A Multi Omics-Based Analysis of Breast Cancer

Word count: 6 461

Boris Vandemoortele
Tristan Vanneste

Supervisors: Prof. Dr. Ir. Tim De Meyer, Ir. Louis Coussement

Academic year: 2020 - 2021

UNIVERSITEIT GENT

FACULTEIT BIO-INGENIEURSWETENSCHAPPEN

Abstract

The combination of multiple single experiments has the potential to reveal new insights and mechanics in complex diseases. Therefore, we performed a multi omics-based study to gain further insight into breast cancer. We analysed differences in gene expression between healthy and tumour tissue/cells with microarrays and RNA sequencing (RNAseq) and linked this to epigenetic differences (more specifically DNA methylation) with genome-wide methylation profiling. Additionally, we analysed the estrogen receptor cistrome to gain insight into endocrine treatment unresponsiveness in hormone receptor-positive (HR^+) breast cancers.

Introduction

Among women, breast cancer is the most common cause of cancer-related death worldwide in both developing and developed countries[1]. 1.15 million breast cancer cases are diagnosed each year worldwide, making breast cancer one of the most prevalent cancers in the world today[2]. Every year, over 411,000 deaths result from breast cancer, which accounts for over 1.6% of all female deaths worldwide. Breast cancer incidence is rising in developing countries due to increased life expectancy and adaptation of the western lifestyle. Despite the common misconception that the majority of breast cancers are occurring in wealthy countries, most breast cancer-related deaths in fact occur in developing rather than in developed countries. *Ferlay et al* estimated that the global incidence and mortality could even increase in the future[3]. Finding an efficient and cost-effective treatment for breast cancer is thus an urgent unmet medical need.

As in most cancers, early detection and an accurate diagnosis are paramount for disease outcome. Early detection improves prognosis greatly, as physicians estimate that 70-80% of patients with early stage, non-metastatic tumours are curable. In Belgium, women between 50 and 70 years old can get a free mammography every two years to screen for breast cancer. If they belong to a risk group, they are allowed to screen annually and starting from a younger age. Also, a genetic test for the *BRCA1* and *BRCA2* genes, of which inherited mutations are strong indicators of breast and ovarian cancer risk[4], is free for risk groups. Regular screening is however more difficult in developing countries, and cancer is often only detected in a very late and metastatic stage in these countries. The five year survival rate drops to 26% for breast cancers that developed to the metastatic stage[5]. An exact mechanism that drives metastasis is yet to be unravelled, but some important new insights have been gained recently[6]. Next to the occurrence of metastasis, also the cancer subtype is important for disease prognosis. Breast cancer is classically categorized into three major subtypes. These subtypes are based on the presence of two molecular markers, being estrogen or progesterone receptor (ER or PR), which are both hormone receptors (HRs), and human epidermal growth factor 2 (ERBB2). 70% of patients is $HR^+/ERBB2^-$, 15-20% is $HR^-/ERBB2^+$ and 15% is triple negative[7]. The median survival rate is lowest for metastatic triple negative breast cancer.

Distinct cancer subtypes require distinct treatment methods. Patients presenting with HR positive tumours receive endocrine therapy. Tamoxifen, a synthetic ER inhibitor, is today's gold standard of selective ER modulators (SERMs). Its mechanism of action is competitive inhibition of binding of estrogen to the ER[8]. A second possibility is the

administration of aromatase inhibitors such as *Astranazole.* They decrease circulating estrogen levels by inhibiting the conversion of androgens to estrogen by the aromatase enzyme. However, this kind of treatment is only applicable to postmenopausal women[9]. Unresponsiveness to endocrine therapy has been described but the underlying mechanisms are not completely understood yet. We found that the ER cistrome differs genome-wide between samples obtained from tumours that were responsive and non-responsive. Further research into this changing cistrome might help us understand endocrine therapy resistance in the future. $ERBB2^+$ breast cancers are usually treated with ERBB2-targeted antibodies in combination with chemotherapy. The triple negative subtype, which is associated with the highest mortality, is difficult to treat and the only FDA approved treatment is chemotherapy[10].

In conclusion, breast cancer is a disease far from perfectly understood and needs further research to clarify the exact mechanisms that drive the disease. Since cancers are known to completely change the genomic, transcriptomic and epigenomic landscape of cells, is an omics-based approach trivial to better understand the cancer in case. We, therefore, aim to identify differences in gene expression between healthy tissue/cells and breast cancer tissue/cells, by analysing a microarray experiment and an RNAseq experiment. As the transformation of healthy cells to cancer cells is often accompanied by epigenetic changes, a genome-wide methylation profiling will be performed as well. If possible, differential methylation of genes will be linked to differential gene expression. Additionally, the ER cistrome was analysed to gain insight into endocrine treatment unresponsiveness in HR^+ breast cancers.

Materials and methods

Microarray transcription profiling

The microarray transcription profiling data was obtained from ArrayExpress[11] under the ArrayExpress experiment identifier E-GEOD-15852[12] in the raw format. 43 paired samples were collected for both tumour and normal tissues from breast cancer patients. In total 86 samples were analysed for gene expression by using an Affymetrix genechip U133A. Explorative Quality Control (QC) was performed on the normal data and the log-transformed data with the *arrayQualityMetrics* package. Background correction as well as quantile normalization were performed on the raw data with Robust Multi-array Average (RMA). A design matrix was made for the disease with a blocking factor for patients. Finally, differential expression (DE) was detected using *limma*[13]. Adjusted p-values were computed according to the Benjamini-Hochberg procedure, with a threshold of 0.05 for marking significant probes. Specific gene annotations were obtained using the BioMart query function.

Methylation profiling by array

Genome-wide profiling of DNA methylation in 4 pairs of matched tumour tissue and normal breast tissue were obtained from the Gene Expression Omnibus (GEO)[14] data repository under the identifier GSE101443[15]. These 8 samples were analysed with an Illumina HumanMethylation450 BeadChip. Probes for which called p-values were insufficient were filtered out together with NA values. An initial analysis was performed on the average methylation percentages for tumour tissue vs control tissue with a

Welch t-test. The raw data was normalized using the *dasen* function in R. MethyLumiM objects were made from the normalized data and used for further analysis. Differential methylation (DM) analysis was also performed with *limma* and the adjusted p-values were computed according to the Benjamini-Hochberg procedure with an alpha threshold of 0.05 for significant probes. The design matrix was constructed for the factor condition and included a blocking design for the factor patient. Finally, the annotation was obtained using the package *ChAMPdata* in R. Significantly DM genes, with a more lax cut-off (alpha = 0.10), and significantly DE genes were analysed using Enrichr, which uses a statistic test based on the Fisher exact test (FET) to calculate pathway and gene ontology enrichment[16].

RNAseq

Datasets were obtained from the *European Nucleotide Archive (ENA)*[17] under project ID PRJNA142887[18]. This dataset RNA comprises sequencing results from both normal breast cells (HMEC) and breast cancer cells (HCC1954). Quality control was performed using fastqc[19] and overrepresented sequences were removed using Trimmomatic's[20] paired-end Illuminaclip function. Seed mismatches were set to two, palindrome clip threshold was 30 and simple clip threshold was 10. The human hg38 reference genome and genome annotation were downloaded from the ensemble genome browser and a pseudotranscriptome was built using Kallisto[21], after which a genome index was made using standard settings. Counts per transcript were summarized to counts per gene using the *scaledTPM* option from *tximport* and then normalized using the trimmed mean of M-values (TMM) method in *EdgeR*[22]. Since no biological replicates were present, we used a user-defined dispersion value of 0.4. to identify DE genes. Threshold values for DE genes were set at an absolute log fold change (LFC) > 2 and false discovery rate (FDR) < 0.05. Finally, the Enrichr web application was used for gene ontology and pathway analysis.

ChIPseq

Datasets were obtained from the *ENA* under project ID PRJNA175144[23]. Only datasets from ER samples were downloaded. Fastqc was used to perform QC, after which overrepresented sequences were removed using Trimmomatic's Illuminaclip. A genome index was made using Bowtie2[24] and the same human reference genome was used as described in the RNAseq experiment. Aligning reads to the hg38 genome was performed with default settings. A small output formatting step was required to make this data compatible with our further analysis in HOMER[25]. Tag directories were made and peaks were called using HOMER and its inbuild hg38 reference genome. We used a minimum fold enrichment over the input of 8 for peak calling. Only uniquely aligned reads were preserved. Peaks from replicates were merged so that all peaks appearing in at least one replicate were retained and the overlap between genes associated to those peaks was visualized by the VIB Venn diagram tool. Gene ontology enrichment of these genes was performed in Enrichr and enriched motifs in these peaks were identified using standard settings in HOMER.

Results

A substantial number of genes were found to be differentially expressed in tumour tissue

After loading the data from ArrayExpress, we performed multiple QC on the raw data, log-transformed data and normalized data. We used the *RMA expression measure* function in R to perform background correction and quantile normalization. We found 5333 probes to be statistically significant after correcting for multiple testing with the Benjamini-Hochberg procedure ($\alpha < 0.05$) (figure 1, A). However, we have to be careful when interpreting that many statistically significant probes since we do not expect that many biologically relevant results. If we look at statistically significant probes and their LFCs, we see that most LFCs are rather low. We choose to filter out those probes that had an absolute LFC < 1 to only keep those probes for which DE is likely to be biologically relevant. After filtering for biologically relevant probes and probes with gene annotation we found 39 significantly DE genes (FDR <0.05 and an absolute LFC > 1.0). 9 genes had a positive LFC and 30 genes had a negative LFC. A positive LFC stands for higher expression in the tumour tissue than in the healthy (normal) tissue. When we did not filter for the LFC, there were 4069 genes (after removing duplicates) found to be statistically significant, of which 2621 genes were significantly upregulated and 1495 genes were significantly downregulated.

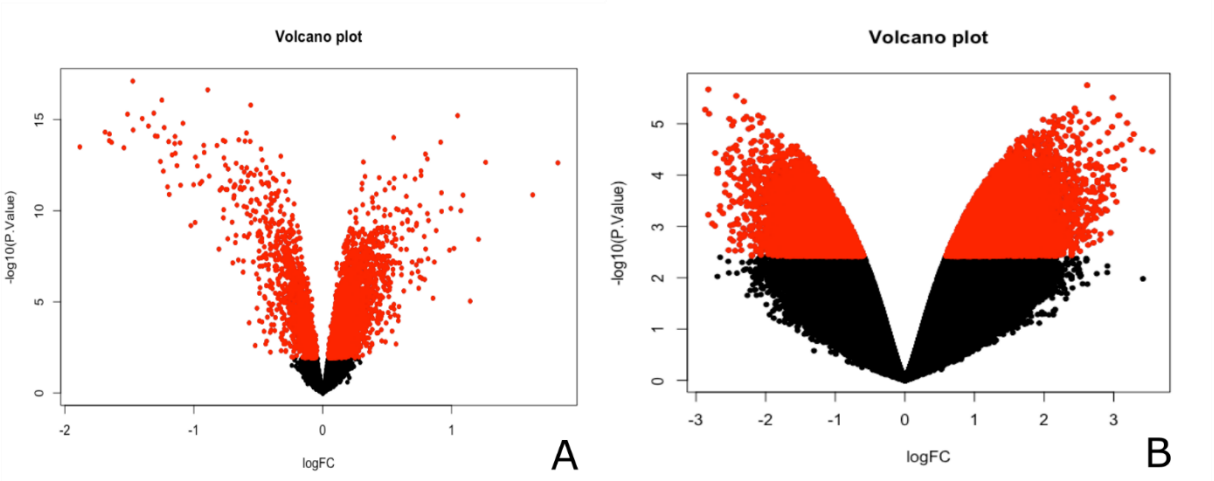

Figure 1: Volcano plots from array expression and methylation profiling of breast cancer vs normal tissue. *A: Volcano plot showing the results of expression profiling. Red: probes that are statistically significant for DE analysis (FDR < 0.05) with a blocking design for patients (n = 43) using limma. We found 5333 significant probes; B: Volcano plot from methylation profiling. Red: probes that are statistically significant for the differential methylation analysis (FDR < 0.10) with a blocking design for patients (n = 4) using limma. We found 18540 significant probes.*

Using a more lax FDR cut off for differential methylation analysis leads to many significant hits

After filtering out probes that had too little counts and removing NA values, we found zero probes to be statistically significant after correcting for multiple testing with the Benjamini-Hochberg procedure at an FDR level of 0.05. For further analysis, we used an FDR level cut off from 0.10 to compare our results across different methods. We found 18540 probes to be statistically significant at the FDR level of 0.10 using *limma*

(Figure 1, B). After removing duplicates and probes that did not have a gene annotation, there were 5917 probes left. 3862 of those 5917 probes had a positive LFC and 2744 had a negative LFC. Since we were only interested in the biologically most relevant probes, we filtered out those probes that had an absolute (LFC) < 2. From the 5917 significant probes, 622 probes had an absolute LFC larger than 2. After removing duplicates, 250 probes were found that had a positive LFC and 104 genes had a negative LFC. Note that a positive LFC matches with higher methylation in tumour tissues compared to normal tissue, but the biological role of this epigenetic difference is often not straightforward.

HCC1954 cells change their expression pattern to mediate cell survival.

Finding differentially expressed genes between healthy HMEC cells and cancerous HCC1954 cells was difficult since no biological replicates were available in this study. Results should thus be critically evaluated before making any conclusions. When using a fixed estimated dispersion value of 0.4, 239 genes were found to be significantly upregulated and 184 genes were significantly downregulated (FDR < 0.05 and an absolute LFC > 2). We then examined to which cellular processes and functions these genes were associated using Enrichr. Table 1 contains an overview of some selected hits that we believe to be interesting.

Table 1: Gene Ontology (GO) analysis of differentially expressed genes in HCC1954 cells vs HMEC cells. *239 and 184 genes were respectively up- and downregulated in HCC1954 cells compared to HMEC cells (n=2, FDR < 0.05 and absolute LFC > 2). Shown are the interesting hits of GO analysis in Enrichr with respective genes belonging to the given categories and associated p-values of enrichment (modulated FET).*

GO category	associated genes	p-value
	Upregulated genes	
Glucuronidation	UGT1A10, UGT1A1, UGT1A5, UGT1A4, UGT1A3, UGT1A7	5.154E-7
Positive regulation of monocyte chemotaxis	CXCL10, CCL5, CXCL17, S100A7	4.188E-5
	Downregulated genes	
Cytokines and inflammatory response	IL1A, CSF2, CXCL1, TNF	8.847E-5
TNF signalling	CSF2, VEGFC, CXCL1, CXCL3, TNF	3.507E-3
Estrogen signalling pathway	KRT27, KRT16, KRT14, CALML3, HBEGF	8.786E-3
Regulation of cell proliferation	FGFBP1, TNFRSF6B, IRS1, EIF5A2, PINX1, VEGFC, IRS2, CXCL1, CXCL3, EREG, TNFRSF10D, FOSL1, IL1A, CCND2, IGFBP6, SOX7, HBEGF	5.00E-5

We believe our results show that HCC1954 cells are reprogrammed towards a more proliferating and immunosuppressive state, which is in line with cancer progression. TNF signalling for example, which is classically known as a potent mediator of cell death in cancer [26], was found to be downregulated in HCC1954 cells. We also found that other inflammatory response genes were significantly enriched in the list of downregulated genes (Table 1). The reduced inflammatory signalling we observe might protect cancer cells from the body's immune system. However, we also found that upregulated genes were enriched for positive regulation of chemotaxis. The role of inflammation in cancer is not straightforward, since it can both positively and negatively affect tumour growth. The stage of cancer development is often a definitive

factor to determine whether the presence of an inflammatory (micro-)environment will either positively or negatively affect tumour growth[27]. Finally, the list of downregulated genes was enriched for genes associated with the regulation of cell proliferation. We thus show that HCC1954 cells maximize their survival chances by increasing cell proliferation and limiting inflammation.

