Department of Genome-Oriented Bioinformatics

Technical University of Munich

Ludwig-Maximilians-Universität München

Bachelor's Thesis in Bioinformatics

Analysis of mRNA-protein Differential Expression Correspondence in Breast invasive carcinoma and Colorectal adenocarcinoma

Martin Pavlov

Department of Genome-Oriented Bioinformatics

Technical University of Munich

Ludwig-Maximilians-Universität München

Bachelor's Thesis in Bioinformatics

Analysis of mRNA-protein Differential Expression Correspondence in Breast invasive carcinoma and Colorectal adenocarcinoma

Analyse der mRNA-Protein-Differenzialexpressionkorrespondenz bei invasivem Brustkrebs und kolorektalem Adenokarzinom

Author: Martin Pavlov
Supervisor: Evans Kataka
Advisor: Prof. Dr. Dmitrij Frishman
Submitted: 15.10.2018

I confirm that this bachelor's thesis is my own work and I have documented all sources and material used.

Date **Signature**

Abstract

A portion of the RNA synthesized in every cell is specific to that cell type. Hence, differential mRNA expression carries biological significance and is assumed to be reflected by similar changes in protein levels. Most large-scale studies into the matter, however, have failed to report such concordance, which is usually attributed to advanced modes of regulation applied between transcript and protein product. On another note, cancer is one of the most researched diseases nowadays. Tumor cells are observed to undergo significant changes in their transcriptome and proteome expression patterns when compared to adjacent normal tissue. Clinical patient expression data plays a crucial role in documenting this divergence. However, investigating mRNA-protein differential expression concordance in cancer patients has been mostly overlooked as valuable asset for understanding the reprogrammed expressome of a cancer cell. Hence, in this study we integrated two large-scale patient-derived gene expression datasets with complementary protein expression data, both of which containing paired cancer-healthy samples, in order to determine mRNA-protein expression correspondence measured between tumor and normal tissue in Breast invasive carcinoma or Colorectal adenocarcinoma patients. Additionally, we consolidated the CORUM complex database, a comprehensive human protein complex collection of data, in order to cross-reference our differential expression results with it and identify variable protein complexes, which change composition in cancer cells. Taking into account the scarce paired patient protein expression data available, we report differential expression of 3 547 genes and 42 proteins, five (5) of which (11.9%) also differentially expressed at the transcriptome level, in breast cancer, and 12 616 genes and 1 119 proteins, with 883 (78.9%) coverage at the gene level, in colorectal cancer. Furthermore, we identified pathways and cellular compartments rich in differential expression, as well as several variable protein complexes with possible implications for cancer development. Our results highlight the importance of both mRNA and protein paired patient expression data for cancer research and suggest the application of the methodology described here for further cancer types, once the required datasets are made available.

Abstrakt

Ein Teil der in jeder Zelle synthetisierten RNA ist spezifisch für diesen Zelltyp. Daher trägt die differentielle mRNA-Expression biologische Bedeutung und wird angenommen, dass sie durch ähnliche Veränderungen im Proteingehalt reflektiert wird. Die meisten groß angelegten Studien zu diesem Thema sind jedoch misslungen, eine solche Übereinstimmung festzustellen, die in der Regel auf fortgeschrittene Regulierungsmechanismen zwischen Transkript und Proteinprodukt zurückzuführen ist. Auf einer anderen Seite ist Krebs eine der am meisten erforschten Krankheiten heutzutage. Es wird beobachtet, dass Tumorzellen im Vergleich zu dem angrenzenden Normalgewebe signifikante Veränderungen in ihren Transkriptom- und Proteomexpressionsmustern erfahren. Klinische Patientenexpressionsdaten spielen eine entscheidende Rolle bei der Dokumentation dieser Abweichung. Die Untersuchung der mRNA-Protein-Differentialexpressionskonkordanz bei Krebspatienten wurde jedoch meist als wertvolles Hilfsmittel zum Verständnis der umprogrammierten Expression einer Krebszelle übersehen. Daher haben wir in dieser Studie zwei groß angelegte, von Patienten stammende Genexpressionsdatensätze mit komplementären Proteinexpressionsdaten integriert, die beide gepaarte krebsgesunde Proben enthalten, um die mRNA-Proteinexpressionskorrespondenz zu bestimmen, die zwischen Tumor und Normalgewebe bei Brustkrebs oder kolorektalem Adenokarzinom gemessen wird. Zusätzlich haben wir die CORUM-Komplexdatenbank, eine umfassende Sammlung von Humanproteinkomplexdaten, konsolidiert, um unsere Ergebnisse der differentiellen Expression mit ihr zu vergleichen und variable Proteinkomplexe zu identifizieren, deren Zusammensetzung in Krebszellen verändert wird. Unter Berücksichtigung der knappen verfügbaren gepaarten Patientenproteinexpressionsdaten berichten wir über die differentielle Expression von 3 547 Genen und 42 Proteinen, von denen fünf (5) (11,9%) auch auf Transkriptomebene, bei Brustkrebs, und 12 616 Genen und 1 119 Proteinen, mit 883 (78,9%) Abdeckung auf Genebene, bei kolorektalem Adenokarzinom. Darüber hinaus identifizierten wir Pathways und zelluläre Kompartimente, die reich an differentieller Expression sind, sowie mehrere variable Proteinkomplexe mit möglichen Auswirkungen auf die Krebsentwicklung. Unsere Ergebnisse unterstreichen die Bedeutung sowohl der gepaarten patientenbezogene mRNA als auch Protein Expressionsdaten für die Krebsforschung und schlagen die Anwendung der hier beschriebenen Methodik für weitere Krebsarten vor, sobald die erforderlichen Datensätze verfügbar sind.

