**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

Logo, company name

Description automatically generated

Logo, company name

Description automatically generated

# 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 countries1. 1.15 million breast cancer cases are diagnosed each year worldwide, making breast cancer one of the most prevalent cancers in the world today2. 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 future3. 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 risk4, 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 stage5. An exact mechanism that drives metastasis is yet to be unravelled, but some important new insights have been gained recently6. 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 negative7. 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 ER8. 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 women9. 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 chemotherapy10.

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 ArrayExpress11 under the ArrayExpress experiment identifier E-GEOD-1585212 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 GSE10144315. 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 enrichment16.

## RNAseq

Datasets were obtained from the *European Nucleotide Archive (ENA)*17 under project ID PRJNA14288718. This dataset RNA comprises sequencing results from both normal breast cells (HMEC) and breast cancer cells (HCC1954). Quality control was performed using fastqc19 and overrepresented sequences were removed using Trimmomatic’s20 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 Kallisto21, 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 PRJNA17514423. 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 Bowtie224 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 HOMER25. 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 (α < 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.

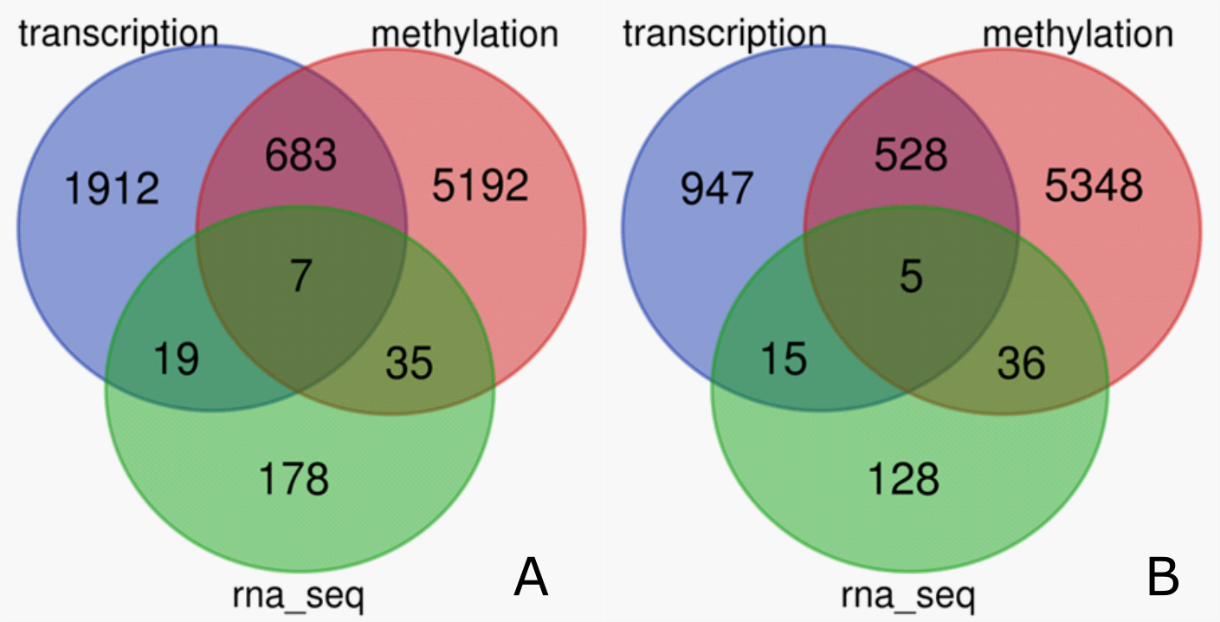
**Table 1: Gene Ontology (GO) analysis of differentially expressed genes in HCC1954 cells vs HMEC cells.** 239 and 184 genes were respectively up- and downregulated in HCC1954 cells compared to HMEC cells (n=2, FDR < 0.05 and absolute LFC > 2). Shown are the interesting hits of GO analysis in Enrichr 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 growth27. 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.

