Project AHTA

Boris Vandemoortele

Tristan Vanneste

21/11/2020

# Abstract

The combination of multiple single experiments has the potential to reveal new insights and mechanics in complex diseases. Therefore, we performed an omics-based study to further analyse breast cancer. We analysed the transcription differences between healthy and tumor cells with microarrays and RNA seq and linked this to the epigenomic differences (more specifically the methylation differences) with genome wide methylation profiling. Additionally, we analysed the ER cistrome to gain insight into endocrine treatment unresponsiveness in HR+ breast cancers.

## Introduction

Among women, breast cancer is the most common cause of cancer related death worldwide in both developing and developed countries (Anderson *et al.*, 2008). 1.15 million breast cancer cases are diagnosed each year worldwide, making breast cancer one of the most prevalent cancers in the world today (Parkin *et al.*, 2005). Every year, over 411,000 deaths result from breast cancer, this accounts for more than 1.6% of the female deaths from all causes. Breast cancer incidence is increasing 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, the majority of breast cancers deaths in fact occur in developing rather than developed countries. *Ferlay et al* (Ferlay J, Bray F, Pisani P, 2002) estimated that the global incidence and mortality could even increase in the future. 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 do this yearly 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 (Fackenthal and Olopade, 2007), is free for risk groups (addendum, figure x (er zou een flowchart bestaan maar ik vind die niet terug)). 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 (Peart, 2017). An exact mechanism that drives metastasis is yet to be unravelled, but some important new insights have been gained recently (Krøigård *et al.*, 2018). Next to the occurrence of metastasis, also the cancer subtype is important for disease prognosis. Breast cancer is classically categorized into tree 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 (HR), and human epidermal growth factor 2 (ERBB2). 70% of patients is HR+/ERBB2-, 15-20% is HR-/ERBB2+ and 15% is triple negative (Waks and Winer, 2019). 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). It’s mechanism of action is a competitive inhibition of binding of estrogen to the ER (Nazarali and Narod, 2014). Another possibility is the administration of aromatase inhibitors such as *Astranazole.* They decrease circulation estrogen levels by inhibiting the conversion of androgens to estrogen by the aromatase enzyme. However, this kind of treatment is only applicable to post-menopausal women (Schneider *et al.*, 2011). 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-targetted 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 (Collignon *et al.*, 2016).

In conclusion, breast cancer is a disease far from perfectly understood and needs further research to clarify the exact mechanisms behind the disease. Since cancers are known to complete change the genomic, transcriptomic and epigenomics landscape of cells, is an omics-based approach trivial to fully understand the cancer in case. We, therefore, aim to identify differences in gene expression between healthy cells and breast cancer cells, by analysing a microarray experiment and an RNA sequencing experiment. As 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 and gene promotors will be linked to differential gene expression. Additionally, the ER cistrome was analysed in order to gain insight into endocrine treatment unresponsiveness in HR+ breast cancers.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5802368/>

<https://acsjournals.onlinelibrary.wiley.com/doi/full/10.1002/cncr.23844>

# Materials and methods

## Microarray transcription profiling

Microarray transcription profiling is used to determine differentially expressed genes in breast cancer tissue compared to normal breast tissue.

The microarray transcription profiling data is obtained from ArrayExpress (Athar *et al.*, 2019) under the ArrayExpress experiment identifier E-GEOD-15852 in the raw format. 43 paired samples were collected from both tumor and normal tissues from multiethnic breast cancer patients. In total 86 samples were analysed for gene expression by using 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. After the pre-processing another QC was performed on the data. A design matrix was made for the disease with a blocking factor for patients. Finally, differential expression (DE) was analysed using *limma* (Ritchie *et al.*, 2015).  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 with the BioMart query function.

## Methylation profiling by array

The epigenomic differences between tumor tissue and normal breast tissue were analysed with methylation profiling on a microarray. Genome-wide profiling of DNA methylation in 4 pairs of matched tumor tissue and normal breast tissue was obtained from the Gene Expression Omnibus (GEO) (Edgar *et al.*, 2002) data repository under the identifier GSE101443. These 8 samples were measured with an Illumina HumanMethylation450 BeadChip. Probes for which calling p-values were insufficient were filtered together with NA values. An initial analysis was performed on the average methylation percentages for tumor tissue vs control tissue with a 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. The differential methylation analysis was also performed with LIMMA and the adjusted p-values were computed according to the Benjamini-Hochberg procedure. The design matrix was constructed with the factor condition and a blocking design for the patient. Finally, the annotation was obtained using the package *ChAMPdata* (3) in R. The significant genes, with an more lax cut-off were analysed using Enrichr(1,2).

