harbors activating BRAF mutations, over 90 % are at codon 600, and among these, over 90 % are a single nucleotide mutation resulting in substitution of glutamic acid for valine (BRAFV600E: nucleotide 1799 T > A; codon GTG > GAG).

Collapse Section



Source:
https://www.researchgate.net/figure/Role-of-BRAF-in-A-normal-and-B-mutated-state-in-the-activation-of-MAPK-pathway-MAPK_fig1_277084891

II. Coding Section

We start the coding part by downloading the count matrix file and the sample annotation file from the database on the NCBI website (National Center for Biotechnology Information):

1. *Count matrix*

	Cell1	Cell2	…	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
…	.	.	.	.
…	.	.	.	.
…	.	.	.	.
GeneM	25	0	.	0

Source:
https://hbctraining.github.io/scRNA-seq/lessons/02_SC_generation_of_count_matrix.html

The count matrix has genes along the rows and samples along the columns. The elements in the matrix give the number of reads which could be uniquely aligned to a given gene for a given sample.

Input for counting = multiple BAM files + 1 GTF file

Simply speaking, the genomic coordinates of where the read is mapped (BAM) are cross-referenced with the genomic coordinates of whichever feature you are interested in counting expression of (GTF), it can be exons, genes or transcripts.

Output of counting = A count matrix, with genes as rows and samples are columns

These are the "raw" counts and will be used in statistical programs downstream for differential gene expression.

2. *Sample annotation*

The sample annotation file contains rows that represent the 95 samples associated with the samples (columns) of the count matrix. Also, it contains columns that represent different factors that characterize the samples, such as the site where the melanoma is (liver, lung, etc.), the type of melanoma (braf mutated or braf wild type), cohort (A, B, C), among others.

RNA-sequenced gene expression profiling was performed on 95 biopsies from 65 patients who are classified in cohorts depending on the treatment they received and the condition in which they are. Those cohorts are:

a. Cohort A: initially included 6 patients who received durvalumab 3 mg/kg every 2 weeks (Q2W) plus standard doses of dabrafenib plus trametinib. Once the safety of this dose was assessed, an additional 20 patients received durvalumab at 10 mg/kg Q2W, of which 7 were enrolled and treated in the second dose-escalation phase and 13 were enrolled and treated as part of the dose-expansion phase.
   The median age in cohort A was 49.0 years (range: 23–71), which is consistent with younger age observed in patients with BRAFV600-mutant melanoma

b. Cohort B: 20 patients were enrolled and received durvalumab 10 mg/kg Q2W concomitantly with trametinib 2 mg every day (QD); 6 patients were enrolled and treated in the dose-escalation phase and 14 patients were enrolled and treated in the dose-expansion phase.
   Cohort B had a median age of 68.0 years (range: 31–85)

c. Cohort C: 22 patients received the same combination but sequentially (trametinib: days 1–42; durvalumab: from day 29); 7 patients were enrolled and treated in the dose-escalation phase and 15 patients were enrolled and treated in the dose-expansion phase (Supplementary Fig. 1).
   Cohort C had a median age of 63.0 years (34–84)

The majority of patients—21 (80.8%), 18 (90.0%), and 18 (81.8%) for cohorts A, B, and C, respectively—had stage IV metastatic melanoma.

Median lactate dehydrogenase (LDH) levels at baseline for cohorts A, B, and C were 206.5 units per liter (U/L), 215.0 U/L, and 225.0 U/L, respectively; levels were elevated in approximately one-third of patients per cohort.

In cohort A, 10 (38.5%) patients had received prior systemic therapies, which was a lower proportion than those seen in cohort B (13 [65.0%]) and cohort C (15 [68.2%]).

Prior treatments in cohorts B and C included immunotherapy in 12 (60.0%) and 11 (50.0%) patients, respectively

Characteristic	Cohort A: Durvalumab 3 or 10 mg/kg + dabrafenib + trametinib (*n* = 26)	Cohort B: Durvalumab 10 mg/kg + trametinib (concurrent) (*n* = 20)	Cohort C: Durvalumab 10 mg/kg + trametinib (sequential) (*n* = 22)
Age, median (range), years	49.0 (23–71)	68.0 (31–85)	63.0 (34–84)
Male sex, *n* (%)	14 (53.8)	13 (65.0)	11 (50.0)
LDH level[a], *n(%)*			
Normal	15 (57.7%)	14 (70.0%)	12 (54.5%)
High	9 (34.6%)	6 (30.0%)	7 (31.8%)
Missing	2 (7.7%)	0	3 (13.6%)
Race			
Black or African American	0	0	1 (4.5)
White	20 (76.9)	15 (75.0)	13 (59.1)
Not collected	6 (23.1)	5 (25.0)	6 (27.3)
Mutation status,n (%)			
BRAF-wild type	0	19 (95.0)	22 (100.0)

In the last row from the previous table we can see that the samples can also be classified in terms of the type of melanoma: BRAF-mutated Melanoma or BRAF-Wild Type Melanoma.