Table of Contents

1 Introduction

Every cell nucleus in the human body contains the full genome first assembled in the zygote. Cells differentiate, because only a portion of all genes is expressed, although the unused genes retain their potential for being expressed. Moreover, a fraction of the RNA synthesized in each cell is unique to that cell type. For that reason differential mRNA expression is considered to imply biological significance. Furthermore, it is assumed that this is reflected by similar changes in protein levels [1][2]. Proteins as biological units are a matter of attentiveness in various areas of computational biology, including diagnostic biomarker discovery and personalized medicine research. Nevertheless, research into mRNA-protein differential expression correspondence results in mostly immensely low concordance, creating uncertainties in inference from only mRNA expression data [3][4][5]. This divergence is generally accredited to further modes of regulation occurring between transcript and translation product [6]. Current large-scale proteomic studies have recognized and mapped proteins to around 85% of the protein-coding genes in human, many of which turned out to show tissue-specific expression [7]. Clarifying the effect of tissue-specific protein expression is a big step in the direction of comprehending differentiation-driven phenotypic modulation in diverse pathophysiological environments.

On another note, most similar studies targeting different types of cancer, one of the most researched diseases nowadays, are based solely on transcriptome or proteome level expression, rather than incorporating both into the same study [8][9][10]. Recent scientific efforts have established the relationship between differentially expressed mRNA and mRNA-protein correlations in ovarian cancer [3]. It has been detected, that differentially expressed mRNAs have significantly better correlation with their protein. However, research on large-scale mRNA-protein differential expression correspondence in cancer is still considerably new, providing undisputed uniform results on the matter that other studies have to rely on [5]. Furthermore, most of the paired cancer-healthy expression data from cancer that is publicly available is measured from cell line cultures rather than from actual patients, which in most cases doesn't reflect the expression profiles in vivo [11]. This is of importance, because techniques like precision personalized oncology and personalized drug development, which revolve around patient expression profiling, are currently making great strides in the domain of research and treatment regarding various cancer types, given their relatively short development history [12][13]. Additionally, paired expression data containing tumor as well as adjacent normal tissue is crucial to cancer studies, such as this one, because unaffected tissue samples are used as normal controls to identify genes and proteins that are differentially expressed between both states. The importance of normal tissue in identifying cancer-related somatic alterations has already been established by research into DNA single-base substitutions as well as insertions and deletions [14]. Additionally, scientific verification has indicated how transcriptomic data from adjacent normal tissue is an advantage for biomarker detection and tumor classification [15][16]. Evidence for that is the discovery of recurring gene expression signatures in hepatocellular carcinoma (HCC) associated with patient survival, which were present in tumor-adjacent healthy tissue, but not in the tumor tissue itself [15].

In order to assess similarity and dissonance of gene and protein expression in cancer paired patient-derived data, we consolidated RNA-Seq gene expression data combined with corresponding Reverse Phase Protein Array expression data for 14 relevant cancer types from The Cancer Genome Atlas (TCGA) [17], as well as protein expression data for colorectal cancer provided by the Clinical Proteomic Tumor Analysis Consortium (CPTAC) [18]. Unfortunately, because of the lack of actual healthy samples in most of the proteomic datasets, we were able to perform a complete pairwise cancer-healthy comparison on both transcriptomic and proteomic levels for two types of cancer in particular- Breast invasive carcinoma (BRCA) and Colorectal adenocarcinoma (COADREAD).

BRCA as one of the most widely spread cancer types is the second cause of cancer death in women around the world [19]. Thus it is a subject of extensive research into mRNA and protein differential expression analysis and patterning, which has been able to identify and verify various cancer-specific pathways and expression profiles later used in the treatment of the disease [20][21][22]. However, to what extend those differential expression patterns concord between transcriptomic and proteomic levels and between cancer and normal tissue has been mostly ignored as valuable insight for the field of oncology.

The term colorectal cancer describes the characterized tumors found in the lining of the large intestine. While both genders are equally at risk, this cancer type is more prominent among men, since they are more likely to develop rectal cancer. Technological advancements regarding high throughput data generation in the domain of biomedicine has allowed comprehensive characterization of its genomic, transcriptomic and proteomic variations [23]. Nevertheless, systematical comparison of gene-protein differential expression concordance in COADREAD has been overlooked as an asset for diagnosis and therapeutic strategies for individualized care of patients.

This project's main goal was to adopt the publicly available paired patient data mentioned above in order to assess mRNA-protein expression correspondence, describe differential expression patterns specific to a cancer type, discuss larger processes made up of the activities of differentially expressed gene products, identify cellular compartments rich in differential expression, as well as cross-reference differentially expressed proteins to the CORUM *Homo Sapiens* protein complex database [24] in order to distinguish protein complexes modulated by cancer-type specific protein expression.

In this study we report 3 547 genes and 42 proteins differentially expressed in breast cancer and 12 616 genes and 1 119 proteins in the colorectal dataset. For each cancer type, we were able to identify multiple pathways enriched in differential expression at the protein level and measure the degree of compliance with mRNA differential expression. Ultimately, we discuss several variable protein complexes, such as the Respiratory Chain Complex I, and the possible importance of their variability for cancer development.

2 Methods and Materials

2.1 mRNA expression data

We downloaded normalized mRNA expression (TCGA RNA-SeqV2) for 14 cancer types from the Broad Institute (http://gdac.broadinstitute.org/) [25]. We then filtered for any cancer type having paired healthy and cancer expression data and obtained 14 cancer types with at least ten (10) patient-healthy expressome pairs. These cancer types were: Bladder urothelial carcinoma (BLCA), Breast Invasive Carcinoma (BRCA), Colon Adenocarcinoma (COAD), Colorectal Adenocarcinoma (COADREAD), Head and Neck squamous cell carcinoma (HNSC), Kidney chromophobe (KICH), Kidney renal clear cell carcinoma (KIRC), Kidney renal papillary cell carcinoma (KIRP), Liver hepatocellular carcinoma (LIHC), Lung Adenocarcinoma (LUAD), Lung squamous cell carcinoma (LUSC), Thyroid carcinoma (THCA), Prostate Adenocarcinoma (PRAD) and Stomach and Esophegael carcinoma (STES).

2.2 RPPA protein expression data

Reverse Phase Protein Array (RPPA) data with Gene annotation corresponding to 13 of the cancer types mentioned in chapter 2.1 (excluding Colorectal adenocarcinoma — COADREAD) was downloaded from the same source. Subsequently, we looked for paired cancer-healthy samples for each cancer type to ensure adequate comparison to the mRNA expression analysis. Only Breast Invasive Carcinoma (BRCA) consisted of paired healthy and corresponding cancer samples. This dataset covered 226 distinct proteins measured across 45 pairs of cancer-healthy samples. Furthermore, 212 of them (93.8%) were also present in their gene form in the breast cancer gene expression dataset. The remaining twelve (12) protein expression datasets (BLCA, COAD, HNSC, KICH, KIRC, KIRP, LIHC, LUAD, LUSC, THCA, PRAD and STES), together with their corresponding mRNA expression profiles were discarded. Subsequent analysis was restricted to BRCA and COADREAD only.