* 1. [Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013;128(14)](http://www.ncbi.nlm.nih.gov/pubmed/23586463)
  2. [Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, McDermott MG, Monteiro CD, Gundersen GW, Ma'ayan A. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research. 2016;*](https://www.ncbi.nlm.nih.gov/pubmed/27141961)[*gkw377*](http://doi.org/10.1093/nar/gkw377).
  3. Yuan Tian, Tiffany Morris, Lee Stirling and Andrew Teschendorff (2020). ChAMPdata: Data. Packages for ChAMP package. R package version 2.20.0.
* Ik vind deze referentie niet terug in mijn ref manager

## RNAseq

Datesets were obtained from the *European Nucleotide Archive* (Leinonen *et al.*, 2011)  under project ID PRJNA142887 (Hon *et al.*, 2012). This dataset RNA comprises sequencing results from both normal breast cells (HMEC) and breast cancer cells (HCC1954). Fastq files were downloaded to the Ugent Porthos server and quality control was performed using fastqc (Andrews, 2015). Overrepresented sequences were added to an adapter file and further used in Trimmomatic’s (Bolger *et al.*, 2014) paired end Illuminaclip function. Seed mismatches were set to two, palindrome clip threshold was 30 and simple clip threshold was 10. This allowed us to obtain data of adequate quality for further processing, which was again checked by fastqc. The human hg38 reference genome and genome annotation were downloaded from the ensemble genome browser (download link in addendum, figure x) and a pseudotranscriptome was built using Kallisto (Bray *et al.*, 2016).  We then made sure that every transcript was identified by an Ensembl transcript id. A genome index was built using the standard settings in Kallisto. Reads were quantified using standard settings, which means no bootstrap was performed. The obtained transcript abundance files were then pasted together to make further processing in R possible. *EdgeR* (Robinson *et al.*, 2009) was used for the identification of differentially expressed genes. Count data was normalised using the TMM method and differential expression was and the dispersion value was fixed at 0.4. This value is proposed in the EdgeR manual in the case that no replicates are available in an experiment with human subjects. Differential expression was eventually tested by an exact T-test, which is the best solution when no replicates are available(volgens mij moet dit bij de resultaten of addendum). Threshold values for differential expression were set at an absolute log fold change > 2 and FDR < 0.05. Finally, Ensembl transcript IDs were linked to gene names using the *org.Hs.eg.db* package. Finally, the Enrichr web application (Chen *et al.*, 2013) was used for gene ontology analysis, in which transcripts that could not be linked to a gene name were left out.

## ChIPseq

Datasets were obtained from the European Nucleotide Archive under project ID PRJNA175144 (Jansen *et al.*, 2013). Only datasets from ER samples were downloaded to the Ugent Porthos server (detailed overview in addendum, figure x). Fastqc was used to perform quality control, after which overrepresented sequences were removed using Trimmomatic’s illuminaclip. A genome index was built using Bowtie2 (Langmead and Salzberg, 2012) and the same human genome as described in the RNAseq experiment. Aligning reads to the hg38 reference genome was performed with default settings. A small output formatting step was required to make this data compatible with our further analysis in HOMER (addendum, figure x line y). Tag directories were made and peaks were called using HOMER (Benner *et al.*, 2017) and the inbuilt hg38 reference genome. We used a minimum fold enrichment over 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.

(Differential peaks between both conditions were identified using a minimal fold enrichment over background of 1.2. enkel als we deze gebruiken).

# Results

## Microarray transcription profiling

After loading the data from ArrayExpress we performed multiple QC with the raw data, log transformed data and the normalized data. We used the Robust Multi-Array Average expression measure to perform background correction and quantile normalization. Differential expression analysis was done using *limma* because we argued that *limma* has more power than *sam*. A design matrix was used with the disease as factor and the patients as blocking factor. We found 5333 probes (see Figure x) to be statistically significant correcting for multiple testing with the Benjamini-Hochberg procedure (alpha= 0.05).