A multi omics-based analysis of the expression from healthy cells/tissue vs cancerous cells/tissue

By combining the results from the methods discussed above, one can get a multi omics-based approach. We separated those genes that we found to be either up- or downregulated in breast cancer through expression profiling (both via array and RNAseq) since they have different biological meanings. We compared the upregulated genes from our microarray and RNAseq analysis to all significant genes from the DM analysis. We choose to use all significant genes from our DM analysis because a higher degree of methylation can both have a positive or negative influence on gene expression, often depending on the exact position of methylated cytosine residues in the gene [28]. We also decided to compare all statistically significant genes for each analysis, without setting an LFC threshold, since it is less relevant to do a comparison with a small number of genes. We found that only 26 genes were upregulated in both the array and RNAseq experiment, of which seven genes were also differentially methylated (figure 2, A). The overlap between downregulated genes was even smaller with only 20 genes overlapping between our array and RNAseq experiment, of which five genes were also differentially methylated (figure 2, B). There are several reasons why this degree of overlap might be so low. First of all, the array experiment was performed on whole tissue samples, while the RNAseq experiment was performed on a cell line. Whole tissues contain multiple cell types, of which each has its own gene expression profile. Thus, next to the expression of cancer cells, also the expression profile of other cells such as immune cells will be measured. Also, no biological replicates were present in the RNAseq study.

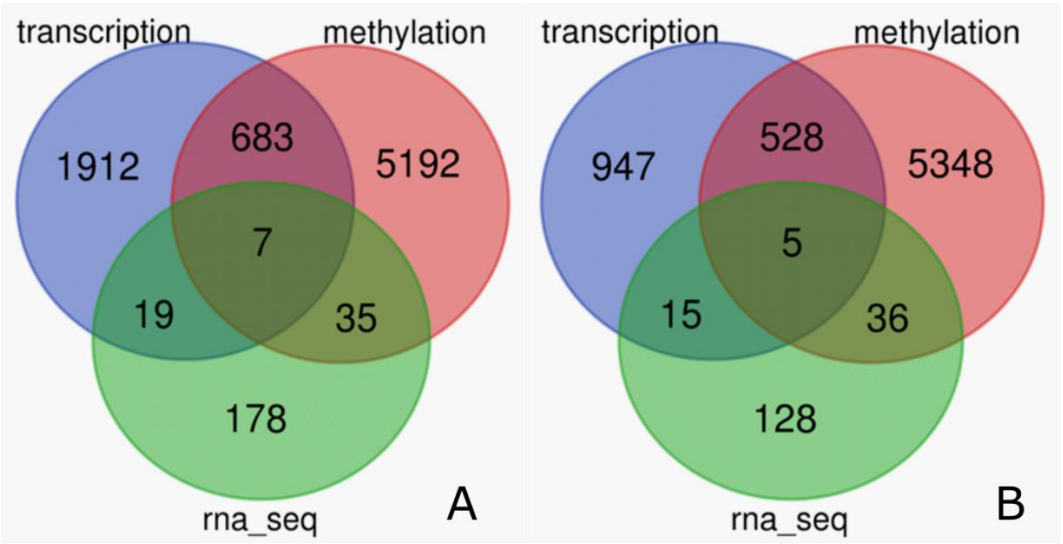

Figure 2: Overlap between expression and methylation profiling experiments. *A: Venn diagram showing overlap between genes that were upregulated in RNAseq (LFC > 0, FDR < 0.05) and array expression profiling (LFC > 0, FDR < 0.05) with genes associated to differentially methylated probes (FDR < 0.10). B: Venn diagram showing overlap between genes that were downregulated in RNAseq (LFC < 0, FDR < 0.05) and array expression profiling (LFC < 0, FDR < 0.05) with genes associated to differentially methylated probes (FDR < 0.10).*

Using Enrichr, we examined to which functions and cellular processes the overlapping genes between our DE and DM analysis were associated. This GO analysis was done without using the genes from the RNAseq analysis since we argued that it is better to do the GO analysis on a larger number of overlapping genes. Table 2 contains an overview of the significant hits. Significantly enriched functions or pathways of overlapping genes are mostly correlated with the regulation of apoptosis, cell death and cell differentiation. These functions are as expected and hint towards some of the molecular changes occurring in tumour cells, e.g., delaying cell death and apoptosis.

***Table 2: GO analysis of the overlap of statistically significant genes between the differential transcription and methylation analysis**. 690 and 533 genes were respectively up- and downregulated in tumour tissues compared to normal tissues. Shown are the interesting hits of GO analysis in Enrichr with respective genes belonging to the given categories and associated p-values of enrichment (modulated FET, significance thresholds were set as described in the methods section, without LFC cut off).*

GO category	associated genes	p-value
	Upregulated genes	
Regulation of apoptotic process	FLT1, FAIM2, NDRG1, EGFR, FOXO1, MALT1, GNA13, WNT11, OPA1, GRK5, TCTN3, TRIM2, NUP62, KDR, FYN, MAP4K4, MAP3K3, BNIP3L, GRID2, TGFB1, ANGPT1, GADD45B, DUSP1, MGMT, DAPK2, BNIP3, HIP1R, LILRB1, IGF1, CIDEC, TNFRSF1A, GCLC, IL6, BCL6, BIN1, DDAH2, BCL2, SPRY2, TEK, NF2, ARHGEF7, MET, DDR2, FGFR1	3.4097009900666743E-6
Negative regulation of apoptotic process	FLT1, FAIM2, EGFR, FOXO1, MALT1, WNT11, OPA1, GRK5, NUP62, KDR, MAP4K4, BNIP3L, ANGPT1, MGMT, CAV1, BNIP3, PRKCA, IGF1, DAB2, GCLC, IL6, DDAH2, BCL2, SPRY2, TEK, MET, DDR2, FGFR1	1.1782632329780712E-4
Negative regulation of programmed cell death	KANK2, FLT1, FAIM2, EGFR, FOXO1, MALT1, WNT11, MECOM, OPA1, GRK5, NUP62, KDR, MAP4K4, BNIP3L, ANGPT1, MGMT, BNIP3, IGF1, DAB2, GCLC, IL6, DDAH2, BCL2, SPRY2, TEK, MET, DDR2, FGFR1	5.3679903028852865E-6
	Downregulated genes	
FSH regulation of apoptosis	CDKN1C, COL15A1, PDXK, PTGER2, CREM, CREBL2, COX7A1, HK2, HSD11B1, AKAP12, RGS5, ABLIM1, GRK5, GNLY, LEPR, CD36, FLNC, IGFBP6, ATP9A, MAP4K4, BNIP3L, ANGPT1, DUSP1, ARPC4, MAPK14, TGFBR2, GPRC5B, DAB2, ZEB1, S100A4	2.2113980539716847E-11
Positive regulation of cell differentiation	TGFB1, TGFB1I1, RARRES2, ZBTB16, CTNNBIP1, LRP5, CREBL2, IGF1, MAPK14, TGFBR2, TMEM100, IL6, DAB2, NEUROD4, KDR, PPARG, CD36, APOB, DDR2	1.154058829433845E-6

Also, a total amount of 2100 genes were found in the list of statistically significant DM genes when selecting for genes linked to the promoter or the first exon region (features ‘TSS’ and ‘1stExon’). As a higher degree of methylation in the promoter region of a gene is often linked to transcriptional silencing, we went looking for the amount of overlap between up- and downregulated genes and genes with either a higher or lower degree of methylation in their promoter. We saw that out of 188 genes that had a higher degree of methylation in their promoter in tumour tissue, 125 genes had a lower expression in tumour tissue compared to normal tissue. Also, 119 out of 168 genes with a lower degree of methylation in their promoter in tumour tissue, had a higher expression in cancer tissue compared to normal tissue. This suggests that epigenetic changes in cells can indeed affect gene expression and drive their transformation into cancer cells.

We also did a GO analysis on those 125 and 119 overlapping genes that had either a lower or higher expression respectively in the tumour tissue. Some interesting results were found for the genes that had a lower expression as well as a higher degree of methylation in their promoters in tumour tissue. These genes were mostly linked with positive regulation of cell differentiation while the genes which had a higher expression in tumours were correlated with the regulation of transcription and again apoptotic processes.

Table 3: Overlap between genes which are DM in their promoter or first exon and DE genes (microarray). *As expected, methylation in the promoter region of a gene is often correlated to decreased expression of that respective gene. Out of 188 genes which had increased methylation in cancer tissue, 125 had a decreased expression in cancer tissue compared to normal tissue. Finally, of the 168 genes that had lower methylation in the promoter of the tumour tissue, 119 had an increased expression in cancer tissue compared to the normal tissue.*

	HIGHER EXPRESSION IN TUMOUR	**LOWER EXPRESSION IN TUMOUR**
HIGHER METHYLATION IN TUMOUR PROMOTER	63 overlapping genes	125 overlapping genes
LOWER METHYLATION IN TUMOUR PROMOTER	119 overlapping genes	49 overlapping genes

ER signalling seems to be intact in the highly diminished ER cistrome of aromatase treatment unresponsive tumours

A common cause of breast cancer is the uncontrolled proliferation of estrogen receptor-positive breast cells. Next to today's gold standard tamoxifen, new SERMs are being developed. One of those is *Astranazole,* an aromatase inhibitor. This class of molecules aims to block the conversion of androgen into estrogen performed by the enzyme aromatase[29]. Our dataset contained 2 groups of patients (n=10), one of which was responsive to aromatase inhibitor treatment and one who did not respond. After ChIPseq targeted against the ER, peaks were called using HOMER. After merging peaks within each group, we found 37372 peaks in the group that responded to aromatase treatment. Only 6854 peaks were found in the other group, which indicates that the ER has a severely compromised DNA binding capacity in this group. To further explore this phenomenon, we annotated peaks to the nearest gene and studied the compositions of these gene sets. 4034 genes were associated with peaks in both groups, while 10350 genes were only found in the responsive group and 242 genes were only found in the non-responsive group (figure 3, A). GO analysis revealed that genes involved in the nuclear receptor transcription pathway were enriched (p-value 0.003347) in the peak subset that was only observed in samples from non-responsive patients. This observation reflects the mode of action of aromatase inhibitors, namely blocking ER signalling, as ER signalling only remained functional in the non-responsive group after administration of the drug.

Motif analysis revealed a higher enrichment of ER elements (EREs) (figure 3, B) in peaks found in responsive samples compared to in peaks found in non-responsive samples, with 14,50% and 9,44% of target sequences containing the motif respectively. This supports our previous observations that ER DNA binding might be impaired in non-responsive samples. Strangely, the top motifs that are enriched in both conditions seem to be Fox-like (figure 3, C) motifs, while only a limited similarity between Fox-like motifs and ERE motifs seems to be present.

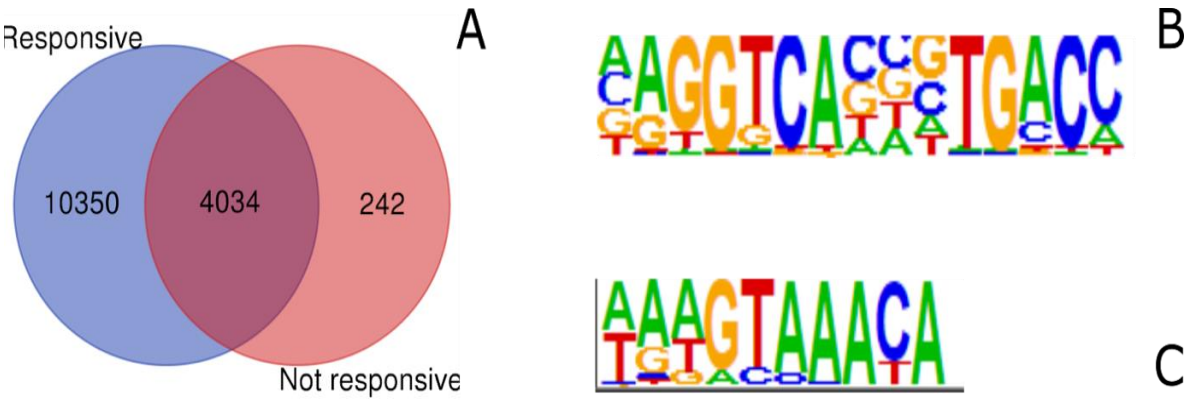

Figure 3: ChIP sequencing reveals differences in ER DNA binding pattern between aromatase inhibitor responsive and non-responsive tumours. *A: Venn diagram showing the overlap between genes associated with peaks found in aromatase inhibitor responsive vs non-responsive tumours. Peaks were called by HOMER using a minimal fold enrichment over background threshold of 8 and peaks occurring in at least one replicate were merged within each group (n = 10). 4034 were found in both groups, while only 242 genes were uniquely found in the group non-responsive to aromatase inhibitor treatment. 10350 genes were associated with peaks that were only found in the group responsive to aromatase treatment, B: ERE motif identified by HOMER motif analysis, C: Fox1a motif identified by HOMER, this motif is unexpectedly the top hit in both conditions.*

ChIP sequencing directed against the ER thus revealed a decreased ER DNA binding capacity in tumour samples that did not respond to aromatase treatment compared to samples that did, which is reflected by a lower number of peaks and a lower enrichment of ERE motifs in these peaks. Genes associated with peaks that are solely found in non-responsive samples are related to ER signalling, which suggests that ER signalling is still functional even after inhibition of estrogen production.

Discussion

Breast cancer affects millions of women each year and is the cause of over 1.6% of the total amount of female deaths annually. The number of yearly diagnosed cases is estimated to keep rising and this is especially a problem in developing countries where regular screening is not yet possible. Finding new cost-effective treatments for breast cancer is thus an urgent unmet medical need. We aimed to gain further insight into the complex mechanics of breast cancer by applying a multi omics-based data analysis approach.