2.3 CPTAC protein expression data

A mixed-sample protein expression dataset consisting of 30 normal colon samples, provided by the Vanderbilt University, and 90 solid colorectal tumor samples, provided by TCGA was obtained from the supplementary material of a publication, studying the proteogenomics of human colon and rectal cancer [4]. This dataset contains quantile-normalized and log-transformed spectral count data for 3718 proteins, measured in both normal and cancer samples, 3598 (96.7%) of which had their respective encoding gene expression captured in the COADREAD mRNA dataset.

2.4 CORUM protein complex data

The complete CORUM dataset (01.7.2018 release) consisting of 4264 annotated protein complexes, having 17 260 protein members, from which 6121 unique, was downloaded and prepared for analysis. The data provided consists of Complex CORUM ID, Complex Name, Organism, Complex Name Synonyms, Cell Line, Subunits (Uniprot IDs), Subunits (Entrez IDs), Protein complex purification method, GO ID, GO description, FunCat ID, FunCat description, Subunits (Protein name), Subunits (Gene name), SWISSPROT organism, Subunits comment, Complex comment, Subunits (Gene name synonyms), Disease comment and PubMed ID. Of the 4264 protein complexes, 2847 refer to human protein complexes. The mean complex size in the dataset amounts to 4.047842 members.

2.5 Identification of differentially expressed genes from mRNA expression data

For differential gene expression analysis of BRCA data we used the TCGAbiolinks [26] R package, as it is a robust and efficient way to facilitate analysis of TCGA data. We carried out the analysis as follows: We first extracted healthy samples based on their TCGA barcodes. We then paired the healthy samples with their corresponding cancer samples and used these barcodes to query the Genomic Data Commons (GDC) [93] as well as prepare the expression data for differential expression analysis (DEA) in TCGAbiolinks. Subsequently, a filtering step was applied to ensure that the following analysis is carried out on mRNA with mean abundance across all samples, higher than a chosen quantile mean. A threshold of 0.25 was chosen for this, as it was shown to best fit our data. Next, the actual DEA was carried out, using the edgeR [27] statistical method, where a negative binomial generalized log-linear model to the read counts was fitted for each gene. The log2-fold-change cutoff was set at 1 and a p-value threshold of 0.05 was considered statistically significant. Differentially expressed genes fulfilling the above criteria were extracted and prepared for subsequent enrichment analysis.

For COADREAD, we took a different approach, as it is not supported by the TCGAbiolinks package and is not eligible for a GDC query and subsequent DEA as described above. We used the xseq [28] R package to filter out genes with low or insignificant expression in the dataset as follows: first, we separated the 51 cancer samples and the 382 healthy ones. We then log2-transformed the expression values for both sample groups. Next, we used the EstimateExpression function of xseq on each group, which incorporates a mixture modelling approach to estimate whether a gene is expressed, giving it a weight ranging between 0 and 1. All genes with a weight less than 0.8 (corresponding to expression value of 4.5 for cancer and 2.7 for healthy (Figure 1) were considered to be not expressed in the respective sample group and all genes with a weight less than 0.8 in both cancer and healthy sample groups were filtered out of the dataset for subsequent analysis. The genes, that were found to be expressed

in only one of the sample groups, were assigned an expression value of 0 across all samples of the other group to ensure a bigger contrast in the DEA.

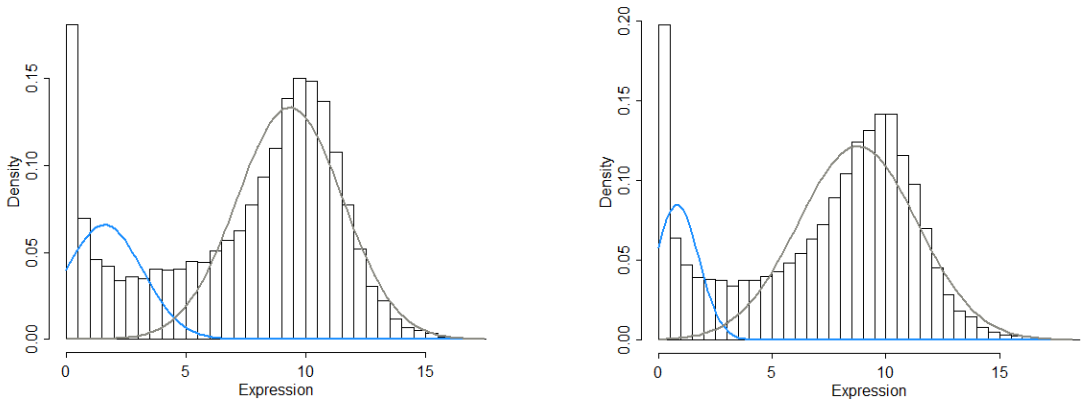

Figure 1: Expression distribution in COADREAD for cancer (left) and healthy (right) samples

The EnrichmentBrowser [29] R package, consolidating LIMMA [30] was chosen as tool for the differential expression analysis. We used Benjamini-Hochberg [31] for multiple testing correction and p-value adjustment. Genes with adjusted p-value lesser than 0.05 and absolute log2-fold-change greater than 2 were considered differentially expressed.

2.6 Identification of differentially expressed proteins from RPPA data

For DEA on the BRCA RPPA data we used RPPApipe [32], as it offers a convenient analysis pipeline specific for RPPA data analysis. The expression data was uploaded onto the job server and two paired sample classes were defined: the 45 healthy samples in the dataset and their respective cancer counterparts. Subsequently, the following parameters were set for preprocessing of the expression matrix. “Median centering” was chosen as scaling method and the option, allowing controls to be used as reference for scaling was disabled, since our data does not contain control samples. Missing values were set to be imputed from the k-Nearest Neighbor and log(2)-transformation of the expression values was turned off, as it was observed to produce NA values. Additional Ensembl ID annotation was performed on the proteins in the dataset to make cross-referencing of the output to the mRNA DEA results easier. We used LIMMA [30] to perform differential expression analysis and Benjamini-Hochberg [31] for multiple testing correction. A p-value threshold of 0.05 and a Fold-change cutoff of 2 were set as requirement for a protein to be considered differentially expressed.