- Processing the date of the Sample Annotation file:

It is required that the rows of the sample annotation match with the columns of the count matrix. This means that the files will be connected by the samples. In order to make this possible we saved the sample annotation file as an excel file so that we can make all the changes needed.

The most fundamental change was to replace the "-" for "." in the row names.

Then, we deleted the columns that contain the same information for every row. They are the following:

- ➔ Status: public on Sep 24 2020
- ➔ Type: SRA
- ➔ Source name: Melanoma
- ➔ Organisms: Homo sapiens
- ➔ Sample type: tumor biopsy
- ➔ Molecule ch1: total RNA
- ➔ Extract protocol ch1: Qiagen
- ➔ Extract protocol ch1.1: Takara Bio SMART-Seq: SMART-Seq v4 Ultra Low Input RNA Kit

- ➔ Extract protocol ch1.2: RNA sequencing was performed on Illumina HiSeq 2500 with 2x75 bp reads
- ➔ Data processing: RNAseq reads were aligned using HISAT2 to the human reference genome GRCh38
- ➔ Taxid: 9606
- ➔ Description: polyA RNA
- ➔ Data processing.1: FPKM values were generated using Stringtie.
- ➔ Data processing.2: Genome_build: GRCh38
- ➔ Data processing.3: Supplementary_files_format_and_content: tab-delimited text file includes FPKM values for each sample
- ➔ Platform_id: GPL16791
- ➔ Contact_name: Antoni,,Ribas
- ➔ Contact_name: aribas@mednet.ucla.edu
- ➔ Contact_phone: (310) 206-3928
- ➔ Contact_department: Medicine, HemOnc Division
- ➔ contact_institute: University of California, Los Angeles
- ➔ contact_address: 9-954 Factor Bldg
- ➔ contact_city: Los Angeles
- ➔ contact_state: CA
- ➔ contact_zip.postal_code: 90095-1678
- ➔ contact_country: USA
- ➔ data_row_count: 0
- ➔ instrument_model: Illumina HiSeq 2500
- ➔ library_selection: cDNA
- ➔ library_source: transcriptomic
- ➔ library_strategy: RNA-Seq
- ➔ supplementary_file_1: NONE

3. Results DESeq

By applying the DESeq function we obtain a large DESeq data set whose results can be extracted with the method “results” from the same package in R.

Since we wanted to compare the differential expressions of the genes in each cohort, we made use of contrasts. That led to an excel file with 3 slides: each one compares the base mean, log2 fold change, p value, p adjusted value, statistics and lfcSE of 2 cohorts. This means that we have cohort A vs cohort B, B vs C and A vs C.

Those parameters will be described in the following paragraph:

The “lfcSE” value is the standard error of the log2FoldChange and “Stats” stands for statistics.

3.1. Log2 Fold change

Fold change is the number of times a gene is over-expressed (or under), compared to some baseline (your control, or the reference gene, etc.). A sample could be 100X more expressed, or 1/100th the expression of the baseline. Because this is hard to show in a graph, we plot in log. It "flattens" the data out to make it more visible.

For example, if logFC = -0.5, then FC = 2-0.5, or 0.7071, which means about 70% of the baseline, not 50%... 50% reduction in expression would be a logFC of -1.

3.2. P value

In this analysis the p value represents the relationship between the changes in the expression of the genes within a cohort. This means that, If it is closer to 0, it will indicate that the expression of the genes in one cohort is relatable to the expression of the genes in the other cohort we are comparing with.

By default, independent filtering is performed to select a set of genes for multiple test correction which maximizes the number of adjusted p-values less than a given critical value alpha (by default 0.1). The filter used for maximizing the number of rejections is the mean of normalized counts for all samples in the dataset. Several arguments from the filtered_p function of the genefilter package (used within the results function) are provided here to control the independent filtering behavior.

In DESeq2 version >= 1.10, the threshold that is chosen is the lowest quantile of the filter for which the number of rejections is close to the peak of a curve fit to the number of rejections over the filter quantiles. 'Close to' is defined as within 1 residual standard deviation. The adjusted p-values for the genes which do not pass the filter threshold are set to NA.