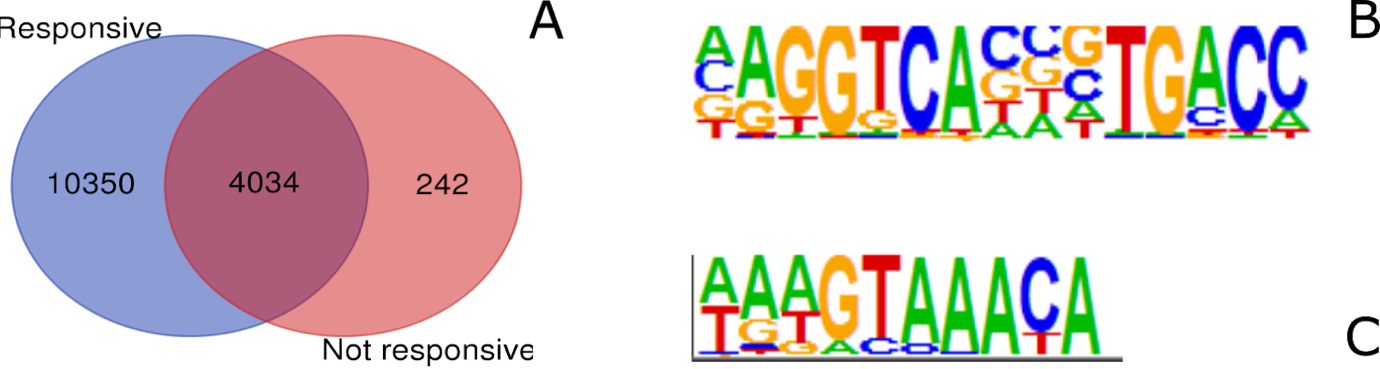
***Table 3: Overlap between genes which are DM in their promoter or first exon and DE genes (microarray).*** *As expected, methylation in the promoter region of a gene is often correlated to decreased expression of that respective gene. Out of 188 genes which had increased methylation in cancer tissue, 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 aromatase29. 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 process30. 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 tissue31. 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 tissue32. 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 tumour33. 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 paper12.