2.7 Identification of differentially expressed proteins from CPTAC protein expression data

Since our protein expression data coming from CPTAC is already quantile-normalized and log-transformed, no further normalization or filtering steps were applied. For the DEA, we used an in-house script incorporating LIMMA [30] and Benjamini-Hochberg [31] for p-value adjustment. Proteins with a p-value lesser than 0.05 and an absolute log2-fold-change greater than 1 were considered differentially expressed.

2.8 Gene Set Enrichment Analysis

The next step was to cluster the differentially expressed proteins into functional categories. We used the DAVID (Database for Annotation, Visualization and Integrated Discovery) online module for our Gene Set Enrichment Analysis (GSEA) [91]. For each of the two cancer types, the DE Protein Set was uploaded onto the job server separately. Ensembl IDs were selected as identifier. For COADREAD, 31 of the 1119 protein IDs couldn't be mapped to DAVID IDs, whereas all of the 42 BRCA proteins were successfully mapped, resulting in 40 DAVID IDs, since in this protein set we have two pairs of proteins, which are represented by the same genes, but have different modification (ENSG00000137154 and ENSG00000197122). The input protein set previously used for DEA, containing all protein IDs was uploaded as background for the enrichment analysis. "Functional Annotation Chart" was chosen as results viewing option. After the lists were submitted, we extracted results for Gene Ontology's (GO) [94] three categories of interest– Biological Process, Cellular Component and Molecular Function class databases. Those showed good coverage of our gene sets: GOTERM_BP_ALL included 94.7% of the COADREAD DE proteins and 100.0% of the BRCA ones, GOTERM_CC_ALL- 97.1% from COADREAD and 100.0% for BRCA and GOTERM_MF_ALL with 94.9% for COADREAD, 100.0% for BRCA. For BRCA, we also downloaded results based on GAD disease classes, as well as UniProtKB [92] Keywords reflecting functional categories.

2.9 GO Term Clustering

In order to cluster GO terms from GSEA and to be able to interpret them in a meaningful way in the "Results and discussion" panel, we used REVIGO [33]. REVIGO summarizes long lists of GO terms by finding a representative subset of the terms and then offers multiple visualizations, like multidimensional scaling or graph-based views, which reveal distance between clusters as well as their significance. For our purposes, we loaded the GO term sets together with their corresponding Benjamini-Hochberg-corrected p-values as significance measurement. GO terms with p-value higher than 0.05 were not used.

2.10 Identification of Variable Protein Complexes from the CORUM Database

In this part of the project we aim at assessing the proportion of members found to be differentially expressed in either breast or colorectal cancer protein expression datasets for each protein complex individually. In order to achieve this we adapted a recent methodology developed for this purpose specifically and published by Ori et al. [34] We started by ruling out all complexes that are not observed in *Homo Sapiens,* as well as small protein complexes with less than five members, because including them would create a bias when calculating the proportion of differential expression in the complex. This step left us with 604 protein complexes. After that, in order to discriminate redundant complexes with very similar composition, we applied another filtering step which required sorting of the complexes by size, from biggest to smallest. In consecutive order, starting from the top, we iteratively compared one complex to all subsequent ones and removed those, which shared more than 50% of their members with it. This step was repeated until the end of the list and resulted in a subset of 363 non-redundant protein complexes. Subsequently, we investigated which of those complexes were sufficiently represented in our breast and colorectal protein expression datasets. A stringent threshold was applied, requiring a protein complex to have at least 50% quantified members in the respective cancer type dataset in order to be considered of interest for further analysis. This moreover reduced the dataset to 171 complexes quantified in colorectal cancer and only 5 in breast carcinoma, again owing to the small size of the initial RPPA dataset. In the final step we examined what segment of the quantified proteins in a complex was also found to be differentially expressed by the DEA described in 2.4 and 2.5. The requirement from the original methodology [34] was used, requiring that at least 20% of the quantified members have to be also differentially expressed in order for a complex to be considered “variable”. In this step we also extracted the direction of regulation for each differentially expressed member in order to interpret the results in a more comprehensive manner. In the end we identified 83 variable complexes in colorectal cancer and 3 in breast cancer.

3 Results and discussion

3.1 Differentially expressed genes in Breast invasive carcinoma (BRCA)

DEA results for BRCA were extracted (Supplementary Table S1). Of the 20 531 genes in the original dataset, 3 547 (17,2%) were found to be differentially expressed. The amount of up-regulated (log2-fold-change > 1) and down-regulated (log2-fold-change > 1) genes in the dataset is relatively identical (Figure 2).