Chart, scatter chart

Description automatically generated

Figure 1. Volcano plot for the differential analysis of the tumor vs normal tissue with a blocking design for patients. In red are the probes that are statistically significant for an adjusted p-value of 0.05 based on the Benjamini-Hochberg procedure.

From the distributions of the p-values we could argue that there probably is a statistically significant effect distinguishing between tumor and normal tissue (Figure x). Keeping in mind that this definitely does not prove the hypothesis but only hints towards it. We have to be carefull with interpreting that many statistically significant probes since we do not expect that many significant results. If we look at the statistically significant probes and their log fold changes, we see that the log fold changes are rather low. We chose to filter out the probes that have a lower absolute log fold change than 1 to eliminate biologically not relevant probes. After filtering for the not biologically relevant probes and the probes with gene annotation we still found 39 probes to be significant (adjusted p-value of <0.05 and an absolute log fold change > 1.0).

Chart, histogram

Description automatically generated

Figure 2. Distributions of the p-values from the limma analysis between tumor and normal tissue with a blocking design for patients.

We did not perform an Enrichr analysis on these probes because we only had 39 probes that were statistically and biologically relevant.

## Methylation profiling by array

We found 0 probes to be significant on the 0.05 level for differential methylation, instead 1786 probes had an adjusted p-value of exact 0.061. 358 probes had an absolute log fold change bigger than 2 and an adjusted p-value of 0.061. For these 358 probes only 250 had a gene annotation. When selecting for genes in promoter regions (the features ‘TSS’ and ‘1stExon’) we found a total amount of significant (alpha <= 0.061) 82 genes. A positive log fold change points to a higher methylation in the tumour samples than in the control. From the 250 significant differentially methylated genes, 182 had a positive log fold change (higher methylation in tumour than in control) and 68 had a negative log fold change (lower methylation in tumour than in control).

Table 1. Top five largest absolute log fold changes for the differentially methylated genes in four pairs of matched tumour and normal breast tissue at the alpha level of 0.061.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Probe\_ID | Chrom | Pos | Gene | Feature | logFC | Adj.p.val |
| cg14763548 | 20 | 25062447 | VSX1 | 1stExon | 3.079867 | 0.0615913 |
| cg0454300 | 11 | 71955332 | PHOX2A | TSS200 | 3.192828 | 0.0615913 |
| cg08380311 | 19 | 3435252 | NFIC | Body | 3.286706 | 0.0615913 |
| cg10698928 | 8 | 65290320 | MIR124-2 | TSS1500 | 3.114323 | 0.0615913 |
| cg15963552 | 3 | 99594931 | C3orf26 | Body | 3.036477 | 0.0615913 |

We further looked at the functions of these genes using Enrichr (1,2). In Enrichr, we found some interesting results regarding the molecular function of these differential methylated genes (Table 2). We found that the differentially methylated genes had a significant influence on DNA binding and hormone activity. Since there are 182 significant genes with a positive log fold change and only 68 with a negative log fold change, we can carefully conclude that the higher methylation in tumour samples leads to different hormone activity and DNA binding. The significant difference in genes with a hormone activity function could possibly be explained by the fact that breast cancer is often driven by estrogen signaling (3). The higher methylation in tumour cells leads to difference in hormone activity which could explain the different estrogen signal in tumour cells.

Table 2. GO analysis of differentially methylated genes in four pairs of tumour and normal breast tissue.

|  |  |  |
| --- | --- | --- |
| GO category | associated genes | p-value |
|  | **Positive log fold changes Genes** |  |
| Regulatory region DNA binding | NR5A2;SALL3;WT1;ZNF536;ARRB1;MEF2D;VSX1;FOXA2 | 0.0001042 |
| transcription regulatory region DNA binding | NR5A2;SALL3;WT1;ZNF536;ARRB1;MEF2D;VSX1;FOXA2 | 0.002973346456056022 |
| sequence-specific DNA binding | ZNF274;NR5A2;SALL3;WT1;ZNF536;POU2F2;MEF2D;NR2F6 | 0.0040712830436499614 |
| hormone activity | NPY;PENK;TRH | 0.016561213055675875 |
|  | Negative logfold changes |  |
|  |  |  |