In general, we analysed two microarrays that contained data from cancer and healthy tissue, one RNAseq experiment containing expression data from a cancer cell line and a ChIPseq experiment targeted against the ER. After analysis of the array expression

data, we found that 5333 probes were significantly DE. However, when applying a minimal absolute LFC filter of 1 to screen for likely biological relevance and annotating for the correct gene, only 39 probes/genes remained. The methylation profiling array revealed that 5917 genes were DM in cancer tissue compared to healthy tissue (FDR < 0.10). Keeping in mind that using a less strict FDR threshold will lead to a larger number of false positive hits. Of these genes, 622 had an absolute LFC > 2. Using Enrichr, we found some fascinating results regarding the functions of overlapping genes from both the DE and DM analysis. For example, we found that genes that were both upregulated and DM were enriched for functions in the negative regulation of cell death. Since a higher degree of methylation in a gene's promoter is often associated with a decreased expression, we performed some additional research into these overlapping genes. We found that out of 188 genes with a higher degree of methylation in their promoter, 125 indeed had a lower expression in cancer tissue. A similar but opposite trend was observed when focussing on genes with a lower degree of methylation in their promoter. These observations might reflect some of the mechanisms at play in the transformation from healthy cells into cancer cells, as epigenetic changes are often put forward as drivers of this process[30]. Of course, these results were obtained from two different experiments, so a degree of caution is needed when comparing these. In addition, we only looked at DM for the epigenetic changes but there are several more, for example, histone modification to look into.

In the RNAseq experiment, we found that 423 genes in total were DE. The GO categories associated with these genes reflected some mechanisms which are likely to be important drivers for tumour growth (e.g., decreased TNF signalling and less regulation of cell proliferation). However, when comparing results from the expression profiling array to the RNAseq experiment, we found surprisingly little overlap. Only 19 and 15 genes were found to be up- or downregulated respectively in both analyses. Several reasons might lie at the basis of this phenomenon. One major factor of impact is the sample that was used for the analysis itself. The array expression profiling experiment was performed on a whole tissue sample, whereas the RNAseq experiment was performed on a cell line. Trivially, a whole tissue sample contains multiple cell types. Proliferating tumours often have high energy demands, which must be compensated by increased delivery of oxygen and nutrients to the tissue[31]. This is often mediated by the occurrence of hypoxia, which in its turn leads to the production of VEGF and the formation of additional blood vessels in the tumorous tissue[32]. Additionally, the body's immune system can detect the tumour as a malignant structure and try to attack it. This is characterised by the infiltration of monocytes and other immune cells into the tumour[33]. Of course, these cells will also contribute to the overall expression profile of tumour tissues. A second important factor is the fact that the RNAseq experiment did not contain any biological replicates. If replicates are available, EdgeR will apply an empirical Bayes procedure to squeeze tag/gene-wise dispersion values towards a global or common dispersion value. However, no biological replicates were available in this study. Therefore, a fixed dispersion value of 0.4 was chosen, as suggested in the EdgeR manual. Consequently, every gene will have an equal estimated dispersion value, which is of course not an accurate reflection of the biological situation and will in its turn lead to less accurate results. Lastly, the cell line used in this RNAseq experiment only reflects one type of breast cancer, namely HR negative breast cancer. Samples from the array expression profiling experiment might be originating from a different type of breast cancer altogether, but we did not find any additional information on this subject in the original paper[12].

The ChIPseq experiment revealed a clearly distinct ER cistrome between tumours that were responsive to aromatase inhibitor treatment and samples obtained from tumours that were not responsive. GO analysis of genes associated with peaks that were only present in the non-responsive group suggested that ER signalling was still active after treatment and might thus still drive tumour growth. One possible reason for this might be that the ER of tumours in this group carries a mutation which renders it constitutively active, even when no ligand (estrogen) is bound to the receptor. However, motif analysis returned a fox-like motif as the top hit. As only a limited similarity between EREs and the fox-like motif seems to be present and the presence of certain unexpected motifs can indicate a lower quality of the experiment, results should be interpreted with caution.

In conclusion, we found an overlap of only 12 statistically significant genes when comparing the expression profiling by array, methylation profiling by array and RNAseq experiment. This is rather low compared to the number of significant genes reported by each method itself. We argue that the low number of overlapping genes is partly explained by the low power of the RNAseq experiment (no biological replicates) and the fact that the RNAseq is done on a cell line while the other analyses were performed on tissue samples. Also, since the experiments were not done on the exact same patients, biological variation is still not fully excluded from the analyses. A better design could potentially lead to more relevant results.

References

1. Anderson, B. O. *et al.* Guideline implementation for breast healthcare in low-income and middle-income countries: Overview of the breast health global initiative Global Summit 2007. in *Cancer* (2008). doi:10.1002/cncr.23844.
2. Parkin, D. M., Bray, F., Ferlay, J. & Pisani, P. Global Cancer Statistics, 2002. *CA. Cancer J. Clin.* (2005) doi:10.3322/canjclin.55.2.74.
3. Ferlay J, Bray F, Pisani P, et al. Cancer Incidence, Mortality and Prevalence Worldwide. *GLOBOCAN* (2002).
4. Fackenthal, J. D. & Olopade, O. I. Breast cancer risk associated with BRCA1 and BRCA2 in diverse populations. *Nature Reviews Cancer* (2007) doi:10.1038/nrc2054.
5. Peart, O. Metastatic breast cancer. *Radiol. Technol.* (2017).
6. Krøigård, A. B. *et al.* Identification of metastasis driver genes by massive parallel sequencing of successive steps of breast cancer progression. *PLoS One* (2018) doi:10.1371/journal.pone.0189887.
7. Waks, A. G. & Winer, E. P. Breast Cancer Treatment: A Review. *JAMA - Journal of the American Medical Association* (2019) doi:10.1001/jama.2018.19323.
8. Nazarali, S. A. & Narod, S. A. Tamoxifen for women at high risk of breast cancer. *Breast Cancer: Targets and Therapy* (2014) doi:10.2147/BCTT.S43763.
9. Schneider, R. E., Barakat, A., Pippen, J. & Osborne, C. Aromatase inhibitors in the treatment of breast cancer in post-menopausal female patients: An update. *Breast Cancer Targets Ther.* (2011) doi:10.2147/BCTT.S22905.
10. Collignon, J., Lousberg, L., Schroeder, H. & Jerusalem, G. Triple-negative breast cancer: Treatment challenges and solutions. *Breast Cancer: Targets and Therapy* (2016) doi:10.2147/BCTT.S69488.
11. Athar, A. *et al.* ArrayExpress update - From bulk to single-cell expression data. *Nucleic Acids Res.* (2019) doi:10.1093/nar/gky964.
12. Pau Ni, I. B. *et al.* Gene expression patterns distinguish breast carcinomas from normal breast tissues: The Malaysian context. *Pathol. Res. Pract.* (2010) doi:10.1016/j.prp.2009.11.006.
13. Ritchie, M. E. *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* (2015) doi:10.1093/nar/gkv007.
14. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* (2002) doi:10.1093/nar/30.1.207.
15. Collignon, E. *et al.* Immunity drives TET1 regulation in cancer through NF-kB. *Sci. Adv.* (2018) doi:10.1126/sciadv.aap7309.
16. Chen, E. Y. *et al.* Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* (2013) doi:10.1186/1471-2105-14-128.
17. Leinonen, R. *et al.* The European nucleotide archive. *Nucleic Acids Res.* (2011) doi:10.1093/nar/gkq967.
18. Hon, G. C. *et al.* Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome Res.* (2012) doi:10.1101/gr.125872.111.
19. Andrews, S. FASTQC A Quality Control tool for High Throughput Sequence Data. *Babraham Inst.* (2015).
20. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina

sequence data. *Bioinformatics* (2014) doi:10.1093/bioinformatics/btu170.
21. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* (2016) doi:10.1038/nbt.3519.
22. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* (2009) doi:10.1093/bioinformatics/btp616.
23. Jansen, M. P. H. M. *et al.* Hallmarks of aromatase inhibitor drug resistance revealed by epigenetic profiling in breast cancer. *Cancer Res.* (2013) doi:10.1158/0008-5472.CAN-13-0704.
24. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* (2012) doi:10.1038/nmeth.1923.
25. Benner, C., Heinz, S. & Glass, C. K. HOMER - Software for motif discovery and next generation sequencing analysis. *Http://Homer.Ucsd.Edu/* (2017).
26. O'Malley, W. E., Achinstein, B. & Shear, M. J. Action of bacterial polysaccharide on tumors. ii. damage of sarcoma 37 by serum of mice treated with serratia marcescens polysaccharide, and induced tolerance. *J. Natl. Cancer Inst.* (1962) doi:10.1093/jnci/29.6.1169.
27. de Visser, K. E. & Coussens, L. M. The inflammatory tumor microenvironment and its impact on cancer development. *Contributions to microbiology* (2006) doi:10.1159/000092969.
28. Tirado-Magallanes, R., Rebbani, K., Lim, R., Pradhan, S. & Benoukraf, T. Whole genome DNA methylation: Beyond genes silencing. *Oncotarget* (2017) doi:10.18632/oncotarget.13562.
29. Avvaru, S. P. *et al.* Aromatase Inhibitors Evolution as Potential Class of Drugs in the Treatment of Postmenopausal Breast Cancer Women. *Mini-Reviews Med. Chem.* (2018) doi:10.2174/1389557517666171101100902.
30. Futscher, B. W. Epigenetic Changes During Cell Transformation. in (2013). doi:10.1007/978-1-4419-9967-2_9.
31. Muz, B., de la Puente, P., Azab, F. & Azab, A. K. The role of hypoxia in cancer progression, angiogenesis, metastasis, and resistance to therapy. *Hypoxia* (2015) doi:10.2147/hp.s93413.
32. Carmeliet, P. VEGF as a key mediator of angiogenesis in cancer. *Oncology* (2005) doi:10.1159/000088478.
33. Laviron, M., Combadière, C. & Boissonnas, A. Tracking monocytes and macrophages in tumors with live imaging. *Frontiers in Immunology* (2019) doi:10.3389/fimmu.2019.01201.

Addendum

Transcription profiling microarray

```r
#set working directory
#https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-15852/
setwd("~/Documents/Bioinformatics/Applied high-throughput analysis/project_AHTA/Transcription profiling of human breast tumors and their paired normal tissues")

## Load packages
library(affy)
library(arrayQualityMetrics)
library(ArrayExpress)
library(limma)
library(biomaRt)

## Import Data
####################

## Load in the data
## Download data to your working directory
getAE("E-GEOD-15852", type = 'raw')

# load in the expressionFeatureSet object
BreastCancer <- ArrayExpress("E-GEOD-15852")

## Reads in all .cel files and takes phenoData from the ExpressionFeatureset we loaded using https_ArrayExpress
BreastCancer <- ReadAffy(phenoData=pData(BreastCancer))
BreastCancer <- ReadAffy() # If ArrayExpress does not work => use this

dim(pData(BreastCancer))
exprs(BreastCancer)
pData(BreastCancer)

```

Figure 4: Loading in the data

```r
# Creating Annotation matrix

disease <- NULL
patients <- rep(1:43, each = 2)


for (i in pData(BreastCancer)$sample){
  if((i %% 2) == 0) {
    disease[i] <- 'Cancer'}
  else{
    disease[i] <- 'Normal'}

}
disease
patients
annotation <- data.frame(disease,patients)
annotation
dim(exprs(BreastCancer))

```

Figure 5: Making the annotation dataframe

```
49
50 ## Quality Control on raw data
51 ####################
52 # raw data
53 arrayQualityMetrics(BreastCancer,outdir="~/Documents/Bioinformatics/Applied high-throughput analysis/project_AHTA/
54                     Transcription profiling of human breast tumors and their paired normal tissues/raw",force=T)
55 # logtransformed data
56 arrayQualityMetrics(BreastCancer,outdir="~/Documents/Bioinformatics/Applied high-throughput analysis/project_AHTA/
57                     Transcription profiling of human breast tumors and their paired normal tissues/rawlog",force=T,do.logtransform=T)
58
59 # Preprocessing of the data
60 BreastCancerRMA<- affy::rma(BreastCancer,background=T)
61
62
63 ## Quality Control on preprocessed data
64 ## QC post preprocessing
65 arrayQualityMetrics(BreastCancerRMA,outdir="~/Documents/Bioinformatics/Applied high-throughput analysis/project_AHTA/
66                     Transcription profiling of human breast tumors and their paired normal tissues/RMA",force=T)       #RMA produces log-transformed data
67
68 head(exprs(BreastCancerRMA))
69 head(exprs(BreastCancer))
70 #annot <- factor(pData(BreastCancerRMA)[,7]) # normal breast tissue and breast tumor tissue
71 annotation
72 dim(annotation) # 86 2
73
```

Figure 6: Quality control steps and quantile normalization

```
74 # this piece is only necessary if you got the object of ArrayExpress otherwise skip this
75 #annotb <- as.double(annot==annot[4]) # we want the breast tumor tissue to be one and the control to be zero
76 #annotb
77 # the problem is we don't account for the persons but can we? bcs we have 86 samples? YES
78 pData(BreastCancerRMA)$Patients<- pData(BreastCancerRMA)$Hybridization.Name
79 pData(BreastCancerRMA)$Patients <- gsub('Normal', '',pData(BreastCancerRMA)$Patients)
80 pData(BreastCancerRMA)$Patients <- gsub('Cancer', '',pData(BreastCancerRMA)$Patients)
81 pData(BreastCancerRMA)$Patients <- gsub('T', '',pData(BreastCancerRMA)$Patients)
82 pData(BreastCancerRMA)$Patients <- gsub('N', '',pData(BreastCancerRMA)$Patients)
83 pData(BreastCancerRMA)$Patients <- gsub(' ', '',pData(BreastCancerRMA)$Patients)
84 pData(BreastCancerRMA)$Patients # this is column with patients every patient has to occur 2 times in this column
85 sum(grepl('BC0155', pData(BreastCancerRMA)$Patients))
86 sum(grepl('BC0117', pData(BreastCancerRMA)$Patients))
87 ID <- factor(pData(BreastCancerRMA)$Patients)
88 length(levels(ID))
89
90 # begin here again
91 annotation$patients <- factor(annotation$patients)
92 annotation$disease <- factor(annotation$disease)
93 annotation
```

Figure 7: Extra annotation for design matrix

```r
## Differential expression by LIMMA
design <- model.matrix(~0+disease+patients, data= annotation)
colSums(design)
colnames(design)[1:2]<-c("Cancer_tissue","normal_tissue")

fit <- lmFit(BreastCancerRMA,design)
cont.matrix <- makeContrasts(CancervsControl=Cancer_tissue-normal_tissue,levels=design)
fit2 <- contrasts.fit(fit,cont.matrix)
fit2 <- eBayes(fit2)

LIMMAout <- topTable(fit2,adjust="BH",number=nrow(exprs(BreastCancerRMA)))
head(LIMMAout)

# MA plot
threshold.sign <- LIMMAout[LIMMAout$adj.P.Val<0.05,]
dim(threshold.sign)
with(LIMMAout, plot(AveExpr, logFC, pch=20,main="MA plot"))
with(subset(threshold.sign),points(AveExpr, logFC,pch=20,col="red"))
# histogram p-values
hist(fit2$p.value, main= 'distributions of the p-values',xlab='p-values')
# Volcano plot
threshold.sign <- LIMMAout[LIMMAout$adj.P.Val<0.05,]
dim(threshold.sign)
with(LIMMAout, plot(logFC, -log10(P.Value), pch=20,main="Volcano plot"))
with(subset(threshold.sign),points(logFC, -log10(P.Value),pch=20,col="red"))