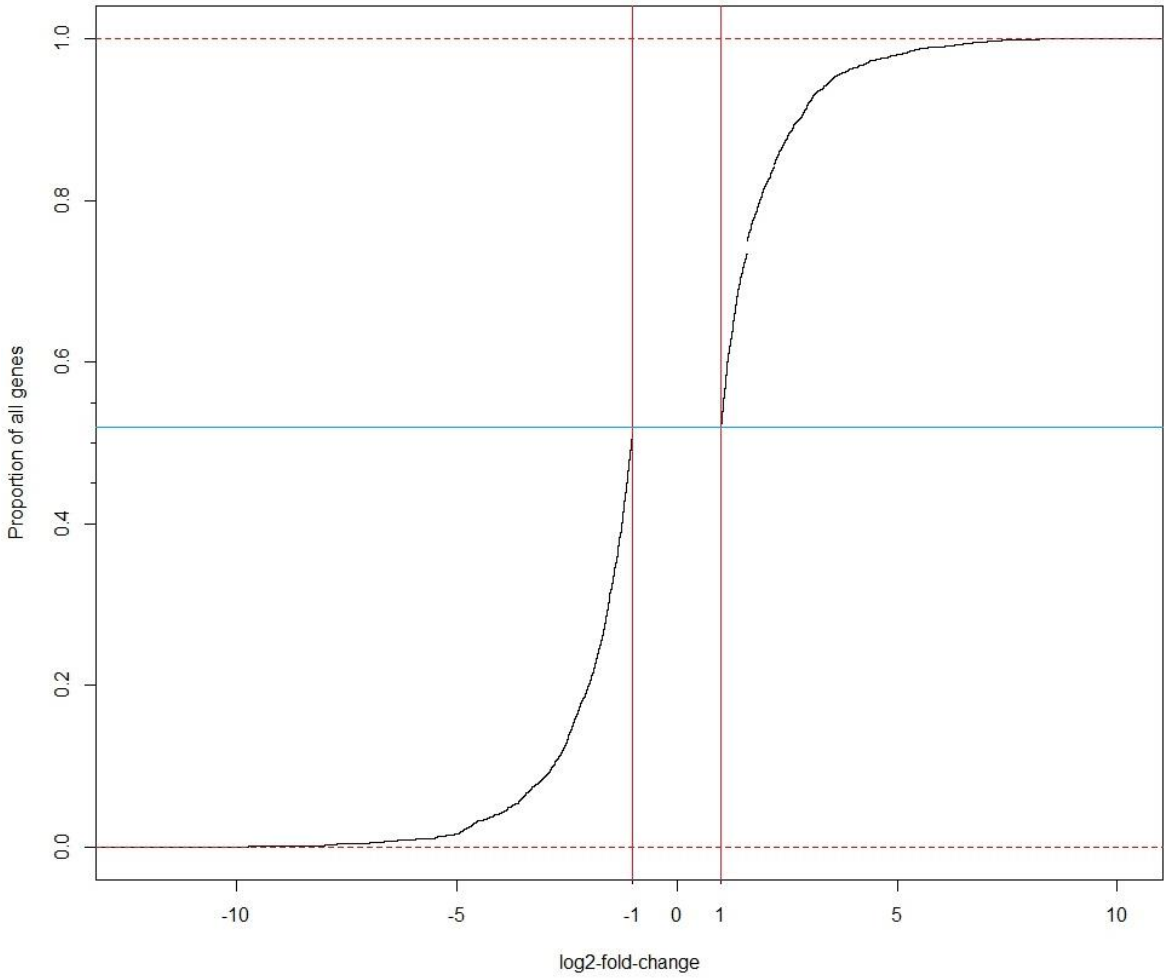

Figure 2: Cumulative distribution of log2-fold-changes of differentially expressed genes in BRCA. Split between values lesser than -1 (1840) and greater than 1 (1707) represented by blue line.

3.2 Differentially expressed genes in Colorectal Adenocarcinoma (COADREAD)

The Colorectal adenocarcinoma gene expression dataset showed a considerably higher ratio of differential expression, compared to BRCA. Of the 20 531 genes in the dataset, 12 616 (61.4%) showed differential expression between cancer and healthy conditions (Supplementary Table S2). From those 12 616 genes, 6 678 turned out to be down-regulated and 5 938 up-regulated.

3.3 Differentially expressed proteins in BRCA

Of the 226 proteins contained in the BRCA RPPA dataset, 42 (18.5 %) were found to be significantly differentially expressed. A volcano-plot (Figure 2) of the differentially expressed proteins was generated, showing how their fold-change correlates with the p-value. To understand the role these differentially expressed proteins may have during tumorigenesis, we first mapped their protein IDs to their corresponding Ensembl IDs and using those Ensembl IDs, they were cross-referenced with the list of differentially expressed genes from the BRCA. Seven (7) of the 42 differentially expressed proteins were found to be up-regulated (Figures 3 and 4) and five (5) showed differential expression at the transcriptome level as well (Supplementary Table S3). The last column of the table refers to the set of protein complexes.

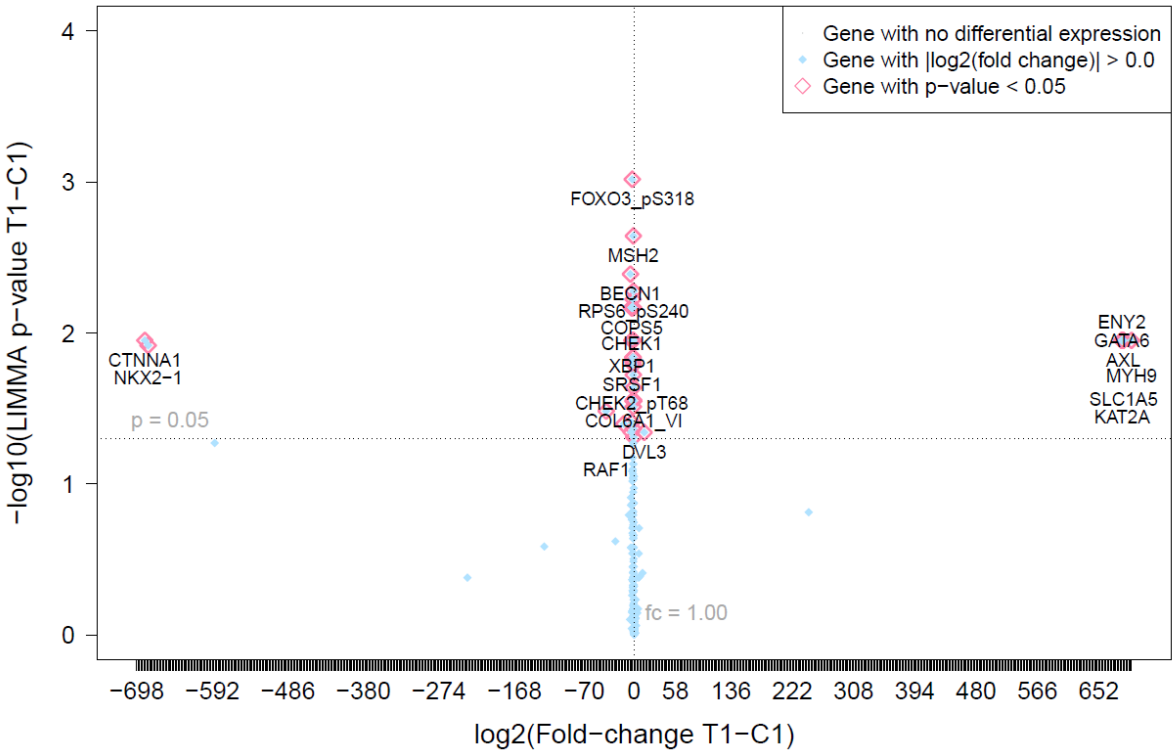

Figure 3: Differentially expressed proteins in BRCA with their respective p-value and fold change