By default, results assign a p-value of NA to genes containing count outliers, as identified using Cook's distance. See the cooksCutoff argument for control of this behavior. Cook's distances for each sample are accessible as a matrix "cooks" stored in the assays() list. This measure is useful for identifying rows where the observed counts might not fit to a Negative Binomial distribution.

A p-value of 0.05 implies that we are willing to accept that 5% of all tests will be false positives. An adjusted p-value (aka a q-value) of 0.05 implies that we are willing to accept that 5% of the tests found to be statistically significant (e.g. by p-value) will be false positives. Such an adjustment is necessary when we're making multiple tests on the same sample.

4. Molecular signature database

Is a database maintained by the Broad Institute of the University of California in San Diego.
It is a very comprehensive collection of gene sets which contains 8 categories:

a. Hallmark gene sets
b. Positional gene sets
c. Curated
d. Regulatory target
e. Computational

f. Entology
g. Oncogenic signature
h. Immunologic signature
i. Cell type signature gene sets

Each category has genes associated with different pathways. The pathway sources or databases are: BIOCARTA, KEGG, PID, REACTOME and WikiPathways
CGP (chemical and genetic perturbations): genes affected by the disease we are analyzing
CP: Canonical pathways: collection of all the pathway sources

If we enter into a source and select a pathway, then we can select one of the member genes. Now we can see all the factors or modulators that are involved in the diferentiation of the preadipocyte (adipocytes are the fat cells) into a differentiated adipocyte. If we detect that these factors are upregulated we can conclude that there is something going on in this field.

There are some pathways which make reference to biological process such as mitotic cell cycle

5. Results - GSVA

In this section we transformed the count matrix into a gene set variation score matrix. Or gene set expression matrix. This score determines if the genes of the gene set are grouped more at the top or more at the button of the gene expression vector

The absolute value of this score is the more distinct when the concentration of the genes are on the top or at the bottom of the sample. So, if a gene is around 0 means that the gene set is distributed roughly equally across the gene expression vector. In the same way, the value -0,25 indicates that there is an accumulation of genes at the bottom.

In this case we have used the Poisson distribution because we are dealing with counts. It is a binomial distribution base and it is used for variables that are discrete. However, in proteomics or microarray we would use a distribution for continuous variables such as Gaussian.

As we extract the results we will see an excel file with columns such as: gs_id, gs_cat (category), gs_subcat, logFC, p value, adjusted p value, gs_exact_source and gs_name.

The variable “gs_name” indicates biological functions that are differentially regulated in a comparison of the cohort A vs B, for example. If the logFC is -0,23 it means that that biological function is down regulated.

The higher the p values are for one slide the less difference is between the 2 compared status of cohort in regards to changes in the differentially expressed pathways. The B_vs_C has the higher p values so in terms of differentially expressed pathways the cohort B and the cohort C are not so different. This makes sense because both share the same braf wild type melanoma.