```

Figure 8: Differential expression analysis using limma

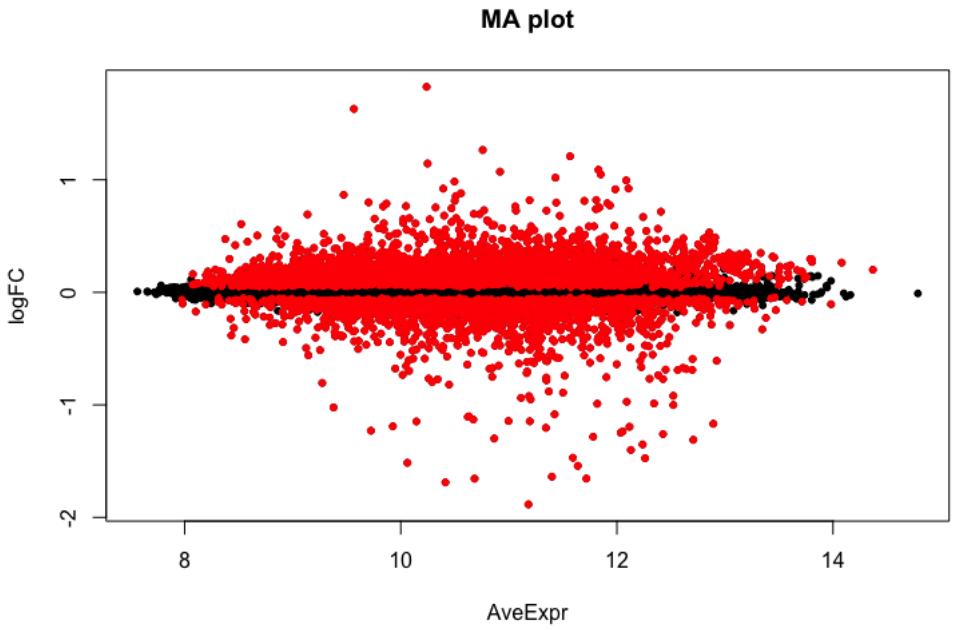

Figure 9: MA plot for the DE using limma for the expression array

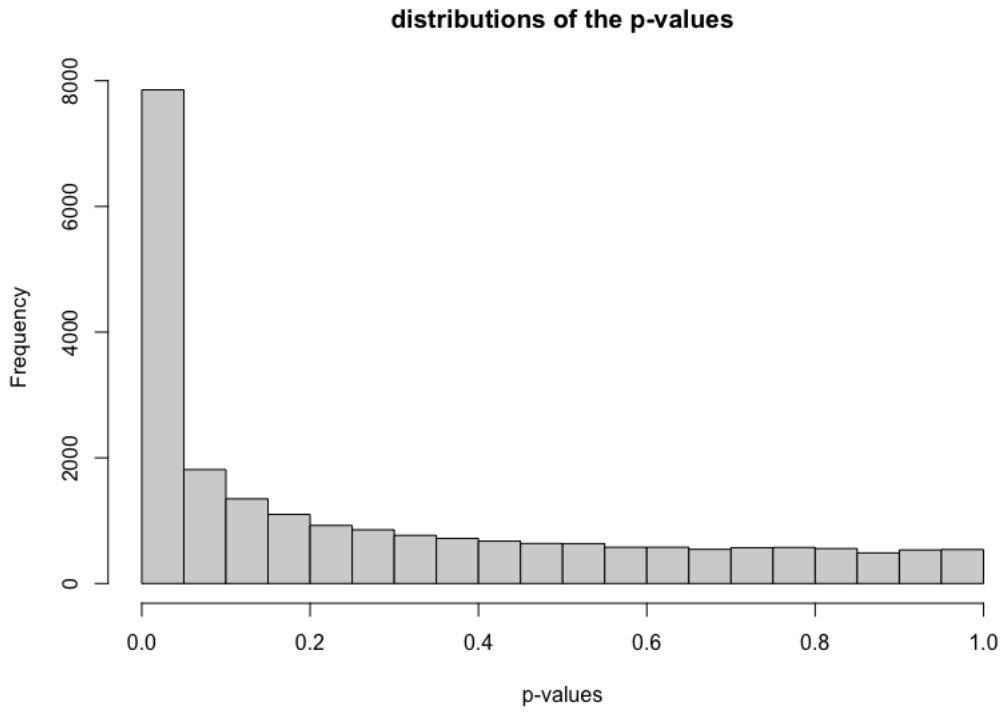

Figure 10: Distribution of the p values from the limma analysis

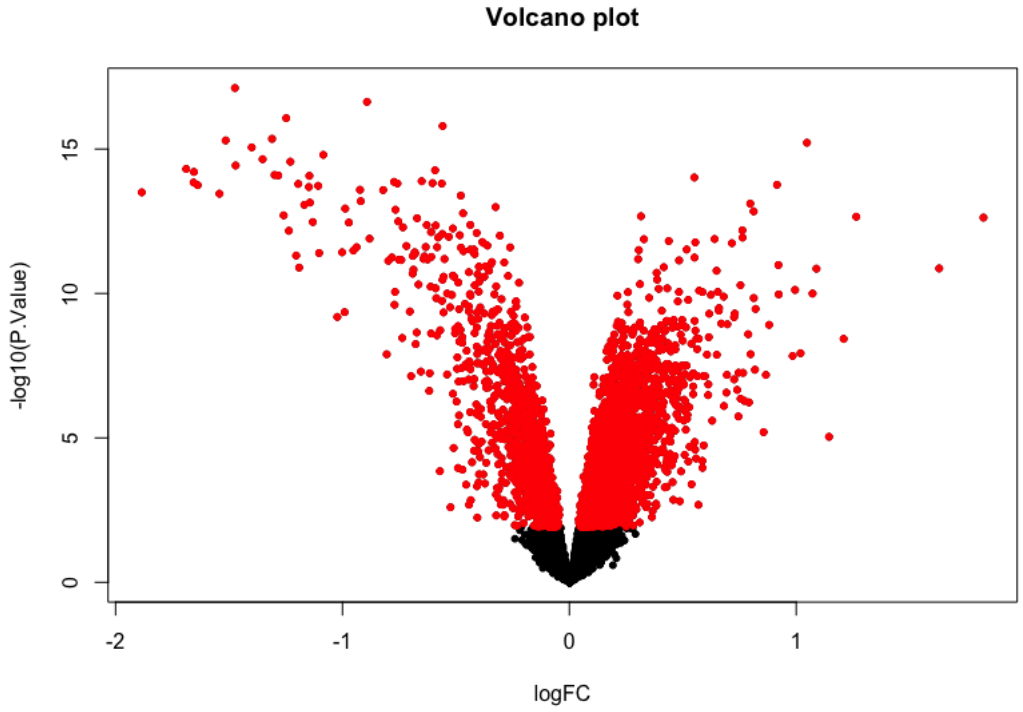

Figure 11: Volcanoplot from the limma analysis

```
# Have a look at the results
significant_pvalues<- LIMMAout[LIMMAout$adj.P.Val<0.05,]
dim(significant_pvalues) # 5333 probes statistical significant
significant_pvalues_1<- LIMMAout[LIMMAout$adj.P.Val<0.05 & abs(LIMMAout$logFC) >1,]
dim(significant_pvalues_1) # 40 probes statistical significant
significant_pvalues_1

significant_pvalues_2<- LIMMAout[LIMMAout$adj.P.Val<0.05 & abs(LIMMAout$logFC) >1.5,]
dim(significant_pvalues_2) # 9 probes statistical significant
head(significant_pvalues_1)

## Load annotation and sort alphabetically on probe name
annotation_BC <- read.table("A-AFFY-33.adf.txt",header=T,sep="\t",skip=17,fill=T)
annotation_BC <- annotation_BC[sort(annotation_BC$Composite.Element.Name,index.return=T)$ix,]

## Check if all probes are present in both sets
dim(annotation_BC)
dim(LIMMAout)

## Double check => "Assumption is the mother of all fuck up's ;)"
sum(annotation_BC$Composite.Element.Name==sort(rownames(LIMMAout)))

## Sort LIMMA output alphabetically on probe name
LIMMAout_sorted <- LIMMAout[sort(rownames(LIMMAout),index.return=T)$ix,]
## Add gene names to LIMMA output
LIMMAout_sorted$gene <- annotation_BC$Composite.Element.Database.Entry.ensembl.
LIMMAout_annot <- LIMMAout_sorted[sort(LIMMAout_sorted$adj.P.Val,index.return=T)$ix,]

```

Figure 12: Gene annotation

```
# Have a look at the results and search for other probesets for your DE genes
head(LIMMAout_annot)
dim(LIMMAout_annot)
significant_pvalues<- LIMMAout_annot[LIMMAout_annot$adj.P.Val<0.05& abs(LIMMAout_annot$logFC) >1,]
adjusted_pvalues<- LIMMAout_annot[LIMMAout_annot$adj.P.Val<0.05,]
dim(significant_pvalues) # 40
dim(adjusted_pvalues) # 5333 7

affyids <- rownames(adjusted_pvalues)

ensembl <- useMart(biomart = "ensembl", dataset = "hsapiens_gene_ensembl")
output_sign <-getBM(attributes = c('affy_hg_u133_plus_2', 'entrezgene_id','hgnc_symbol','chromosome_name', 'start_position', 'end_position'),
                    filters = 'affy_hg_u133_plus_2', #  is a vector of filters that one will use as input to the query.
                    values = affyids, # a vector of values for the filters
                    mart = ensembl)

adjusted_pvalues[,"hgnc_symbol"] <-NA
adjusted_pvalues[,"chromosome_name"] <-NA
adjusted_pvalues[,"start_position"] <-NA
adjusted_pvalues[,"end_position"] <-NA
# filtering of the zero entries

output_sign <- output_sign[!output_sign$hgnc_symbol=="",]
output_sign <- output_sign[!output_sign$affy_hg_u133_plus_2=="",]
count=1
adjusted_pvalues <- adjusted_pvalues[rownames(adjusted_pvalues)%in%output_sign$affy_hg_u133_plus_2,]

for (i in rownames(adjusted_pvalues)){
  adjusted_pvalues$hgnc_symbol[count] <- output_sign[output_sign$affy_hg_u133_plus_2==i,3]
  adjusted_pvalues$chromosome_name[count] <- output_sign[output_sign$affy_hg_u133_plus_2==i,4]
  adjusted_pvalues$start_position[count] <- output_sign[output_sign$affy_hg_u133_plus_2==i,5]
  adjusted_pvalues$end_position[count] <- output_sign[output_sign$affy_hg_u133_plus_2==i,6]
  count=count+1
}

```

Figure 13: Gene annotation for all significant p values

```r
dim(adjusted_pvalues) # 5028 11
head(adjusted_pvalues)

save(adjusted_pvalues,file="adjusted_pvalues_annotation_transcription.Rda") # object = adjusted_pvalues



# for the abs(logFC) > 1
affyids <- rownames(significant_pvalues)

ensembl <- useMart(biomart = "ensembl", dataset = "hsapiens_gene_ensembl")
output_sign <-getBM(attributes = c('affy_hg_u133_plus_2', 'entrezgene_id','hgnc_symbol','chromosome_name', 'start_position', 'end_position'),
                    filters = 'affy_hg_u133_plus_2', #  is a vector of filters that one will use as input to the query.
                    values = affyids, # a vector of values for the filters
                    mart = ensembl)


significant_pvalues[,"hgnc_symbol"] <-NA
significant_pvalues[,"chromosome_name"] <-NA
significant_pvalues[,"start_position"] <-NA
significant_pvalues[,"end_position"] <-NA
# filtering of the zero entries
output_sign <- output_sign[!output_sign$hgnc_symbol=="",]
output_sign <- output_sign[!output_sign$affy_hg_u133_plus_2=="",]
count=1
significant_pvalues <- significant_pvalues[rownames(significant_pvalues)%in%output_sign$affy_hg_u133_plus_2,]

for (i in rownames(significant_pvalues)){
  significant_pvalues$hgnc_symbol[count] <- output_sign[output_sign$affy_hg_u133_plus_2==i,3]
  significant_pvalues$chromosome_name[count] <- output_sign[output_sign$affy_hg_u133_plus_2==i,4]
  significant_pvalues$start_position[count] <- output_sign[output_sign$affy_hg_u133_plus_2==i,5]
  significant_pvalues$end_position[count] <- output_sign[output_sign$affy_hg_u133_plus_2==i,6]
  count=count+1
}
significant_pvalues2 <- significant_pvalues
dim(significant_pvalues2) # 39 11
head(significant_pvalues2)

save(significant_pvalues2,file="Significant_output_annotation_transcription.Rda") # object = significant_pvalues2
sessionInfo()
```