* 1. [Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013;128(14)](http://www.ncbi.nlm.nih.gov/pubmed/23586463)
  2. [Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, McDermott MG, Monteiro CD, Gundersen GW, Ma'ayan A. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research. 2016;*](https://www.ncbi.nlm.nih.gov/pubmed/27141961)[*gkw377*](http://doi.org/10.1093/nar/gkw377).
  3. Xue, M., Zhang, K., Mu, K. *et al.* Regulation of estrogen signaling and breast cancer proliferation by an ubiquitin ligase TRIM56. *Oncogenesis* **8,**30 (2019). https://doi.org/10.1038/s41389-019-0139-x

## HCC1954 cells change their expression pattern to mediate cell survival.

Finding differentially expressed genes between healthy HMEC cells and cancerous HCC1954 cells was difficult due to the fact that 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, 536 genes were found to be significantly upregulated and 497 genes were downregulated (FDR < 0.05 and abs(logFC) > 2). We then examined which cellular processes and functions these genes were associated to using Enrichr. Table 1 contains an overview of some selected hits that we believe to be interesting.

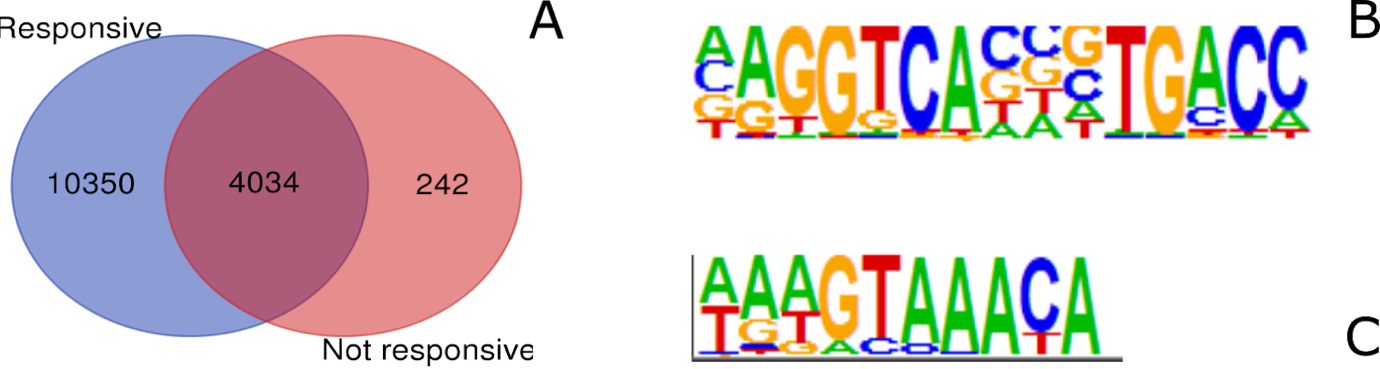
**Table 3: GO analysis of differentially expressed genes in HCC1954 cells vs HMEC cells.** 536 and 497 genes were up- and downregulated respectively in HCC1954 cells compared to HMEC cells (n=1, FDR < 0.05 and abs(logFC) > 2). Shown are the interesting hits of GO analysis in Enrichr with respective genes belonging to the given categories and associated p-values of enrichment.