6. Visualization

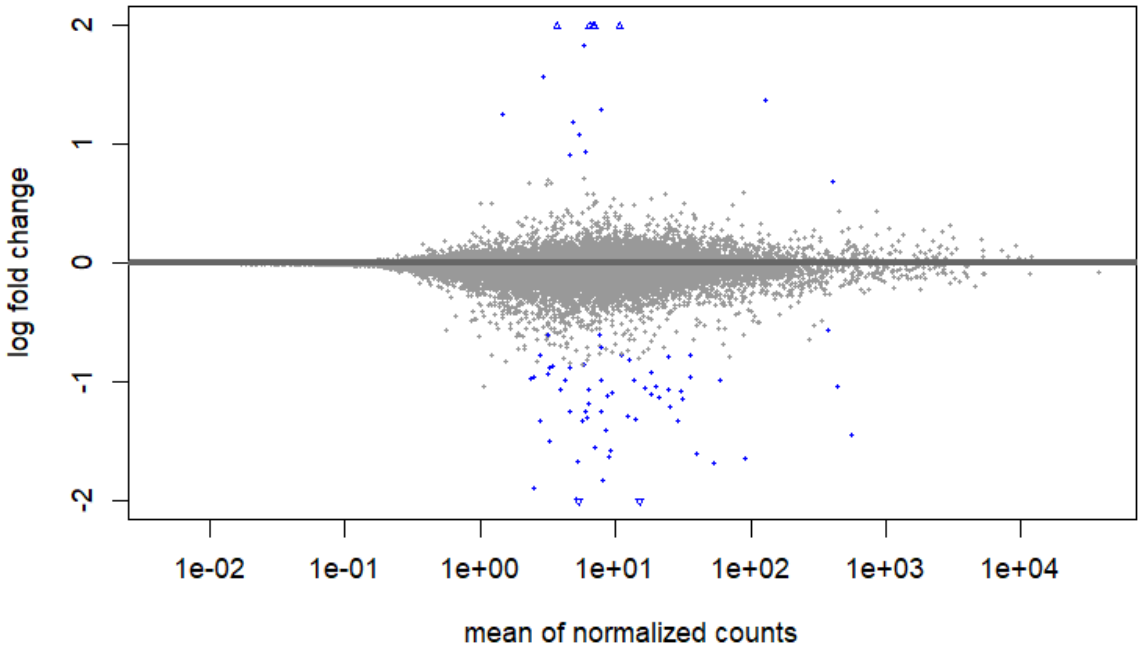


Figure 1: plot MA

This is a diagnostic plot.
The mean of normalized counts takes the mean across all samples for each gene.
This shows if the normalization removed the dependence of the standard deviation from the intensity (gene abundance). The sd should be independent from the intensity .
The logFC does not depend on the intensity because the intensity is the mean of normalized counts. If the logFC would have a trend towards the more intense genes, then there is something wrong. If those factors were dependent from each other we would see a slope, but in this case we see a horizontal line.

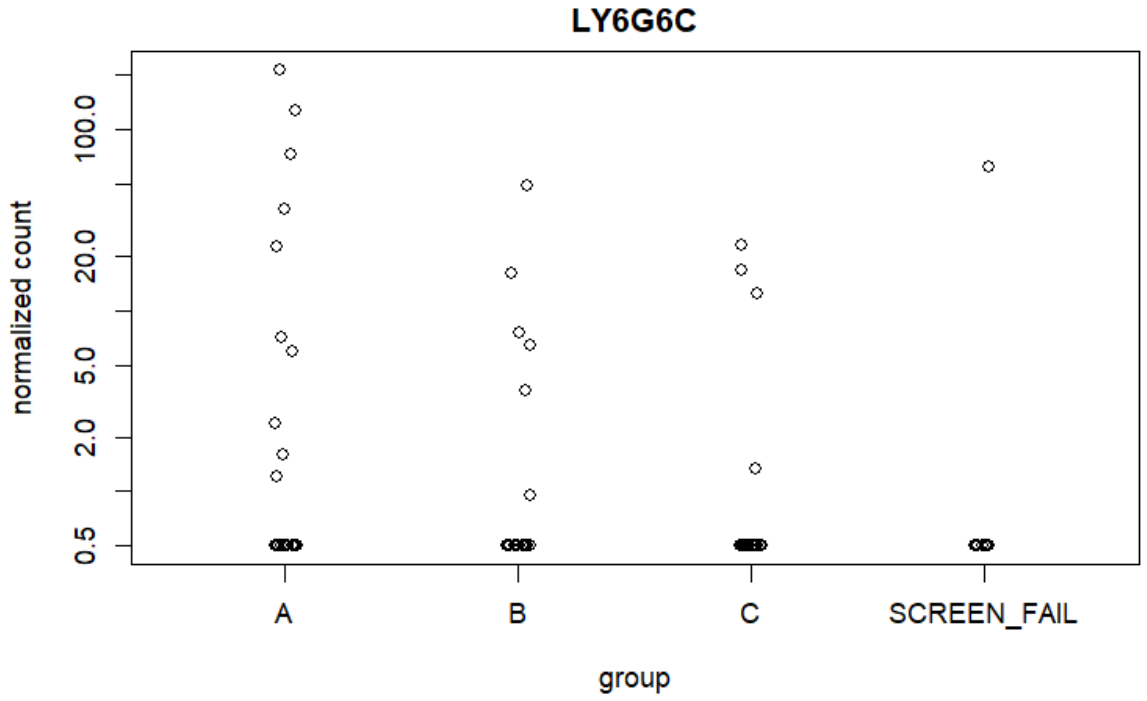


Figure 2: differential expression profile of LY6G6C gene

This is the gene expression which differs the most within the different cohorts. It was selected based on the minimum p adjusted value.