Figure 14: Gene annotation for a selected subset of significant genes (absolute LFC >2)

```
> sessionInfo()
R version 4.0.3 (2020-10-10)
Platform: x86_64-apple-darwin17.0 (64-bit)
Running under: macOS Big Sur 10.16

Matrix products: default
LAPACK: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] parallel  stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] hgu133acdf_2.18.0         biomaRt_2.44.4            limma_3.44.3              ArrayExpress_1.48.0       arrayQualityMetrics_3.44.0 affy_1.66.0               Biobase_2.48.0
[8] BiocGenerics_0.34.0

loaded via a namespace (and not attached):
  [1] colorspace_1.4-1           hwriter_1.3.2              ellipsis_0.3.1             htmlTable_2.1.0            XVector_0.28.0             GenomicRanges_1.40.0
  [7] base64enc_0.1-3            base64_2.0                 rstudioapi_0.11            hexbin_1.28.1              affyio_1.58.0              bit64_4.0.5
 [13] AnnotationDbi_1.50.3       xml2_1.3.2                 codetools_0.2-16           splines_4.0.3              oligoClasses_1.50.4        knitr_1.30
 [19] Formula_1.2-4              jsonlite_1.7.1             annotate_1.66.0            dbplyr_1.4.4               cluster_2.1.0              vsn_3.56.0
 [25] png_0.1-7                  httr_1.4.2                 BiocManager_1.30.10        compiler_4.0.3             backports_1.1.10           assertthat_0.2.1
 [31] Matrix_1.2-18              BeadDataPackR_1.40.0       htmltools_0.5.0            prettyunits_1.1.1          tools_4.0.3                gtable_0.3.0
 [37] glue_1.4.2                 GenomeInfoDbData_1.2.3     reshape2_1.4.4             affxparser_1.60.0          dplyr_1.0.2                rappdirs_0.3.1
 [43] Rcpp_1.0.5                 vctrs_0.3.4                Biostrings_2.56.0          svglite_1.2.3.2            preprocessCore_1.50.0      setRNG_2013.9-1
 [49] iterators_1.0.13           xfun_0.18                  stringr_1.4.0              lifecycle_0.2.0            affyPLM_1.64.0             XML_3.99-0.5
 [55] zlibbioc_1.34.0            scales_1.1.1               hms_0.5.3                  SummarizedExperiment_1.18.2 beadarray_2.38.0          RColorBrewer_1.1-2
 [61] oligo_1.52.1               curl_4.3                   yaml_2.2.1                 memoise_1.1.0              gridExtra_2.3              ggplot2_3.3.2
 [67] gdtools_0.2.2              rpart_4.1-15               latticeExtra_0.6-29        stringi_1.5.3              RSQLite_2.2.1              gcrma_2.60.0
 [73] genefilter_1.70.0          gridSVG_1.7-2              S4Vectors_0.26.1           foreach_1.5.1              checkmate_2.0.0            GenomeInfoDb_1.24.2
 [79] rlang_0.4.8                pkgconfig_2.0.3            systemfonts_0.3.2          bitops_1.0-6               matrixStats_0.57.0         lattice_0.20-41
 [85] purrr_0.3.4                htmlwidgets_1.5.2          bit_4.0.4                  tidyselect_1.1.0           plyr_1.8.6                 magrittr_1.5
 [91] R6_2.5.0                   IRanges_2.22.2             generics_0.0.2             Hmisc_4.4-1                DelayedArray_0.14.1        DBI_1.1.0
 [97] pillar_1.4.6               foreign_0.8-80             survival_3.2-7             RCurl_1.98-1.2             nnet_7.3-14                tibble_3.0.4
[103] crayon_1.3.4               BiocFileCache_1.12.1       jpeg_0.1-8.1               progress_1.2.2             grid_4.0.3                 data.table_1.13.2
[109] blob_1.2.1                 digest_0.6.27              xtable_1.8-4               ff_4.0.4                   illuminaio_0.30.0          openssl_1.4.3
[115] stats4_4.0.3               munsell_0.5.0              askpass_1.1
>
```

Figure 15: Session info for DE analysis script

Methylation profiling by array

```r
#set working directory
setwd("~/Documents/Bioinformatics/Applied high-throughput analysis/project_AHTA/methylation/GSE101443_RAW")
#https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE101443

## Load packages
library('lumi')
library('wateRmelon')
library('ChAMPdata')


methyldata <- readEPIC(getwd())
ID <- c('GSM2703232','GSM2703233', 'GSM2703234', 'GSM2703235', 'GSM2703236','GSM2703237', 'GSM2703238', 'GSM2703239')
condition <- c('Tumour_A','Normal_A','Tumour_B','Normal_B','Tumour_C','Normal_C','Tumour_D','Normal_D')
Patient <- c('A','A','B','B','C','C','D','D')
annotation <- data.frame(ID,condition,Patient)
annotation
## Have a look at the data and annotation
print(methyldata)
print(dim(methyldata))
print(sum(is.na(exprs(methyldata)))) # 1918
print(head(betas(methyldata))) # The "betas" function will retreive the beta values ( = methylation percentages) and the "exprs" function will retreive the M-values.
print(head(exprs(methyldata)))

## Change sampleNames to something more comprehensible
sampleNames(methyldata) <- annotation[,2]
sampleNames(methyldata)
```

Figure 16: Loading in the data for the DM script

```r
## Remove NA values
methyldata <- methyldata[!(rowSums(is.na(exprs(methyldata)))>=1),]
methyldata

## Remove probes for which calling p-value is insufficient
methyldata.pf<-pfilter(methyldata) # removes the probes which we are not sure that are called correctly
#1408 probes were removed

## Comparison of average methylation between control and tumor samples
boxplot(betas(methyldata),las=2)
meth_mean_tumour <- rep(0,8)
meth_mean_control <- rep(0,8)
for (i in 1:ncol(methyldata)){
  if((i %% 2) == 0) { # even
    meth_mean_control[i] <- mean(betas(methyldata)[,i])
  } else {
    meth_mean_tumour[i] <- mean(betas(methyldata)[,i])
  }
}
meth_mean_tumour <- meth_mean_tumour[c(1,3,5,7)]
meth_mean_control <- meth_mean_control[c(2,4,6,8)]

t_test_res <- t.test(meth_mean_control,meth_mean_tumour,var.equal=F)
t_test_res

dat_boxplot <- data.frame(betas = c(meth_mean_tumour,meth_mean_control),
                          group = c('Tumour','Tumour','Tumour','Tumour','Normal','Normal','Normal','Normal'))
boxplot(betas~group,dat_boxplot,las=2)

```

Figure 17: Filtering and comparing the average methylation between controls and tumour samples.

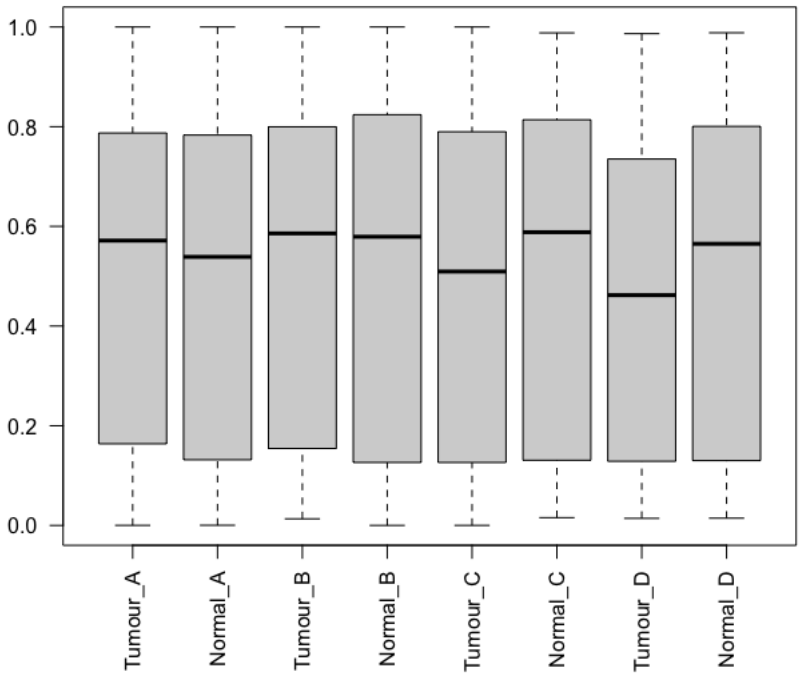

Figure 18: A boxplot of the Betas for each individual sample.

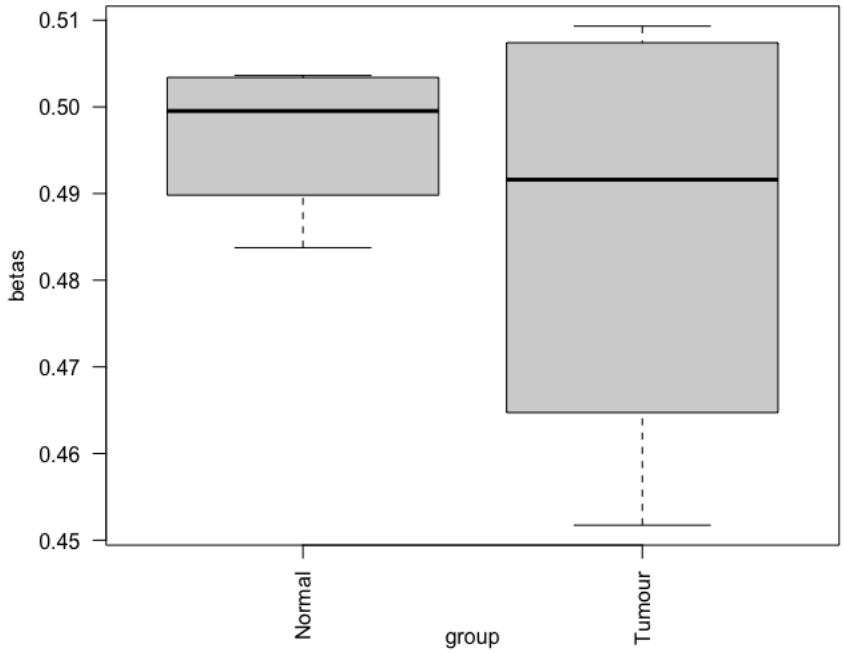

Figure 19: A boxplot of the average Betas for the tumour and normal tissue.

```
> t_test_res <- t.test(meth_mean_control,meth_mean_tumour,var.equal=F)
> t_test_res

        Welch Two Sample t-test

data:  meth_mean_control and meth_mean_tumour
t = 0.74168, df = 3.7048, p-value = 0.5025
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.03019448  0.05127950
sample estimates:
mean of x mean of y
0.4966067 0.4860642
```

Figure 20: Result of the t-test for the average Betas for tumour and normal tissue.

```
## Normalization & QC
########################

## Perform normalization including dye color adjustment
methyldata.dasen.pf <- dasen(methyldata.pf)
#we will use the "dasen" function to adjust for color bias and normalize our data.

## Make methylumi objects to check density and color bias adjustment
#transform both the non-normalized counts as the normalized counts to a MethyLumiM object
methyldataM <- as(methyldata.pf, 'MethyLumiM')
methyldataN <- as(methyldata.dasen.pf, 'MethyLumiM')

## Make QC plot
par(mfrow=c(2,1))
plotColorBias1D(methyldataM,channel="both",main="before",xlab="Beta-value")
plotColorBias1D(methyldataN,channel="both",main="after",xlab="Beta-value")
density(methyldataM,xlab="M-value",main="before")
density(methyldataN,xlab="M-value",main="after")
```

Figure 21: Normalization of the data and QC

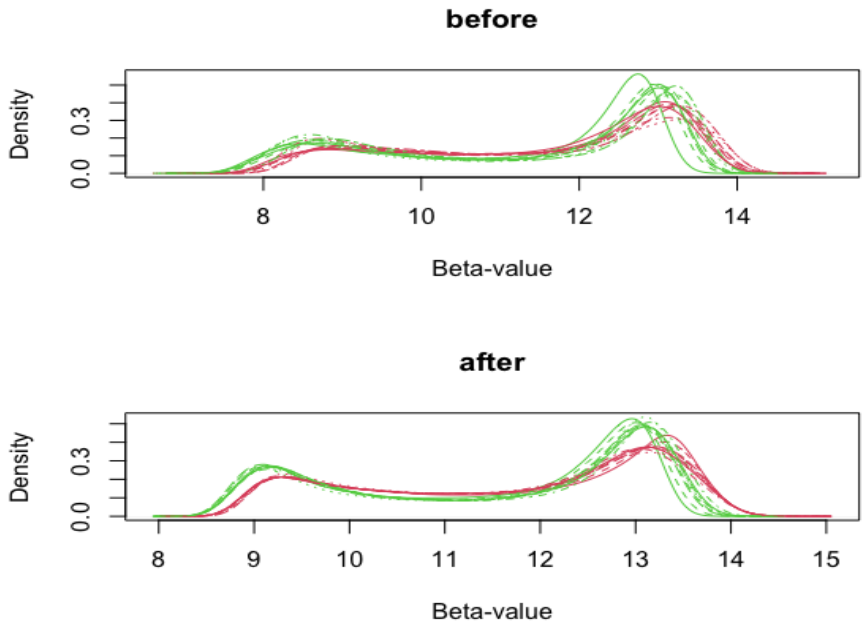

Figure 22: A density plot of the Betas before and after normalization.

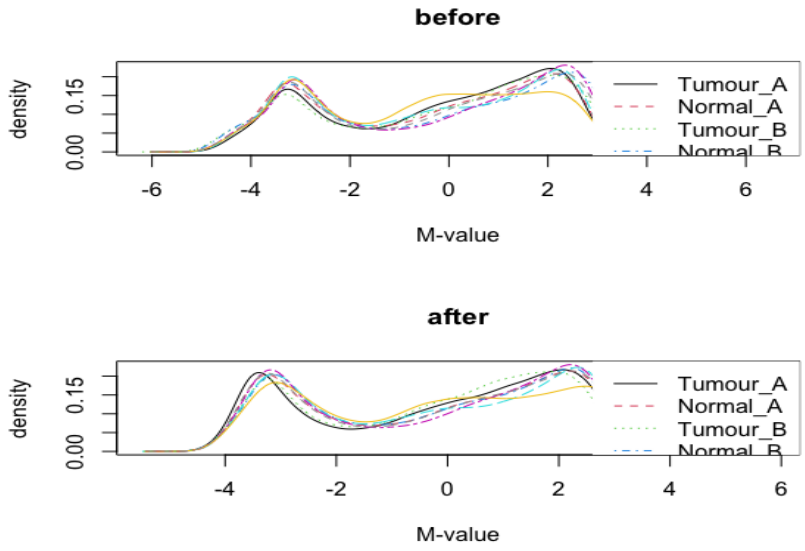

Figure 23: A density plot of the m-values before and after normalization

```r
## Differential methylation analysis: limma
############################
par(mfrow=c(1,1))
## Build design and contrasts
condition <- factor(as.character(c('Tumour','Normal','Tumour','Normal','Tumour','Normal','Tumour','Normal')))
patient <- factor(annotation$Patient)

design2 <- model.matrix(~0+condition+patient)
colnames(design2)[1:2] <- c("Control","Tumor")
design2
cont.matrix2 <- makeContrasts(TumourvsControl=Tumor-Control,levels=design2)

## Limma
fit_2 <- lmFit(methyldataN,design2) # normalized data
fit_2 <- contrasts.fit(fit_2,cont.matrix2)
fit_2 <- eBayes(fit_2)

LIMMAout_2 <- topTable(fit_2,adjust="BH",number=nrow(exprs(methyldata)))
LIMMAout_2

hist(fit_2$p.value, main= 'distributions of the p-values',xlab='p-values')
```