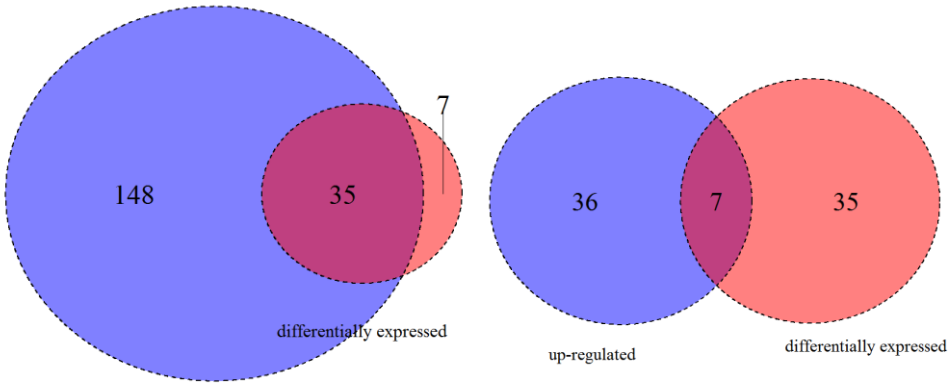

Figure 4: BRCA regulation and differential expression intersections

3.4 Differentially expressed proteins in COADREAD

The colorectal protein dataset, containing expression values for 3588 proteins, produced results, pointing to 1119 proteins being differentially expressed, which makes up a differential expression ratio of 31.1% for the dataset (Supplementary Table S4). Of those 1119 differentially expressed proteins, 883 (78.9%) overlap with the gene DEA results for colorectal cancer, meaning they were found to be differentially expressed on both transcriptome and protein expression levels. Furthermore, the majority- 644 (57.5%) of the proteins were found to be up-regulated (fold-change > 1) (Figure 5).

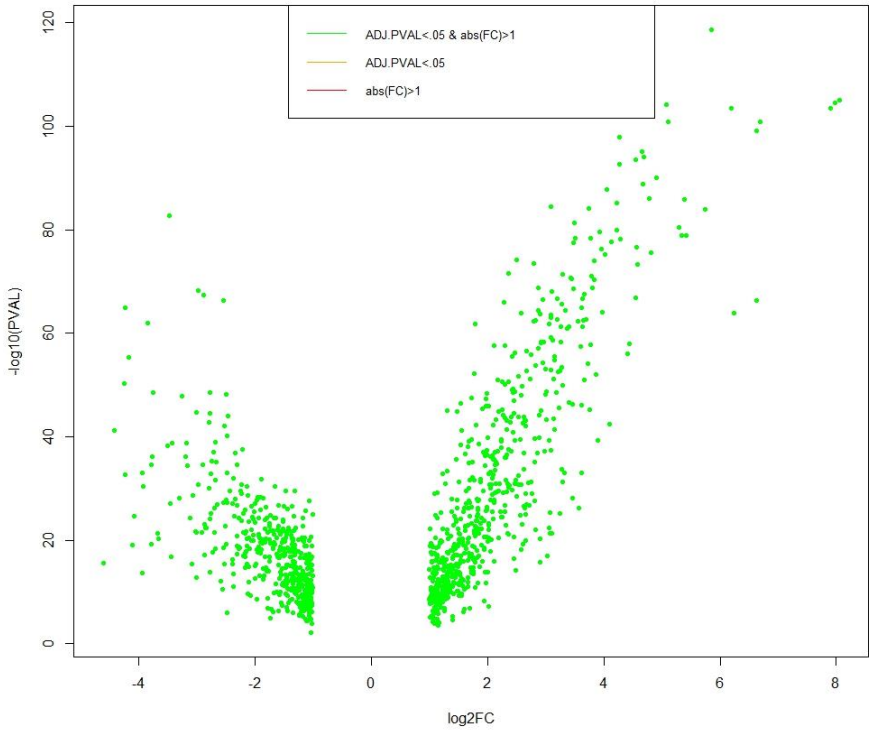

Figure 5: Differentially expressed proteins in COADREAD with their respective p-value and fold change

3.5 GSEA Results in BRCA

Gene Set Enrichment Analysis results were downloaded from DAVID for all categories described above, but owing to the small input protein set for BRCA, some categories were underrepresented with regard to significantly enriched terms or did not yield any significant terms at all (Table 1).

Table 1: BRCA Enrichment Analysis statistics. Terms are considered significant if they are assigned a Benjamini-Hochberg-corrected p-value not higher than 0.05. Last column refers to percentage of protein IDs assigned to each significant term, averaged.

Category	Enriched Terms	Significant Terms	Avg. Enrichment of Significant Terms
GAD disease classes	9	5	39%
UniProtKB Keywords	38	25	30.5%
GO terms: Biological Process	67	13	24.8%
GO terms: Cellular Component	3	0	0%
GO terms: Molecular Function	0	0	0%

3.5.1 Most differentially expressed proteins relevant to cancer

As expected, the differentially expressed protein set in BRCA turned out to be significantly enriched for the GAD disease class “Cancer” (Supplementary Table S5). A predominant part (60.0%) of the proteins are associated with the disease, making it the top hit, followed by the class of neurological diseases with 40.0%. Breast cancer has been observed to systematically imply neurological complications, some of the most common syndromes being cerebellar degeneration, retinopathy and encephalitis [35]. Most of these are mediated by antibodies against known neural antigens, although some cases appear to be mediated by non-humoral mechanisms [35].

In descending order of representation after neurological comes the immune disease class, associated with 32.5% of the proteins. Interestingly, none of those proteins are actually unique to the disease class, with most of them being also relevant to cancer. This interplay is expected, since the connection between cancer and immunity, in particular autoimmunity, is well established. Cancer has been implicated in some autoimmune disorders (AID), such as scleroderma and myositis [36].