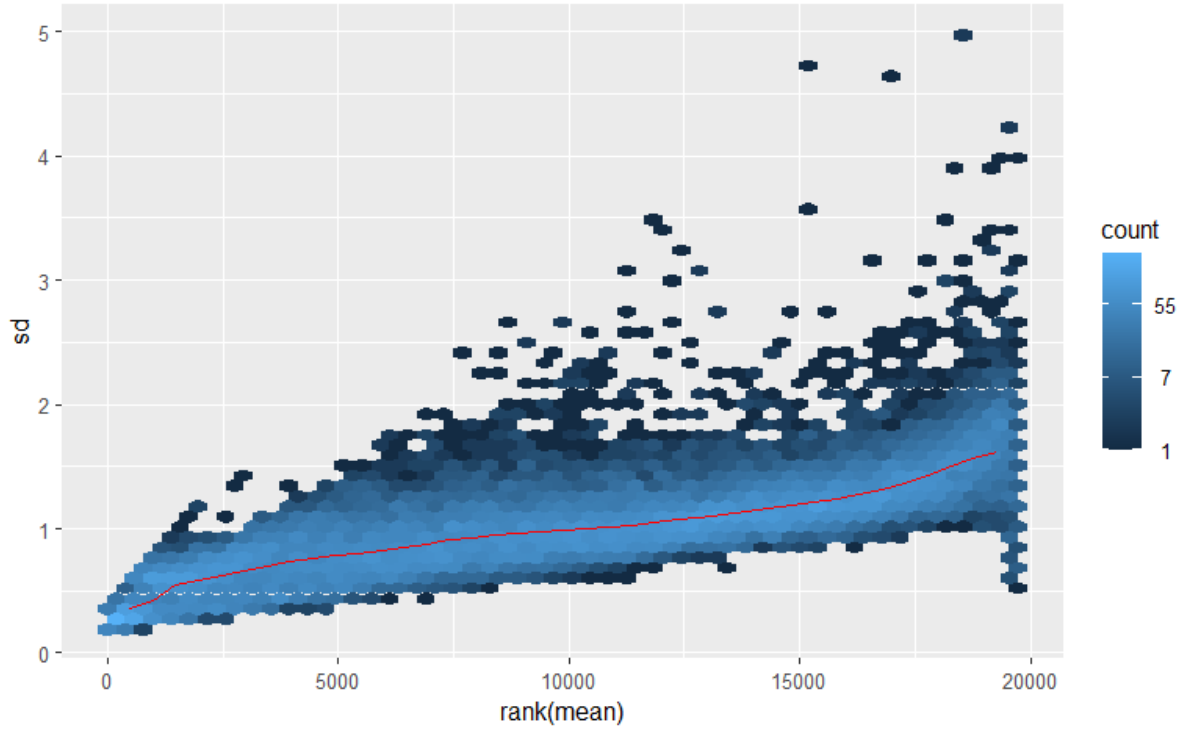


Figure 3: standard deviation vs rank - vsd transformation

This graphic plots the standard deviation of the transformed data, across samples, against the mean, using the variance stabilizing transformation. The scatterplot allows you to visually verify whether there is a dependence of the standard deviation (or variance) on the mean.

The red line depicts the running median estimator (window-width 10%). If there is no variance-mean dependence, then the line should be approximately horizontal.

We are seeing the rank in the x axis which represents the mean of the expression values of every gene. On the left side we have the lowest expression genes, whereas on the right side, we can see the highest expression genes independent from the expression itself. We are analyzing the difference of standard deviation between those 2 groups of genes. Every dot is a gene and the count is the expression.