Figure 24: Differential methylation analysis using limma

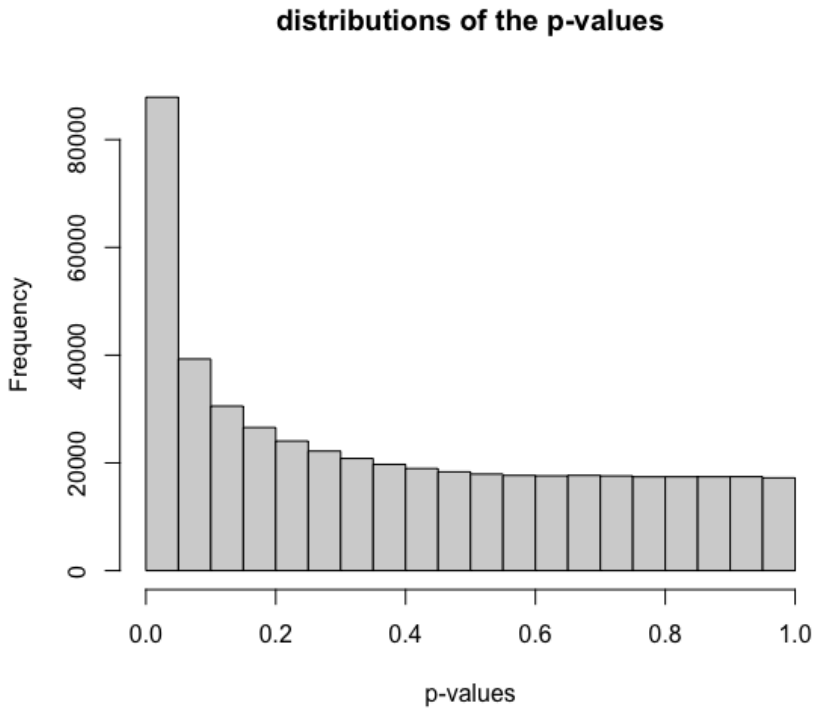

Figure 25: Distributions of the p-values for the limma analysis for DM

```r
# Volcano plot
threshold.sign <- LIMMAout_2[LIMMAout_2$adj.P.Val<0.10,]
dim(threshold.sign)
with(LIMMAout_2, plot(logFC, -log10(P.Value), pch=20,main="Volcano plot"))
with(subset(threshold.sign),points(logFC, -log10(P.Value),pch=20,col="red"))

# MA plot
threshold.sign <- LIMMAout_2[LIMMAout_2$adj.P.Val<0.10,]
threshold.signdown <- LIMMAout_2[LIMMAout_2$adj.P.Val0.10,]
dim(threshold.sign)
with(LIMMAout_2, plot(AveExpr, logFC, pch=20,main="MA plot"))
with(subset(threshold.sign),points(AveExpr, logFC,pch=20,col="red"))


dim(LIMMAout_2[LIMMAout_2$adj.P.Val <= 0.0615913,]) # 1786
dim(LIMMAout_2[LIMMAout_2$adj.P.Val <= 0.10,]) # 18540
dim(LIMMAout_2[abs(LIMMAout_2$logFC) > 2 & LIMMAout_2$adj.P.Val <= 0.10,]) # 933 significant probes


dim(LIMMAout_2[LIMMAout_2$logFC > 2 & LIMMAout_2$adj.P.Val <= 0.10,]) # 701 significant genes
dim(LIMMAout_2[LIMMAout_2$logFC < (-2) & LIMMAout_2$adj.P.Val <= 0.10,]) # 232 significant genes
## Check M-values for top results
exprs(methyldataN)[rownames(methyldataN)%in%rownames(head(LIMMAout_2)),]
betas(methyldataN)[rownames(methyldataN)%in%rownames(head(LIMMAout_2)),]

```

Figure 26: Plots for the limma analysis for DM

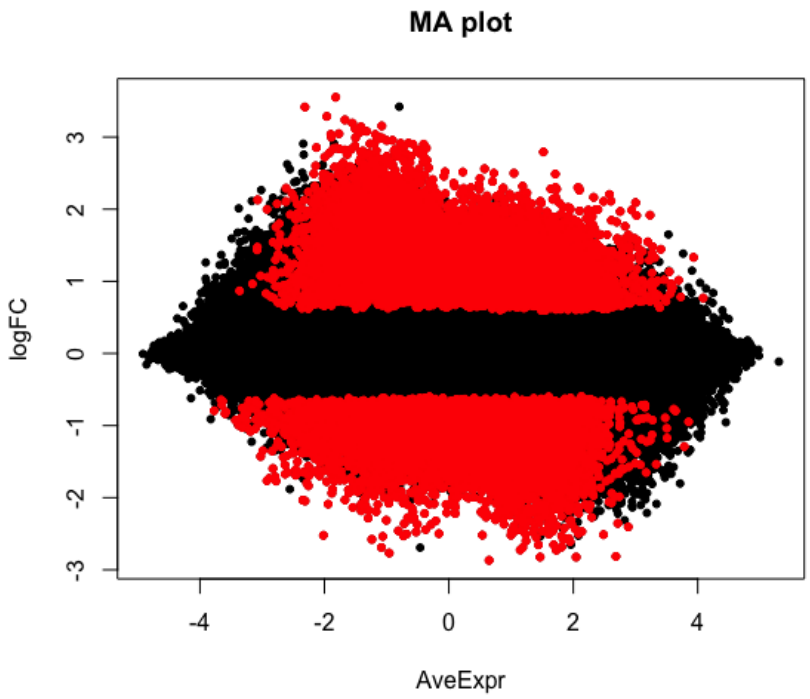

Figure 27: A MA plot for the DM analysis using limma

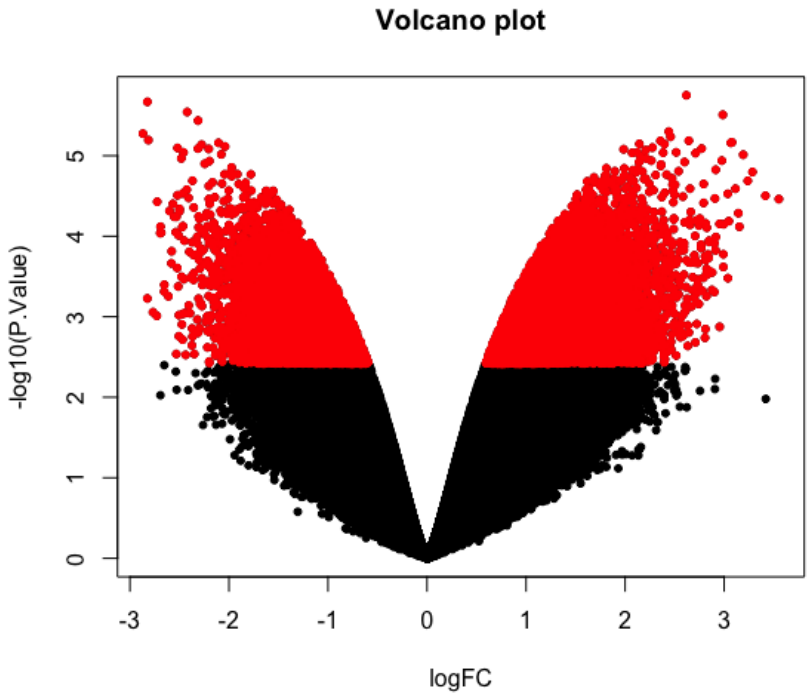

Figure 28: A volcano plot of the DM analysis with limma.

```
## Functional annotation of limma results
############################

## Load annotation and sort alphabetically on probe name
data("probe.features")

annotation_MA <- probe.features
print(head(annotation_MA))
annotation_MA <- annotation_MA[sort(rownames(annotation_MA),index.return=T)$ix,]

## Check if all probes are present in both sets
dim(LIMMAout_2)
dim(annotation_MA) # annotation has more rows than Limma output
sum(LIMMAout_2$Probe_ID%in%rownames(annotation_MA))
sum(rownames(annotation_MA)%in%LIMMAout_2$Probe_ID)
# Also check the reverse so no duplicate rows are present in annotation

## Since more probes are present in the annotation file, remove unnecessary probes
annotation_MA <- annotation_MA[rownames(annotation_MA)%in%LIMMAout_2$Probe_ID,]
dim(annotation_MA)
## Sort LIMMA output alphabetically on probe name
LIMMAout_sorted_2 <- LIMMAout_2[sort(LIMMAout_2$Probe_ID,index.return=T)$ix,]

## Add gene names to LIMMA output
LIMMAout_sorted_2$Gene <- annotation_MA$gene
LIMMAout_sorted_2$Feature <- annotation_MA$feature
LIMMAout_sorted_2$Chrom <- annotation_MA$CHR
LIMMAout_sorted_2$Pos <- annotation_MA$MAPINFO
LIMMAout_sorted_2$Chrom <- as.character(LIMMAout_sorted_2$Chrom)
LIMMAout_sorted_2$Gene <- as.character(LIMMAout_sorted_2$Gene)
LIMMAout_sorted_2$Feature <- as.character(LIMMAout_sorted_2$Feature)

# The data type for these columns is altered to prevent issues further downstream
LIMMAout_annot_2 <- LIMMAout_sorted_2[sort(LIMMAout_sorted_2$P.Value,index.return=T)$ix,c(1,12,13,10,11,4,7,8)]
LIMMAout_annot_2<- LIMMAout_annot_2[!LIMMAout_annot_2$Gene=="",]# filtering
LIMMAout_annot_2
```

Figure 29: Gene annotation for the DM analysis

```r
dim(LIMMAout_annot_2[abs(LIMMAout_annot_2$logFC) > 2 & LIMMAout_annot_2$adj.P.Val <= 0.10,]) # 622
dim(LIMMAout_annot_2[LIMMAout_annot_2$logFC > 2 & LIMMAout_annot_2$adj.P.Val <= 0.10,]) # 481
dim(LIMMAout_annot_2[LIMMAout_annot_2$logFC < (-2) & LIMMAout_annot_2$adj.P.Val <= 0.10,]) # 141
# a positive logFC points to higher methylation in tumor than in control
# a negative logFC points to higher methylation in control than in sample

significant_p_values <- LIMMAout_annot_2[abs(LIMMAout_annot_2$logFC) > 2 & LIMMAout_annot_2$adj.P.Val <= 0.10,]
dim(significant_p_values) # 622 8
# saving results
save(significant_p_values, file= "Methylation_significant.Rda")
foldchanges<- sort(significant_p_values$logFC,decreasing=TRUE)[0:10]
topten <- significant_p_values[significant_p_values$logFC%in%foldchanges,]
topten

adj_pvalues <- LIMMAout_annot_2[LIMMAout_annot_2$adj.P.Val <= 0.10,]
dim(adj_pvalues) # 13179
head(adj_pvalues)
save(adj_pvalues,file="methylation_all_pvalues.Rda")

load('methylation_all_pvalues.Rda')
load('Methylation_significant.Rda')
```

Figure 30: Gene annotation for DM analysis

```r

## Interpretation results
############################

## Select probes in promoter regions
LIMMAout_annot_prom <- significant_p_values[grepl("TSS",significant_p_values$Feature) | (significant_p_values$Feature=="1stExon"),] #TSS = transcription start site
dim(LIMMAout_annot_prom) # 231
head(LIMMAout_annot_prom)

write.table(LIMMAout_annot_prom$Gene,file='Promoter_genes.txt',row.names=FALSE,quote=FALSE,col.names=FALSE)
sessionInfo()

```

Figure 31: Analysis for the promoter regions

```
> sessionInfo()
R version 4.0.3 (2020-10-10)
Platform: x86_64-apple-darwin17.0 (64-bit)
Running under: macOS Big Sur 10.16

Matrix products: default
LAPACK: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats4    parallel  stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] ChAMPdata_2.20.0                            wateRmelon_1.32.0                           illuminaio_0.30.0                           IlluminaHumanMethylation450kanno.ilmn12.hg19_0.6.0
 [5] ROC_1.64.0                                  methylumi_2.34.0                            minfi_1.34.0                                bumphunter_1.30.0
 [9] locfit_1.5-9.4                              iterators_1.0.13                            foreach_1.5.1                               Biostrings_2.56.0
[13] XVector_0.28.0                              SummarizedExperiment_1.18.2                 DelayedArray_0.14.1                         FDb.InfiniumMethylation.hg19_2.2.0
[17] org.Hs.eg.db_3.11.4                         TxDb.Hsapiens.UCSC.hg19.knownGene_3.2.2     GenomicFeatures_1.40.1                      AnnotationDbi_1.50.3
[21] GenomicRanges_1.40.0                        GenomeInfoDb_1.24.2                         IRanges_2.22.2                              S4Vectors_0.26.1
[25] ggplot2_3.3.2                               reshape2_1.4.4                              scales_1.1.1                                matrixStats_0.57.0
[29] lumi_2.40.0                                 hgu133acdf_2.18.0                           biomaRt_2.44.4                              limma_3.44.3
[33] ArrayExpress_1.48.0                         arrayQualityMetrics_3.44.0                  affy_1.66.0                                 Biobase_2.48.0
[37] BiocGenerics_0.34.0

loaded via a namespace (and not attached):
  [1] backports_1.1.10          Hmisc_4.4-1               BiocFileCache_1.12.1      systemfonts_0.3.2         plyr_1.8.6                oligo_1.52.1              splines_4.0.3             BiocParallel_1.22.0
  [9] digest_0.6.27             htmltools_0.5.0           magrittr_1.5              checkmate_2.0.0           memoise_1.1.0             affyPLM_1.64.0            cluster_2.1.0             gcrma_2.60.0
 [17] readr_1.4.0               annotate_1.66.0           beadarray_2.38.0          svglite_1.2.3.2           askpass_1.1               siggenes_1.62.0           prettyunits_1.1.1         jpeg_0.1-8.1
 [25] colorspace_1.4-1          blob_1.2.1                rappdirs_0.3.1            xfun_0.18                 dplyr_1.0.2               crayon_1.3.4              RCurl_1.98-1.2            jsonlite_1.7.1
 [33] hexbin_1.28.1             genefilter_1.70.0         GEOquery_2.56.0           survival_3.2-7            glue_1.4.2                gtable_0.3.0              zlibbioc_1.34.0           BeadDataPackR_1.40.0
 [41] Rhdf5lib_1.10.1           HDF5Array_1.16.1          setRNG_2013.9-1           vsn_3.56.0                bit_4.0.4                 rngtools_1.5              DBI_1.1.0                 Rcpp_1.0.5                xtable_1.8-4
 [49] progress_1.2.2            htmlTable_2.1.0           foreign_0.8-80            bit_4.0.4                 mclust_5.4.6              preprocessCore_1.50.0     Formula_1.2-4             htmlwidgets_1.5.2
 [57] httr_1.4.2                RColorBrewer_1.1-2        ellipsis_0.3.1            ff_4.0.4                  pkgconfig_2.0.3           reshape_0.8.8             XML_3.99-0.5              nnet_7.3-14
 [65] dbplyr_1.4.4              tidyselect_1.1.0          rlang_0.4.8               munsell_0.5.0             tools_4.0.3               generics_0.0.2            RSQLite_2.2.1             stringr_1.4.0
 [73] yaml_2.2.1                knitr_1.30                bit64_4.0.5               beanplot_1.2              oligoClasses_1.50.4       scrime_1.3.5              purrr_0.3.4               nlme_3.1-150
 [81] doRNG_1.8.2               nor1mix_1.3-0             xml2_1.3.2                compiler_4.0.3            rstudioapi_0.11           curl_4.3                  png_0.1-7                 affyio_1.58.0
 [89] tibble_3.0.4              stringi_1.5.3             gdtools_0.2.2             lattice_0.20-41           Matrix_1.2-18             multtest_2.44.0           vctrs_0.3.4               pillar_1.4.6
 [97] lifecycle_0.2.0           BiocManager_1.30.10       data.table_1.13.2         bitops_1.0-6              rtracklayer_1.48.0        R6_2.5.0                  latticeExtra_0.6-29       hwriter_1.3.2
[105] KernSmooth_2.23-17        gridSVG_1.7-2             gridExtra_2.3             affxparser_1.60.0         nleqslv_3.3.2             codetools_0.2-16          MASS_7.3-53               assertthat_0.2.1
[113] rhdf5_2.32.4              openssl_1.4.3             withr_2.3.0               GenomicAlignments_1.24.0  Rsamtools_2.4.0           GenomeInfoDbData_1.2.3    mgcv_1.8-33               hms_0.5.3
[121] quadprog_1.5-8            grid_4.0.3                rpart_4.1-15              tidyr_1.1.2               base64_2.0                DelayedMatrixStats_1.10.1 base64enc_0.1-3
>
```