|  |  |  |
| --- | --- | --- |
| GO category | associated genes | p-value |
| Upregulated genes | | |
| Positive regulation of cell proliferation | LYN, IRS1, OSR1, EIF5A2, COX17, VEGFC, IRS2, LAMC1, TNF, CXCL5, CIAO1, EREG, FOSL1, EFNB2, CCND2, CCPG1, SOX15, HAS2, TSPYL5, HBEGF | 0.001067 |
| Inflammatory response | LYN, TNFRSF6B, PTGER2, CXCL1, PIK3CB, CXCL3, TNF, CXCL5, TNFRSF10D, TNFRSF1A, IL1A, KRT16, CARD18, IKBKG | 0.001310 |
| TNF signalling | CSF2, VEGFC, CXCL1, PIK3CB, IKBKG, CXCL3, TNF, MMP9, CXCL5, BIRC2, TNFRSF1A | 0.00002804 |
| Breast cancer (ER negative) | L3MBTL3, NDUFB3, PPIL3 | 0.02686 |
| Downregulated genes | | |
| Estrogen metabolism | STS, UGT1A1, GSTA1, UGT1A3 | 0.0004162 |
| leukocyte transendothelial migration | CLDN4, CLDN3, CLDN9, RASSF5, CLDN8, CLDN23, PECAM1, PIK3CB, CLDN1, VAV1, ICAM1 | 0.00001845 |
| steroid hormone synthesis \* | UGT1A10, DHRS11, STS, UGT1A1, UGT1A5, UGT1A4, UGT1A3, UGT1A7 | 0.00002834 |
| positive regulation of monocyte chemotaxis | CCR1, CXCL10, CCL5, CXCL17, S100A7 | 0.00001742 |

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 (O’Malley *et al.*, 1962), has also been put forward as a promotor of breast cancer growth via a positive feedback loop of TNFR1/NFκB (Cai *et al.*, 2017). We also found that inflammatory response genes were significantly enriched in the list of upregulated genes (table 1, line 4; p-value = 0.001310). The role of inflammation and cancer is not straightforward, since it can both positively and negatively regulate cancer growth. On the one hand, IL-1, IL-6 and TNF mediate the formation of an inflammatory milieu by modulating the expression of tumour-promoting factors (Multhoff *et al.*, 2012). Of these, IL-1 and TNF were significantly upregulated in HCC1954 cells in our dataset. On the other hand, increased inflammation might recruit immune cells to the tumor microenvironment and therefore cause a decrease in tumor growth. However, we observe a decreased capacity for immune cell migration (table 1, row 9 and 11). We thus show that HCC1954 cells maximize their survival chances by increasing cell proliferation and creating an inflammatory microenvironment, whilst suppressing immune cell recruitment to protect themselves.

## 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. Aromatase inhibitors, which block the conversion of androgen into estrogen performed by the enzyme aromatase (ref zoeken), aim to stop this process. Our dataset contained 2 groups of patients (n=5), 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 (figure 1, A). 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. GO analysis revealed that genes involved in the nuclear receptor transcription pathway were enriched (p-value 0.003347, Bioplanet 2019) in the gene 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 is still functional in these samples.



**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 to 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 = 5). 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 to 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.

Motif analysis revealed a higher enrichment of ER elements (EREs) (figure 1, 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 1, C) motifs, while no major similarity between Fox motifs and ERE motifs seems to be present.

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 to 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 estrogenic production. A mutation in the aromatase enzyme might cause this (voor in de discussie).

# Discussion

# References

Anderson,B.O. *et al.* (2008) Guideline implementation for breast healthcare in low-income and middle-income countries: Overview of the breast health global initiative Global Summit 2007. In, *Cancer*.

Andrews,S. (2015) FASTQC A Quality Control tool for High Throughput Sequence Data. *Babraham Inst.*

Athar,A. *et al.* (2019) ArrayExpress update - From bulk to single-cell expression data. *Nucleic Acids Res.*

Benner,C. *et al.* (2017) HOMER - Software for motif discovery and next generation sequencing analysis. *Http://Homer.Ucsd.Edu/*.

Bolger,A.M. *et al.* (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*.

Bray,N.L. *et al.* (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*

Cai,X. *et al.* (2017) Inflammatory factor TNF-α promotes the growth of breast cancer via the positive feedback loop of TNFR1/NF-κB (and/or p38)/p-STAT3/HBXIP/TNFR1. *Oncotarget*.

Chen,E.Y. *et al.* (2013) Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*.

Collignon,J. *et al.* (2016) Triple-negative breast cancer: Treatment challenges and solutions. *Breast Cancer Targets Ther.*

Edgar,R. *et al.* (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*

Fackenthal,J.D. and Olopade,O.I. (2007) Breast cancer risk associated with BRCA1 and BRCA2 in diverse populations. *Nat. Rev. Cancer*.