3.5.2 Protein phosphorylation as biomarker in breast cancer

Inspecting the Keyword Set from UniProtKB (UP) (Supplementary Table S6), we identified “Phosphoproteins” as the most enriched term with 34 (85%) of the proteins being associated to it and a Benjamini-Hochberg-corrected p-value < 0.01, showing the highest significance of the term for our gene set. Protein phosphorylation is regarded as a very important and widespread molecular regulatory mechanism. The development of phosphoproteins in extracellular vesicles has been proposed as biomarker for breast cancer [37]. Yet, our study suggests that differential expression of such proteins in actual breast tissue may be considered a valid biomarker for breast cancer as well.

3.5.3 Cellular glucose homeostasis pathway

The five Biological Process GO terms with most significant p-value in breast cancer: GO:0071322, GO:0071326, GO:0071333, GO:0071331 and GO:0001678 (Supplementary Table S7) turned out to be clustered together by REVIGO and "cellular glucose homeostasis" was chosen as best representation of this pathway. Furthermore, all five terms are represented by the very same five genes: BAD (ENSG00000002330), ENY2 (ENSG00000120533), RAF1 (ENSG00000132155), FOXO3 (ENSG00000118689) and XBP1 (ENSG00000100219) in our dataset. The only exception to this is "cellular response to carbohydrate stimulus (GO:0071322)" which also includes the MYC (ENSG00000136997) gene, a proto-oncogene encoding a bHLH transcription factor.

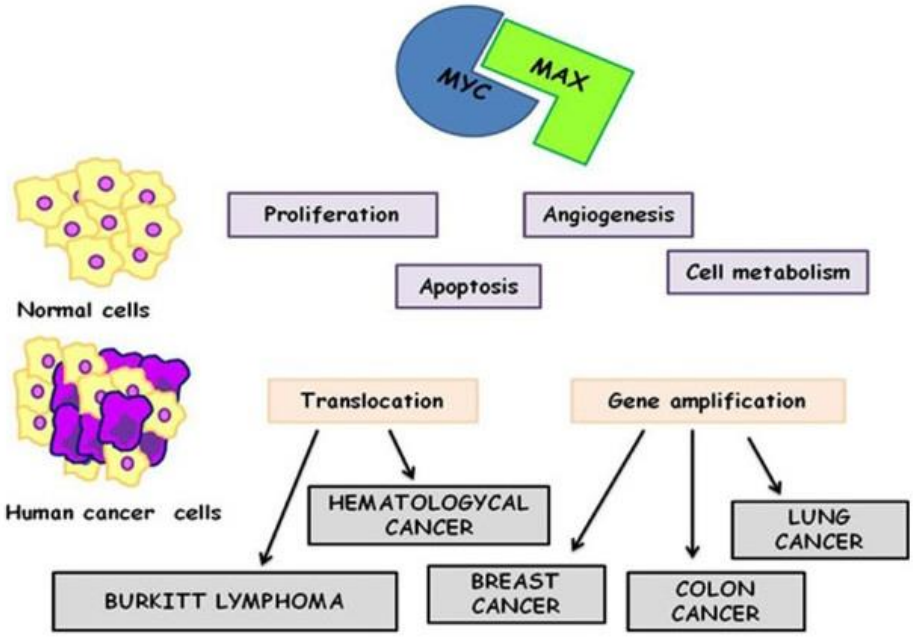

Figure 6: MYC alterations in human cancer overview (figure from paper [38])

It has been documented, that Myc is characterized by constitutive expression in several cancer types [38]. It is also believed to regulate expression of up to 15% of all genes in the human genome by binding enhancer box sequences (E-boxes) [39][40]. Given the fact, that a threshold level of pathological expression of Myc has been established for multiple cancer types [40] (Figure 6), including breast cancer [41], this supports the hypothesis that steady elevated Myc expression drives tumorigenesis [42]. In our mRNA and protein expression datasets, however, Myc was found to be differentially expressed on both levels and actually down-regulated in cancer compared to healthy samples, which is in contradiction to what previous research claims [43]. A possible explanation for this might be the small proteomic dataset, on which the differential protein expression analysis was performed.

3.6 GSEA Results in COADREAD

As a result of the much bigger size of the differentially expressed protein set in COADREAD compared to BRCA (1119 in COADREAD, 42 in BRCA), the enrichment analysis yielded notably more enriched as well as significant enriched (p-value <0.05) GO terms (Table 2).

Table 2: COADREAD Enrichment Analysis statistics for all three GO term categories. Significant terms have Benjamini-Hochberg-corrected p-values < 0.05.

Category	Enriched Terms	Significant Terms	Avg. Enrichment of Significant Terms
GO terms: Biological Process	275	42	8.4%
GO terms: Cellular Component	34	11	7.8%
GO terms: Molecular Function	79	16	12%

After using REVIGO to visualize the significant terms of each GO category separately, we identified multiple terms of high importance (dispensability = 0) in our dataset as follows:

3.6.1 Biological processes

3.6.1.1 Response to xenobiotic stimulus relation to colorectal cancer

The GO term (GO:0009410) with highest uniqueness (0.952) and lowest dispensability (0) in the "biological process" category as assigned by REVIGO refers to "response to xenobiotic stimulus", which by definition includes any process resulting in a change of a cell as a result of xenobiotic compound stimulus. This term was found to be related to 26 differentially expressed proteins in the colorectal dataset in our study (Supplementary Table S8). A few studies have been able to link xenobiotic metabolism to inhibition of colorectal tumorigenesis. One of them states, that polymethoxyflavones (PMFs), xenobiotic proteins exclusively found in citrus plants, inhibit benzo[a]pyrene/dextran sodium sulfate-induced colorectal carcinogenesis by regulating xenobiotic metabolism [44]. Another one provides insight on the anticarcinogenic activity of the cytochrome P450 (CYP), the glutathione S-transferase (GST) family and the UDP-glucuronosyltransferase (UGT) superfamily xenobiotic biotransformation mechanism concerning colorectal cancer using available clinical data [45]. Although some cancer types are found to be associated with and maintained by chronic inflammation [46], sometimes induced by cancerogenic xenobiotics like some of the heavy metals [47], there are currently not many research results linking the term "response to xenobiotic stimulus" directly to colorectal cancer. In our findings, most proteins associated with the term are found to be down-regulated and differentially expressed at the transcriptome level as well (Table 3). These results point to the conclusion that response to xenobiotic stimulus downregulation on both transcriptomic and proteomic levels may play an important role in Colorectal adenocarcinoma development.