Figure 222: Session info for the DM analysis script

Enrichr analysis of the overlapping genes between the DE analysis and the DM analysis

Upregulated genes:
https://maayanlab.cloud/Enrichr/enrich?dataset=8c7db6e42219c7c5d96cc9dbe530ea17

Downregulated genes:
https://maayanlab.cloud/Enrichr/enrich?dataset=c6ee7cca7cb0a62758f68f292d70e048

Upregulated promoter associated genes:
https://maayanlab.cloud/Enrichr/enrich?dataset=6cfb4e553111341520a7754bc14bf72d

Downregulated promoter associated genes:
https://maayanlab.cloud/Enrichr/enrich?dataset=1c607d0a18c55508a409a6856ed8c031

RNA sequencing

```
#/bin/bash

#Download data
mkdir raw
mkdir QC
cd raw

wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR201/SRR201983/SRR201983.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR201/SRR201984/SRR201984.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR201/SRR201985/SRR201985.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR201/SRR201986/SRR201986.fastq.gz

fastqc -o ../QC -threads 4 *.gz
#Check html files, overrepresented sequences are present
#Create an adapters.fa file

#Trimming
java -jar /data/student_homes/public/PracticalSession3/trimmomatic-0.36.jar PE -threads 3 breastCancer_1.fastq.gz \
breastCancer_2.fastq.gz trimmed_breastCancer_1.fastq.gz /dev/null trimmed_breastCancer_2.fastq.gz \
/dev/null ILLUMINACLIP:adapters.fa:2:30:10

java -jar /data/student_homes/public/PracticalSession3/trimmomatic-0.36.jar PE -threads 3 control_1.fastq.gz \
control_2.fastq.gz trimmed_control_1.fastq.gz /dev/null trimmed_control_2.fastq.gz /dev/null ILLUMINACLIP:adapters.fa:2:30:10

#QC to check if trimming was succesfull
fastqc -o ../QC -threads 2 trimmed*.gz
```

Figure 33: shell script to download data, perform quality control and adapter trimming

```
#!/bin/bash

mkdir genome
cd genome

#download hg18 genome and annotation
wget ftp://ftp.ensembl.org/pub/release-101/gtf/homo_sapiens/Homo_sapiens.GRCh38.101.gtf.gz
wget ftp://ftp.ensembl.org/pub/release-101/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz

gunzip *.gz

gtf_to_fasta Homo_sapiens.GRCh38.101.gtf Homo_sapiens.GRCh38.dna.primary_assembly.fa hg18_transcriptome.fa
#Make sure faste identifiers start with an Ensembl ID
sed 's/>[[:digit:]]\+ />/g' hg18_transcriptome.fa > hg18_transcriptome_ENS.fa
kallisto index -i hg18_index hg18_transcriptome.fa
cd ..
mkdir counts
cd raw

for file in $(ls trimmed*gz | sed -e "s/trimmed_//" | sed "s/_.*//" | sort -u ); { kallisto quant -i ../genome/hg18_index \
-o ../counts/$file --threads 3 "trimmed_"$file"_1.fastq.gz" "trimmed_"$file"_2.fastq.gz"; }

# Paste quantification files together to create a count file
cd ../counts
paste breastCancer/abundance.tsv control/abundance.tsv | cut -f 1,2,4,5,9,10 > counts.txt
# Now download this file to your local computer for further analysis in R
```

Figure 34: shell script to quantify reads using Kallisto

```
#!/bin/Rscript

setwd("C:/Users/Boris/Documents/School/2de master/Applied high throughput analysis/Project/RNAseq")
#Create transcripts to gene table
library("org.Hs.eg.db")
transcripts <- as.list(org.Hs.egENSEMBLTRANS2EG)
transcripts <- data.frame(ENTREZID = unlist(transcripts))
transcripts$transcripts <- rownames(transcripts)
transcripts <- transcripts[,c(2, 1)] #First column must contain transcript IDs
#Read count matrix, group transcript from the same gene based on the transcripts table (tx2gene)
library(tximport)
files <- c("Control_abundance.tsv", "Cancer_abundance.tsv")
names(files) <- c("Control", "Cancer")
#Why scaledTPM? -> we're grouping counts of transcripts of different lenght per gene
data <- tximport(files=files, type="kallisto", tx2gene = transcripts, ignoreAfterBar = TRUE, countsFromAbundance = "scaledTPM")
head(data$counts)


## DE analysis
library(edgeR)
counts <- data$counts
bcv <- 0.4 # Value suggested by the EdgeR manual in case there are no replicates for human data
DGE_list <- DGEList(counts=counts, group=1:2)
# Filter out genes with low counts
keep <- filterByExpr(DGE_list)
DGE_list <- DGE_list[keep, , keep.lib.sizes=FALSE]
#TMM normalization and test for DE
DGE_list <- calcNormFactors(DGE_list)
DGE_list$samples
et <- exactTest(DGE_list, dispersion=bcv^2)
res <- topTags(et,n=nrow(data))
write.table(data$counts, "Counts.txt", quote = FALSE, col.names = NA, row.names = TRUE, sep = "\t")
```

Figure35: part of the Rscript that was to load in data and perform DE analysis using EdgeR

```
#Add some more annotation
library("TxDb.Hsapiens.UCSC.hg38.knownGene")
library("org.Hs.eg.db")
annots <- select(org.Hs.eg.db, keys=rownames(res),
                 columns=c("SYMBOL","GENENAME"), keytype="ENTREZID")
resultTable <- merge(res, annots, by.x=0, by.y="ENTREZID")
head(resultTable)
write.table(resultTable, file = "EdgeR_out.txt", sep = "\t", row.names = FALSE, quote = F)

## Volcano plot and histogram of p-values
library(ggplot2)
data <- read.table("EdgeR_out.txt", header=TRUE, sep = "\t", quote="")
##Identify the genes that have a FDR < 0.001
data$threshold = as.factor(data$FDR < 0.05)
##Construct the plot object
g <- ggplot(data=data,
            aes(x=logFC, y =-log10(FDR),
                colour=threshold)) +

  geom_point(alpha=0.4, size=3) +
  xlim(c(-14, 14)) +
  xlab("log2 fold change") + ylab("-log10 FDR") +
  theme_bw() +
  scale_color_manual(values = c("blue", "red")) +
  theme(legend.position="none")

png("Pvalue_histogram.png")
hist(resultTable$PValue, 100, xlab = "P values")
dev.off()
png("Volcano.png")
g
dev.off()
```

Figure 36: part of the Rscript that was used to add gene symbol names to DE genes and create a volcano plot and histogram of p-values.

```r
# Filter out transcripts without annotation
resultTable <- resultTable[!is.na(resultTable$SYMBOL),]
dim(resultTable[(resultTable$logFC > 2),])
dim(resultTable[(resultTable$logFC > 0),])
dim(resultTable[(resultTable$FDR < 0.05),])

#Create gene lists for Enrichr pathway anaylsis
up <- resultTable[(resultTable$FDR < 0.05 & resultTable$logFC > 2),]
dim(up)
down <- resultTable[(resultTable$FDR < 0.05 & resultTable$logFC < -2),]
dim(down)

write.table(up$SYMBOL,"up.txt", row.names = FALSE, sep='\t', quote=F, col.names=F)
write.table(down$SYMBOL,"down.txt", row.names = FALSE, sep='\t', quote=F, col.names=F)

all_up <- resultTable[(resultTable$FDR < 0.05 & resultTable$logFC > 0),]
dim(all_up)
all_down <- resultTable[(resultTable$FDR < 0.05 & resultTable$logFC < 0),]
dim(down)

write.table(all_up$SYMBOL,"all_up.txt", row.names = FALSE, sep='\t', quote=F, col.names=F)
write.table(all_down$SYMBOL,"all_down.txt", row.names = FALSE, sep='\t', quote=F, col.names=F)

```

Figure 37: part of the Rscript that was used to filter out DE genes and create a list of gene names

```
> sessionInfo()
R version 4.0.3 (2020-10-10)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 18363)

Matrix products: default

locale:
[1] LC_COLLATE=Dutch_Belgium.1252  LC_CTYPE=Dutch_Belgium.1252    LC_MONETARY=Dutch_Belgium.1252 LC_NUMERIC=C
[5] LC_TIME=Dutch_Belgium.1252

attached base packages:
[1] parallel  stats4    stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] ggplot2_3.3.2                   TxDb.Hsapiens.UCSC.hg38.knownGene_3.10.0 GenomicFeatures_1.42.0
 [4] GenomicRanges_1.42.0            edgeR_3.32.0                             limma_3.46.0
 [7] txiimport_1.18.0                org.Hs.eg.db_3.12.0                      AnnotationDbi_1.52.0
[10] IRanges_2.24.0                  S4Vectors_0.28.0                         Biobase_2.50.0
[13] BiocGenerics_0.36.0

loaded via a namespace (and not attached):
  [1] MatrixGenerics_1.2.0        httr_1.4.2                  bit64_4.0.5                 assertthat_0.2.1
  [5] askpass_1.1                 BiocFileCache_1.14.0        blob_1.2.1                  GenomeInfoDbData_1.2.4
  [9] progress_1.2.2              pillar_1.4.6                RSQLite_2.2.1               Rsamtools_2.6.0
 [13] lattice_0.20-41             glue_1.4.2                  digest_0.6.27               XVector_0.30.0
 [17] colorspace_1.4-1            Matrix_1.2-18               XML_3.99-0.5                pkgconfig_2.0.3
 [21] biomaRt_2.46.0              zlibbioc_1.36.0             purrr_0.3.4                 scales_1.1.1
 [25] BiocParallel_1.24.0         tibble_3.0.4                openssl_1.4.3               farver_2.0.3
 [29] generics_0.1.0              ellipsis_0.3.1              withr_2.3.0                 SummarizedExperiment_1.20.0
 [33] magrittr_2.0.1              crayon_1.3.4                memoise_1.1.0               xml2_1.3.2
 [37] tools_4.0.3                 prettyunits_1.1.1           hms_0.5.3                   lifecycle_0.2.0
 [41] matrixStats_0.57.0          stringr_1.4.0               munsell_0.5.0               locfit_1.5-9.4
 [45] DelayedArray_0.16.0         Biostrings_2.58.0           compiler_4.0.3              rlang_0.4.9
 [49] grid_4.0.3                  RCurl_1.98-1.2              rappdirs_0.3.1              rstudioapi_0.11
 [53] bitops_1.0-6                labeling_0.4.2              gtable_0.3.0                DBI_1.1.0
 [57] curl_4.3                    R6_2.5.0                    GenomicAlignments_1.26.0    dplyr_1.0.2
 [61] rtracklayer_1.49.5          bit_4.0.4                   readr_1.4.0                 stringi_1.5.3
 [65] Rcpp_1.0.5                  vctrs_0.3.4                 dbplyr_2.0.0                tidyselect_1.1.0
> |
```

Figure 38: session info of the Rscript that was used for RNAseq analysis

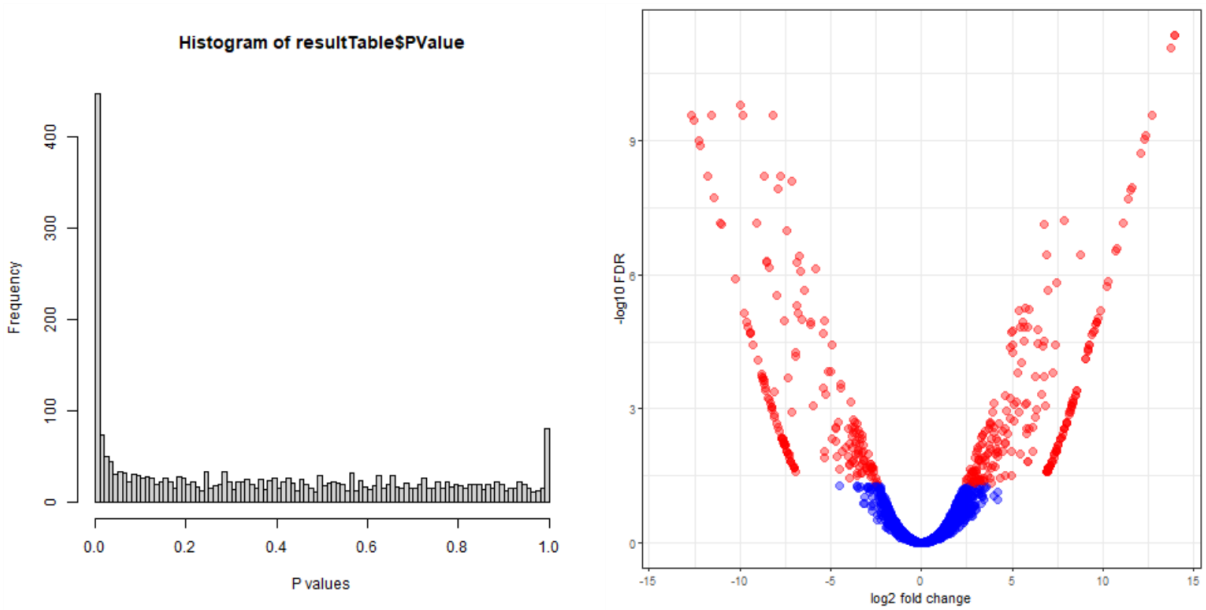

Figure 39: histogram of p-values and volcano plot showing significantly DE genes in red

RNA sequencing gene ontology analysis in Enrichr:

Upregulated genes:
https://maayanlab.cloud/Enrichr/enrich?dataset=07b6432de934c1db58623be24d1b62b5

Downregulated genes:
https://maayanlab.cloud/Enrichr/enrich?dataset=11be9f939ee477c5c2eedf19064f2f52