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

## Text Description automatically generated

Figure 4: Loading in the data

## Graphical user interface, text, application, chat or text message Description automatically generated

Figure 5: Making the annotation dataframe

## Graphical user interface, text, application Description automatically generated

Figure 6: Quality control steps and quantile normalization

## Text Description automatically generated

Figure 7: Extra annotation for design matrix

## Graphical user interface, text, application Description automatically generated

Figure 8: Differential expression analysis using limma

## Chart, scatter chart Description automatically generated

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

## Chart, histogram Description automatically generated

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

## Chart, scatter chart Description automatically generated

Figure 11: Volcanoplot from the limma analysis

## Graphical user interface, text, application, email Description automatically generated

Figure 12: Gene annotation

## Graphical user interface, text, application, email Description automatically generated

Figure 13: Gene annotation for all significant p values

## Graphical user interface, text, application Description automatically generated

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

## Table Description automatically generated

Figure 15: Session info for DE analysis script

## Methylation profiling by array

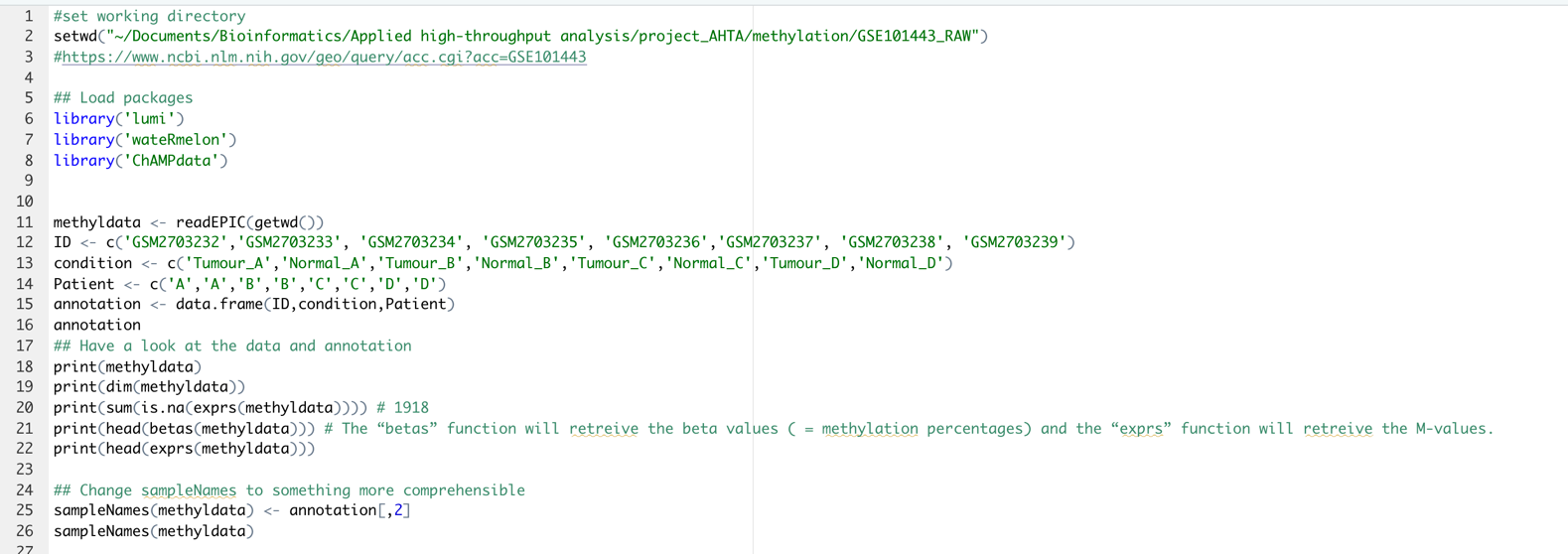


Figure 16: Loading in the data for the DM script

Graphical user interface, text, application, email

Description automatically generated

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

Chart

Description automatically generated

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

Chart, box and whisker chart

Description automatically generated

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

Graphical user interface, text, application

Description automatically generated

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

Graphical user interface, text

Description automatically generated

Figure 21: Normalization of the data and QC

Chart, line chart, histogram

Description automatically generated

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

Chart, line chart, histogram

Description automatically generated

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

Graphical user interface, text, application

Description automatically generated

Figure 24: Differential methylation analysis using limma

Chart, histogram

Description automatically generated

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

Graphical user interface, text, application

Description automatically generated

Figure 26: Plots for the limma analysis for DM

Chart, scatter chart

Description automatically generated

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

Chart, histogram, scatter chart

Description automatically generated

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

Graphical user interface, text, application, email

Description automatically generated

Figure 29: Gene annotation for the DM analysis

Graphical user interface, text, application

Description automatically generated

Figure 30: Gene annotation for DM analysis

A picture containing timeline

Description automatically generated

Figure 31: Analysis for the promoter regions

A picture containing table

Description automatically generated

Figure 32: 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

Afbeelding met tekst

Automatisch gegenereerde beschrijving

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

Afbeelding met monitor, binnen, scherm, schermafbeelding

Automatisch gegenereerde beschrijving

**Figure 34: shell script to quantify reads using Kallisto**Afbeelding met tekst

Automatisch gegenereerde beschrijving

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

Afbeelding met tekst

Automatisch gegenereerde beschrijving

**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.**

Afbeelding met tekst

Automatisch gegenereerde beschrijving

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

Text

Description automatically generated

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

Chart, scatter chart

Description automatically generated

**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

Graphical user interface, text, application

Description automatically generated

Figure 40: Loading in the multiple results

A picture containing text

Description automatically generated

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

Graphical user interface, text

Description automatically generated

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

## Graphical user interface, text, application Description automatically generated

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

## Graphical user interface, text, application Description automatically generated

Figure 44: Gene set analysis

## Graphical user interface Description automatically generated

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

## ChIP sequencing

## Afbeelding met tekst Automatisch gegenereerde beschrijving

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

Afbeelding met tekst

Automatisch gegenereerde beschrijvingAfbeelding met tekst

Automatisch gegenereerde beschrijving

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

Afbeelding met tekst

Automatisch gegenereerde beschrijving

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

Afbeelding met tekst

Automatisch gegenereerde beschrijving

**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>