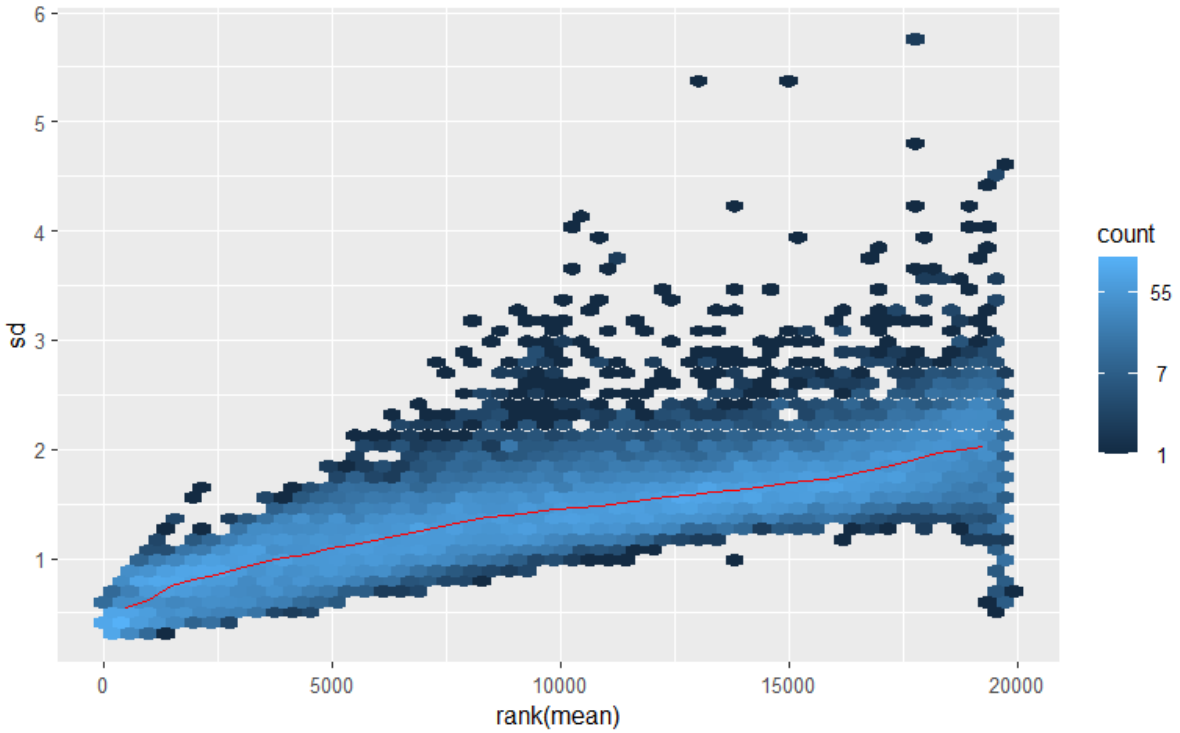


Figure 4: standard deviation vs rank - normal transformation

This graphic plots the standard deviation of the transformed data, across samples, against the mean, using the shifted logarithm transformation.

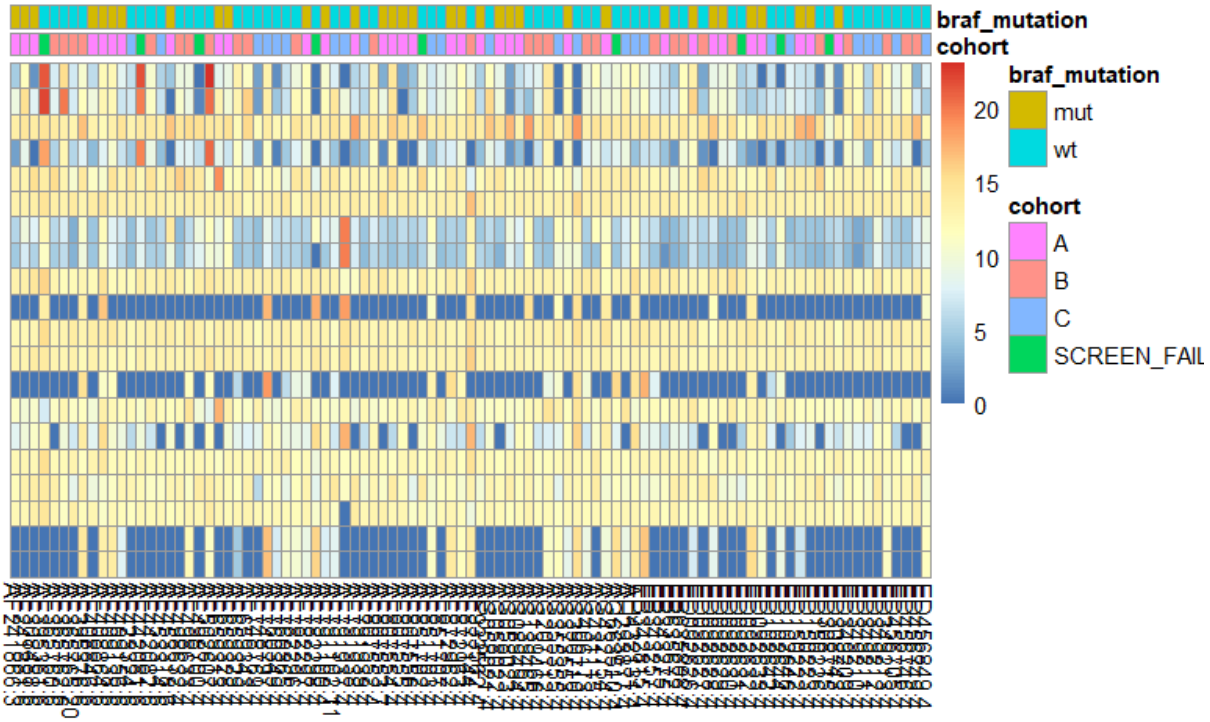


Figure 5: Heat Map

In this graphic we can see the samples classified according to the mutation type and the cohort they belong to. Also we can distinguish their levels of expression by the coloring.