Multi omics-based analysis

```r
#set working directory
setwd("~/Documents/Bioinformatics/Applied high-throughput analysis/project_AHTA/Transcription profiling of human breast tumors and their paired normal tissues")

library(goseq)
library(tidyverse)
library(ggplot2)

load("Significant_output_annotation_transcription.Rda")# significant_pvalues2
load('methylation_all_pvalues.Rda') # adj_pvalues
load("adjusted_pvalues_annotation_transcription.Rda") # adjusted pvalues
load("Methylation_significant.Rda") # significant_p_values


down_RNA<- read.table('down-2.txt', header=FALSE,quote= "")
up_RNA<- read.table('up-2.txt', header=FALSE)

dim(down_RNA) # 184
dim(up_RNA)  #239

total_RNA <- rbind(down_RNA,up_RNA)
head(total_RNA)
write.table(total_RNA,file='RNA_seq_all.txt',row.names=FALSE,quote=FALSE,col.names=FALSE)
dim(total_RNA)

dim(significant_pvalues2) # 39 11
# significant adjusted p values(0.05) for transcription and log fold changes of > 1
dim(adjusted_pvalues) # 5028
# all significant adjusted p values(0.05) for transcription
dim(significant_p_values) #622 8
# significant adjusted p values(0.010) for methylation and log fold changes of > 2
dim(adj_pvalues) # 13179
# all significant adjusted p values(0.10) for methylation

```

Figure 40: Loading in the multiple results

```r
head(up_RNA)
head(adj_pvalues)
head(adjusted_pvalues)
head(significant_pvalues2) # transcription
head(significant_p_values) # methylation

# writing tables => for transcription 39 significant genes
write.table(significant_pvalues2$hgnc_symbol,file='transcription_genes.txt',row.names=FALSE,quote=FALSE,col.names=FALSE)
pos_logfold<- significant_pvalues2[significant_pvalues2$logFC >= 1,]
dim(pos_logfold) # 9 11
# this means higher expression in the tumor samples
neg_logfold<- significant_pvalues2[significant_pvalues2$logFC <= (-1),]
dim(neg_logfold) #30 11
# this means higher expression in the control samples
write.table(pos_logfold$hgnc_symbol,file='transcription_genes_pos_logfold.txt',row.names=FALSE,quote=FALSE,col.names=FALSE)
write.table(neg_logfold$hgnc_symbol,file='transcription_genes_neg_logfold.txt',row.names=FALSE,quote=FALSE,col.names=FALSE)

#writing tables => for all transcription significant genes
write.table(adjusted_pvalues$hgnc_symbol,file='transcription_genes_all.txt',row.names=FALSE,quote=FALSE,col.names=FALSE)
length(unique(adjusted_pvalues$hgnc_symbol)) # 4069
pos_expression <- adjusted_pvalues[adjusted_pvalues$logFC >= 0,]
length(unique(pos_expression$hgnc_symbol)) #2621
write.table(pos_expression$hgnc_symbol,file='transcription_genes_pos_logfold_all.txt',row.names=FALSE,quote=FALSE,col.names=FALSE)
neg_expression <- adjusted_pvalues[adjusted_pvalues$logFC <= 0,]
length(unique(neg_expression$hgnc_symbol)) # 1495
write.table(neg_expression$hgnc_symbol,file='transcription_genes_neg_logfold_all.txt',row.names=FALSE,quote=FALSE,col.names=FALSE)

```

Figure 41: Separating the files according to LFC and writing to txt files

```r
#writing tables => for methylation 622 significant genes
write.table(significant_p_values$Gene,file='methylation_genes.txt',row.names=FALSE,quote=FALSE,col.names=FALSE)
length(unique(significant_p_values$Gene)) #348
length(significant_p_values$Gene) #622
dim(significant_p_values)
pos_logfold<- significant_p_values[significant_p_values$logFC >= 2,]
length(unique(pos_logfold$Gene)) # 250
# this means higher expression in the tumor samples
neg_logfold<- significant_p_values[significant_p_values$logFC <= (-2),]
length(unique(neg_logfold$Gene)) # 104
unique(neg_logfold$Gene)[unique(neg_logfold$Gene)%in%unique(pos_logfold$Gene)]
significant_p_values[significant_p_values$Gene=='SCT',]

# this means higher expression in the control samples
write.table(pos_logfold$Gene,file='methylation_genes_pos_logfold.txt',row.names=FALSE,quote=FALSE,col.names=FALSE)
write.table(neg_logfold$Gene,file='methylation_genes_neg_logfold.txt',row.names=FALSE,quote=FALSE,col.names=FALSE)

#writing tables => for methylation 13179 significant genes
write.table(adj_pvalues$Gene,file='methylation_genes_all.txt',row.names=FALSE,quote=FALSE,col.names=FALSE)
length(unique(adj_pvalues$Gene)) # 5917
pos_logfold<- adj_pvalues[adj_pvalues$logFC > 0,]
length(unique(pos_logfold$Gene)) # 3862
# this means higher expression in the tumor samples
neg_logfold<- adj_pvalues[adj_pvalues$logFC < 0,]
length(unique(neg_logfold$Gene)) # 2744
# this means higher expression in the control samples
write.table(pos_logfold$Gene,file='methylation_genes_pos_logfold_all.txt',row.names=FALSE,quote=FALSE,col.names=FALSE)
write.table(neg_logfold$Gene,file='methylation_genes_neg_logfold_all.txt',row.names=FALSE,quote=FALSE,col.names=FALSE)
```

Figure 42: Separating the files according to LFC and writing to txt files

```r
# Is methylation in promotor region correlated with down regulation of expression?
## Select probes in promoter regions for whole dataset
LIMMAout_annot_prom_all <- adj_pvalues[grepl("TSS",adj_pvalues$Feature) | (adj_pvalues$Feature=="1stExon"),] #TSS = transcription start site
length(unique(LIMMAout_annot_prom_all$Gene)) # 2100
head(LIMMAout_annot_prom_all)

positive_methylation_promotor <- LIMMAout_annot_prom_all[LIMMAout_annot_prom_all$logFC > 0,]
negative_methylation_promotor <- LIMMAout_annot_prom_all[LIMMAout_annot_prom_all$logFC < 0,]

write.table(positive_methylation_promotor$Gene,file='methylation_promotor_pos.txt',row.names=FALSE,quote=FALSE,col.names=FALSE)
write.table(negative_methylation_promotor$Gene,file='methylation_promotor_neg.txt',row.names=FALSE,quote=FALSE,col.names=FALSE)
length(unique(positive_methylation_promotor$Gene)) # 1124
length(unique(negative_methylation_promotor$Gene)) # 999

# => more methylation in tumor promotor
sum(unique(positive_methylation_promotor$Gene)%in%unique(neg_expression$hgnc_symbol)) # 125
sum(unique(positive_methylation_promotor$Gene)%in%unique(pos_expression$hgnc_symbol)) # 63

test_1 <- unique(positive_methylation_promotor$Gene)[unique(positive_methylation_promotor$Gene)%in%unique(neg_expression$hgnc_symbol)] # 125
write.table(test_1,file='methylation_promotor_test1.txt',row.names=FALSE,quote=FALSE,col.names=FALSE)

# => more methylation in normal promotor
sum(unique(negative_methylation_promotor$Gene)%in%unique(neg_expression$hgnc_symbol)) # 49
sum(unique(negative_methylation_promotor$Gene)%in%unique(pos_expression$hgnc_symbol)) # 119

test_2 <- unique(negative_methylation_promotor$Gene)[unique(negative_methylation_promotor$Gene)%in%unique(pos_expression$hgnc_symbol)] # 119
write.table(test_2,file='methylation_promotor_test2.txt',row.names=FALSE,quote=FALSE,col.names=FALSE)
```

Figure 43: Selecting the genes associated with promoter region and writing to txt files

```r
# Gene Set Enrichment Analysis with the overlapping genes between microarray and the methylation
length(unique(adj_pvalues$Gene)) # 5917 methylation

length(unique(pos_expression$hgnc_symbol)) #2621
length(unique(neg_expression$hgnc_symbol)) # 1495

sum(unique(pos_expression$hgnc_symbol)%in%unique(adj_pvalues$Gene)) #690
sum(unique(neg_expression$hgnc_symbol)%in%unique(adj_pvalues$Gene)) # 533

list_pos_genes <- unique(pos_expression$hgnc_symbol)[unique(pos_expression$hgnc_symbol)%in%unique(adj_pvalues$Gene)]
list_neg_genes <- unique(neg_expression$hgnc_symbol)[unique(neg_expression$hgnc_symbol)%in%unique(adj_pvalues$Gene)]

write.table(list_pos_genes,file='list_pos_genes.txt',row.names=FALSE,quote=FALSE,col.names=FALSE) # more expression in tumor
write.table(list_neg_genes,file='list_neg_genes.txt',row.names=FALSE,quote=FALSE,col.names=FALSE) # less expression in tumor

sessionInfo()
```

Figure 44: Gene set analysis

```
> sessionInfo()
R version 4.0.3 (2020-10-10)
Platform: x86_64-apple-darwin17.0 (64-bit)
Running under: macOS Big Sur 10.16

Matrix products: default
LAPACK: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

loaded via a namespace (and not attached):
 [1] zlibbioc_1.34.0       BiocManager_1.30.10   compiler_4.0.3        parallel_4.0.3        tools_4.0.3           affy_1.66.0           yaml_2.2.1            affyio_1.58.0
 [9] Biobase_2.48.0        preprocessCore_1.50.0 BiocGenerics_0.34.0
>
```

Figure 45: Session info from the multi omics-based analysis

ChIP sequencing

```
#!/bin/bash

#Download data
cd raw
#Good outcome
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR568/SRR568483/SRR568483.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR568/SRR568486/SRR568486.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR568/SRR568489/SRR568489.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR568/SRR568492/SRR568492.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR568/SRR568495/SRR568495.fastq.gz

#Poor outcome
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR568/SRR568498/SRR568498.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR568/SRR568501/SRR568501.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR568/SRR568503/SRR568503.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR568/SRR568506/SRR568506.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR568/SRR568509/SRR568509.fastq.gz

#input mix
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR568/SRR568518/SRR568518.fastq.gz

mkdir ../QC
fastqc -o ../QC -threads 2 *.gz
# Trimming is needed
for file in *.gz; { java -jar /data/student_homes/public/PracticalSession3/trimmomatic-0.36.jar SE -threads 2 $file \
trimmed_$file ILLUMINACLIP:adapters.fa:2:30:10 &
if [ $(jobs | wc -l) -gt 2 ]; then
        wait -n
fi
 }

# QC of trimmed data
fastqc -o ../QC -threads 3 trimmed*.gz
```

Figure 46: shell script to download data, perform quality control and trimming

```
#!/bin/bash

mkdir genome
cd genome

#Fasta with genome sequence was already downloaded in the RNAseq experiment
bowtie2-build --threads 3 \
/data/student_homes/Boris.Vandemoortele/AHTA/project/RNAseq/genome/Homo_sapiens.GRCh38.dna.primary_assembly.fa \
hg18
# hg18 is the wrong name, should be hg38 (but it's the right genome so everything's fine

#set bowtie index variable
BOWTIE2_INDEXES=/data/student_homes/Boris.Vandemoortele/AHTA/project/ChIPseq/genome
export BOWTIE2_INDEXES

mkdir aligned
cd raw

for i in $( ls trimmed*.gz | sed s"/trimmed_//" | sed s"/.fastq.gz//" ); { gunzip -c trimmed_$i.fastq.gz \
| bowtie2 -q --threads 2 -x hg18 -U - | samtools view -b | \
samtools sort -o /data/student_homes/Boris.Vandemoortele/AHTA/project/ChIPseq/aligned/$i.bam &
if [ $(jobs | wc -l) -gt 2 ]; then
        wait -n
fi
 }

# bam files now have a number in the chr column, change it to 'chr_number' for further analysis with HOMER
cd ../aligned

for file in *.bam
do
    filename=`echo $file | cut -d "." -f 1`
    samtools view -H $file | sed -e 's/SN:\([0-9XY]\)/SN:chr\1/' -e 's/SN:MT/SN:chrM/' | samtools reheader - $file > ${filename}_chr.bam
done

#.bam files can be removed to save some disk space

# Finally, rename files from accession.bam to condition.bam (check online which accession is which condition)
mv SRR568483_chr.bam good_1.bam
mv SRR568486_chr.bam good_2.bam
mv SRR568489_chr.bam good_3.bam
mv SRR568492_chr.bam good_4.bam
mv SRR568495_chr.bam good_5.bam
mv SRR568498_chr.bam poor_1.bam
mv SRR568501_chr.bam poor_2.bam
mv SRR568503_chr.bam poor_3.bam
mv SRR568506_chr.bam poor_4.bam
mv SRR568509_chr.bam poor_5.bam
mv SRR568518_chr.bam input.bam
```

Figure 47: shell script for mapping to the hg38 reference genome, do some output formatting on bam files and rename files

```
#!/bin/bash

# Create tag directories
mkdir tagdirs
cd aligned
for file in $( ls *.bam | sed s"/.bam//" ); \
{ makeTagDirectory /data/student_homes/Boris.Vandemoortele/AHTA/project/ChIPseq/tagdirs/$file -unique \
-genome hg38 $file.bam &
if [ $(jobs | wc -l) -gt 3 ]; then
        wait -n
fi
 }

cd ..
mkdir peaks
cd tagdirs

# Peak calling
for dir in *; { findPeaks $dir -style factor -o ../peaks/$dir -i input -F 8.0 &
if [ $(jobs | wc -l) -gt 1 ]; then
        wait -n
fi
 }

cd ..
cd peaks
mkdir merged
# Merge peaks from the same condition
mergePeaks good* > merged/good
mergePeaks poor* > merged/poor

#Annotate peaks
cd merged
annotatePeaks.pl good hg38 > good_annotated.txt
annotatePeaks.pl poor hg38 > poor_annotated.txt
```

Figure 48: shell script for peak calling using HOMER, merging peaks per group and annotating those peaks

```
#!/bin/bash
mkdir motifs
cd peaks/diffPeaks
for file in *; { findMotifsGenome.pl $file hg38 /data/student_homes/Boris.Vandemoortele/AHTA/project/ChIPseq/motifs/$file \
-p 2 -S 15 &
if [ $( jobs | wc -l ) -gt 1 ]; then
        wait -n
fi
 }

cd ../merged
for file in *; { findMotifsGenome.pl $file hg38 /data/student_homes/Boris.Vandemoortele/AHTA/project/ChIPseq/motifs/$file \
-p 2 -S 15 &
if [ $( jobs | wc -l ) -gt 1 ]; then
        wait -n
fi
 }
```

Figure 49: shell script for HOMER motif finding

Gene ontology analysis of genes associated to peaks:

Genes associated to peaks from the aromatase inhibitor unresponsive group: https://maayanlab.cloud/Enrichr/enrich?dataset=381b05eb88816ca70959a5f8e8c8aa2b

Genes associated to peaks in both groups: https://maayanlab.cloud/Enrichr/enrich?dataset=67c4fb04642336c4ad2e68b80ea4805f