Ferlay J, Bray F, Pisani P, et al (2002) Cancer Incidence, Mortality and Prevalence Worldwide. *GLOBOCAN*.

Hon,G.C. *et al.* (2012) Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome Res.*

Jansen,M.P.H.M. *et al.* (2013) Hallmarks of aromatase inhibitor drug resistance revealed by epigenetic profiling in breast cancer. *Cancer Res.*

Krøigård,A.B. *et al.* (2018) Identification of metastasis driver genes by massive parallel sequencing of successive steps of breast cancer progression. *PLoS One*.

Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*.

Multhoff,G. *et al.* (2012) Chronic inflammation in cancer development. *Front. Immunol.*

Nazarali,S.A. and Narod,S.A. (2014) Tamoxifen for women at high risk of breast cancer. *Breast Cancer Targets Ther.*

O’Malley,W.E. *et al.* (1962) 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.*

Parkin,D.M. *et al.* (2005) Global Cancer Statistics, 2002. *CA. Cancer J. Clin.*

Peart,O. (2017) Metastatic breast cancer. *Radiol. Technol.*

Ritchie,M.E. *et al.* (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*

Robinson,M.D. *et al.* (2009) edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*.

Schneider,R.E. *et al.* (2011) Aromatase inhibitors in the treatment of breast cancer in post-menopausal female patients: An update. *Breast Cancer Targets Ther.*

Waks,A.G. and Winer,E.P. (2019) Breast Cancer Treatment: A Review. *JAMA - J. Am. Med. Assoc.*