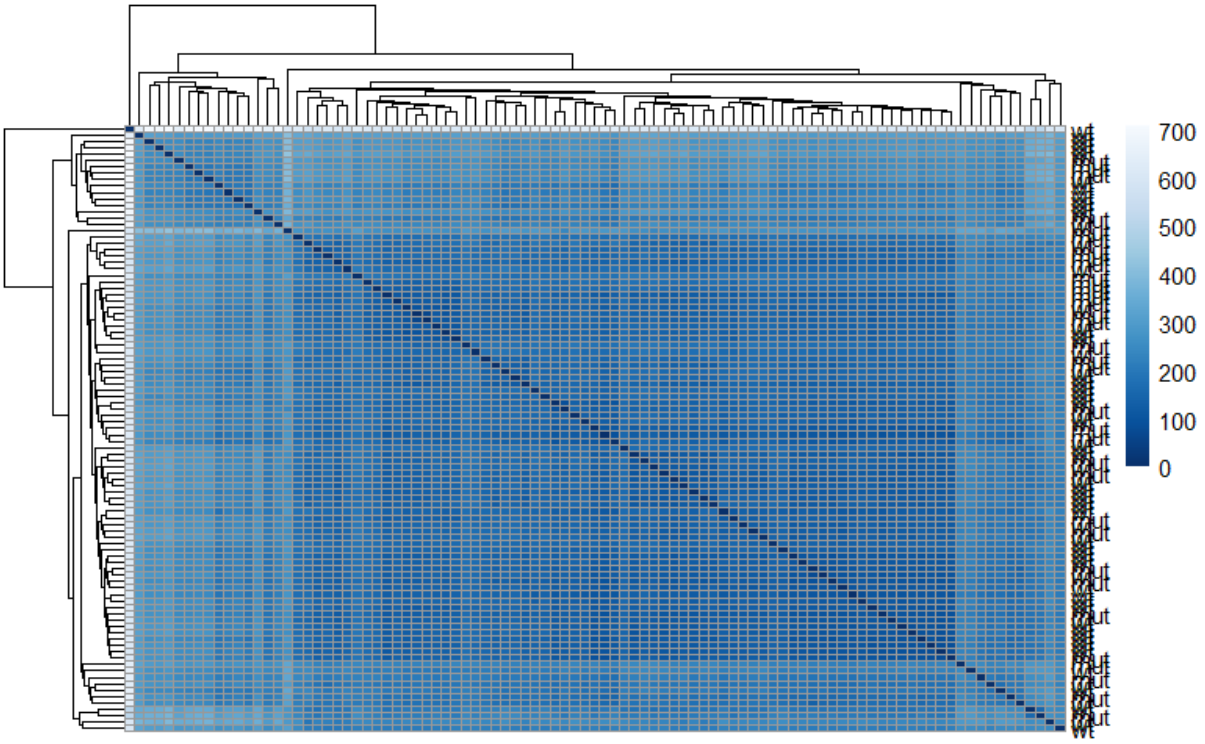


Figure 6: Heat map distances

We have established a transformation to the data (vds) which goes from 0 to 700 counts. There are 3 different groups in regards to the coloring. A heatmap of this distance matrix gives us an overview over similarities and dissimilarities between samples

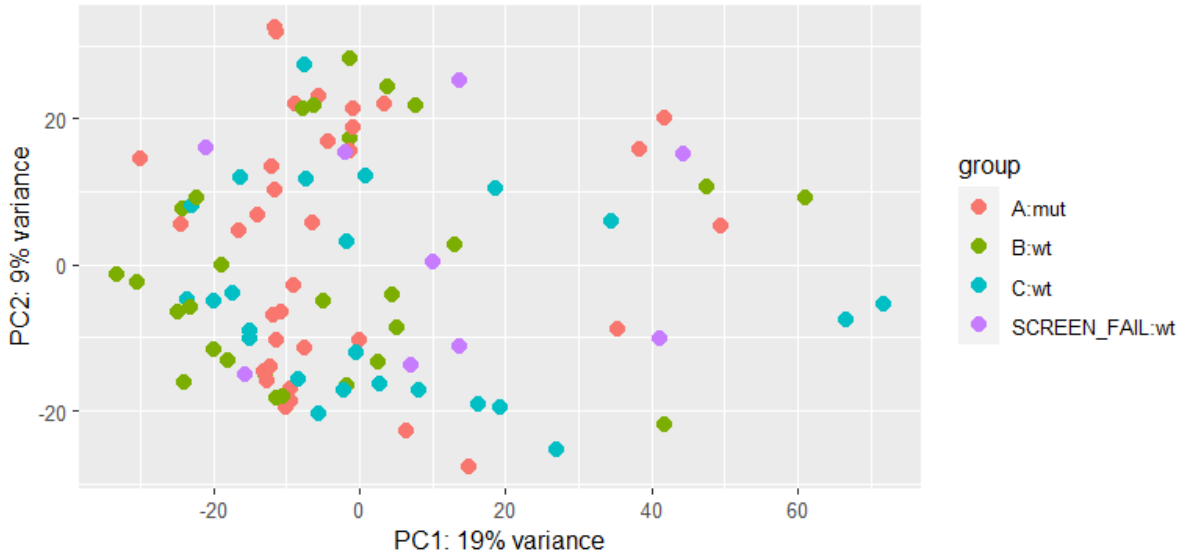


Figure 7: Principal Components Analysis

We can conclude that the BRAF mutation does not have much influence on the gene expression, because the cohort A (has the samples of BRAF mutated melanoma) is not grouped separately from the others.

The principal components represent the dimensions// of the cloud of dots. Each of the components contribute to the variance, in this case, 28%.

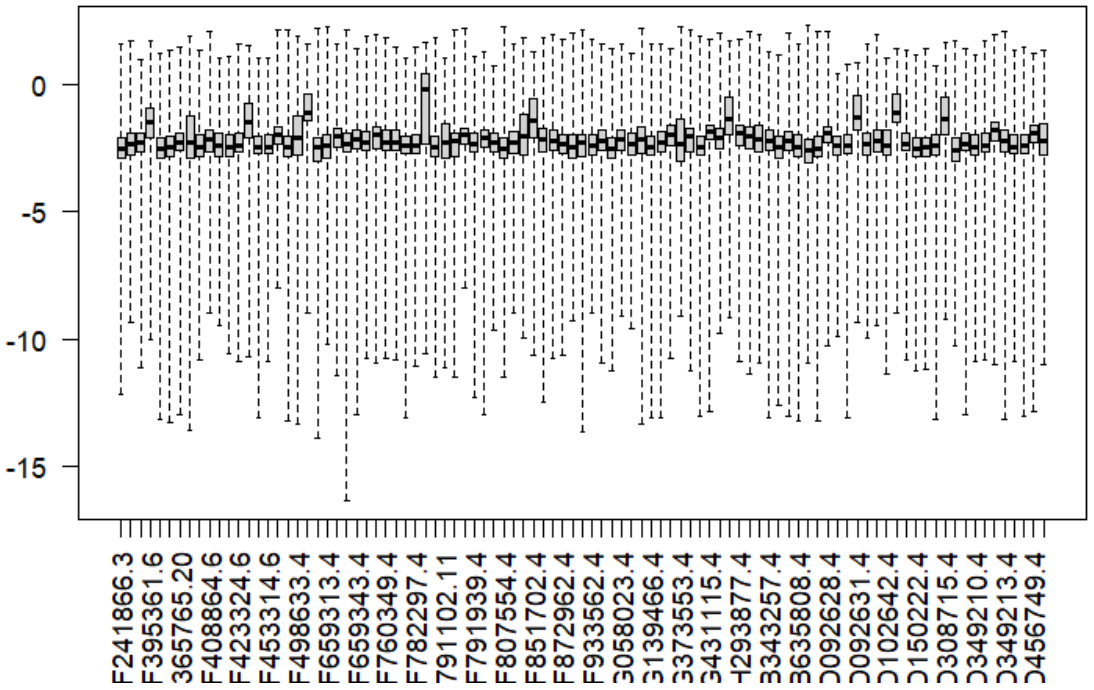


Figure 8: Boxplot

The DESeq function calculates, for every sample, a diagnostic test for outliers called Cook's distance: is a measure of how much a single sample is influencing the fitted coefficients for a gene. A large value of Cook's distance is intended to indicate an outlier count.

Bibliography:

https://warwick.ac.uk/fac/sci/moac/people/students/peter_cock/r/geo/
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2674223/
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE158403
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7721806/
https://medlineplus.gov/genetics/condition/melanoma/#causes