# Addendum

## Introduction

Afbeelding van de flowchart die toont wie in aanmerking komt voor BRCA test

## Methylation profiling by array

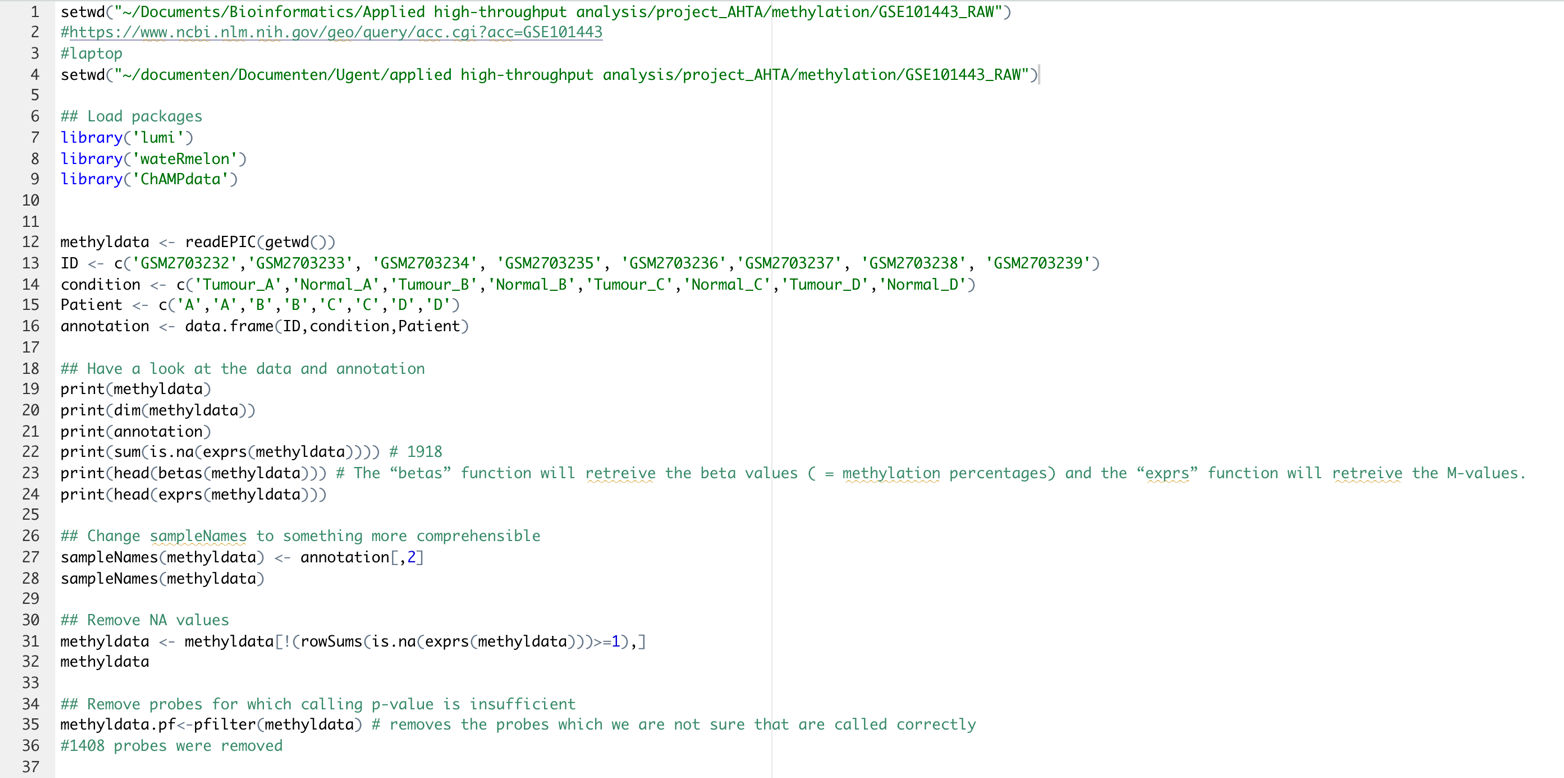


Figure 4. Loading in the data and filtering op probes which are not called correctly.

Graphical user interface, text, application

Description automatically generated

Figure 5. Comparing the average methylation between controls and tumour samples.

Chart

Description automatically generated

Figure 6. A boxplot of the Betas for each individual sample.

Chart, box and whisker chart

Description automatically generated

Figure 7. A boxplot of the average Betas for the tumour and normal tissue.

Text

Description automatically generated

Figure 8. Result of the t-test for the average Betas for tumour and normal tissue.

Graphical user interface, text

Description automatically generated

Figure 9. Normalization of the data.

Chart, histogram

Description automatically generated

Figure 10. A density plot of the Betas before and after normalization.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Figure 9. A density plot of the M-values before normalization

Figure 10. A density plot of the M-values after normalization

Graphical user interface, text, application

Description automatically generated

Figure 11. Differentially methylation analysis with LIMMA.

A picture containing chart

Description automatically generated

Figure 11 A volcano plot of the differential methylation analysis with LIMMA.

Chart, scatter chart

Description automatically generated

Figure 12. MA plot of the differential methylation analysis with LIMMA.

Graphical user interface, text, application, email

Description automatically generated

Figure 13. Loading and linking of the annotation file.

Graphical user interface, text, application

Description automatically generated

Figure 14. Interpretation of the results.

Differentially methylated genes analysis in Enrichr:

https://maayanlab.cloud/Enrichr/enrich?dataset=1e302a0a8bb41eddaf3245ad02633a95

## RNA sequencing

Afbeelding met tekst

Automatisch gegenereerde beschrijving

**Figure 15: shell script to download data, perform quality control and adapter trimming**

Afbeelding met monitor, binnen, scherm, schermafbeelding

Automatisch gegenereerde beschrijving

**Figure 16: shell script to quantify reads using Kallisto**

Afbeelding met tekst

Automatisch gegenereerde beschrijving

**Figure 17: R script to find differentially expressed genes after transcript quantification by Kallisto**

RNA sequencing gene ontology analysis in Enrichr:

Upregulated genes: <https://maayanlab.cloud/Enrichr/enrich?dataset=b796d54de48c91c9624438cbb5a73f29>

Downregulated genes: <https://maayanlab.cloud/Enrichr/enrich?dataset=9eb5354e4fb281a508614d4b3a07296f>

## ChIP sequencing

Afbeelding met tekst

Automatisch gegenereerde beschrijving

**Figure 18: shell script to download data, perform quality control and trimming**

Afbeelding met tekst

Automatisch gegenereerde beschrijvingAfbeelding met tekst

Automatisch gegenereerde beschrijving

**Figure 19: 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 20: shell script for peak calling using HOMER**

Afbeelding met tekst

Automatisch gegenereerde beschrijving

**Figure 21: 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>