Table 3: Differentially expressed proteins in COADREAD, associated with "response to xenobiotic stimulus", divided by regulation and transcriptome level inclusion

Category	Proteins
down-regulated & diff. exp. at transcriptome level	15
down-regulated & not diff. exp. at transcriptome level	1
up-regulated & diff. exp. at transcriptome level	7
up-regulated & not diff. exp. at transcriptome level	3

3.6.1.2 Nucleobase, nucleoside and nucleotide metabolism

Another indispensable significantly enriched GO term in colorectal cancer is “nucleobase-containing small molecule metabolic process” (GO:0055086). This term describes all cellular chemical reactions and pathways involving a nucleobase-containing small molecule: a nucleobase, a nucleoside or a nucleotide. There are 130 differentially expressed proteins connected to this term. Most of them are down-regulated and also differentially expressed at the gene level (Table 4).

Table 4: Differentially expressed proteins in COADREAD, associated with "nucleobase-containing small molecule metabolic process", divided by regulation and transcriptome level inclusion

Category	Proteins
down-regulated & diff. exp. at transcriptome level	57
down-regulated & not diff. exp. at transcriptome level	17
up-regulated & diff. exp. at transcriptome level	43
up-regulated & not diff. exp. at transcriptome level	13

It has been well established that metabolic processes in cancer cells are reprogrammed. Hence, these cells heavily depend on glycolysis for ATP production. A study indicates that defects in the antioxidative defense are more common in aging patiens with colorectal cancer, resulting in the enhanced activity of enzymes like adenosine deaminase (ADA) and xanthine oxidase (XO), which are main parttakers in the purine nucleotides decomposition [48]. The intensity of those metabolic processes and pathological rewirings is believed to depend on the severity and stage of the tumor.

3.6.2 Cellular components

3.6.2.1 Extracellular space

For the “cellular components” GO category, the two significantly enriched terms with highest uniqueness and 0 dispensability were “extracellular region” (GO:0005576, frequency = 25.63%, uniqueness = 0.857) and “extracellular space” (GO:0005615, frequency = 8.46%, uniqueness = 0.825) (Supplementary Table S9). Owing to the close relation of the two terms and the fact, that “extracellular space” is a sub-term of “extracellular region” with a lower occurrence frequency, we chose to discuss it as it is more specific.

Extracellular space refers to the part of a multicellular organism outside the cells proper, which is outside the plasma membranes and occupied by fluid. Therefore, proteins in the interstitial fluid or blood plasma that are emitted by a cell are adherent to this term. In our colorectal cancer dataset, this term has higher representation by down-regulated proteins, amounting to 53.0% down-regulation of the term (Table 5).

Table 5: Differentially expressed proteins in COADREAD, associated with "extracellular space", divided by regulation and transcriptome level inclusion

Category	Proteins
down-regulated & diff. exp. at transcriptome level	54
down-regulated & not diff. exp. at transcriptome level	32
up-regulated & diff. exp. at transcriptome level	44
up-regulated & not diff. exp. at transcriptome level	32

It is known, that primary colorectal cancer, as well as its metastases, secrete exosomes that deviate by content and function from healthy colorectal cell-derived exosomes [95]. A study into the proteome of colorectal cancer-originating exosomes has revealed that signal transduction proteins, such as the proto-oncogene tyrosine-protein kinase Src, were significantly enriched in differential expression [49]. Another study targeting over 600 miRNAs potentially inhibiting colorectal metastases development in the liver found an intriguing interaction between two miRNAs (miR-551a and miR-483) and creatine kinase B [50]. Creatine kinase B (CKB) is an enzyme secreted by metastatic cells into the extracellular space, where it phosphorylates creatine, which is then used for ATP-induced metastases proliferation. CKB was therefore found to be highly expressed in metastatic tissue, which further consolidates the importance of extracellular space as target for research in cancer.

3.6.2.2 Vacuolar lumen

The vacuole has an essential role in cell function and survival. The function and the content of the vacuole (vacuolar lumen) vary greatly according to cell type. Thus, it is a potential target of interest in cancer research, as alterations in the substance of the vacuolar lumen between normal and tumor cells may reveal targets for cancer therapy. As an example supporting this thesis we take an in vitro experiment performed on human colon adenocarcinoma HT-29 cell line [51]. After treating the cell line culture with NH(4)Cl in concentration reflecting the one in large intestine lumen, an increase in the volume of vacuolar lysosomes was observed. In a cascade manner, ornithine decarboxylase activity has been found to be hindered and consequently polyamine synthesis was negatively affected. With polyamines being obligatory for cell growth, this resulted in HT-29 growth inhibition. Another study into phosphatase and tensin homolog (PTEN)-modulated cell morphogenesis in Caco-2 colorectal cancer cell line has captured the changes in intracellular vacuoles as a result of knockdown of PTEN [52]. Caco-2 cells with inactive PTEN have been found to show inhibited cell division control protein 42 homolog (Cdc42) activation, which in turn results in intracellular vacuoles untypical to high-grade colorectal cancer.

In our study the term was found to be strongly represented by down-regulated differentially expressed proteins (68.5%), with most of them also being supported by down-regulation at the gene level (70.8%) (Table 6). These results together with the findings described above suggest that the vacuolar lumen indeed undergoes changes, characteristic to colorectal cancer.

Table 6: Differentially expressed proteins in COADREAD, associated with the GO term "vacuolar lumen", divided by regulation and transcriptome level inclusion

Category	Proteins
down-regulated & diff. exp. at transcriptome level	17
down-regulated & not diff. exp. at transcriptome level	7
up-regulated & diff. exp. at transcriptome level	9
up-regulated & not diff. exp. at transcriptome level	2

3.6.2.3 Respiratory chain complex

It has been established that cancer cells undergo changes in cellular glucose metabolism. An important organelle for this process is the mitochondrion. The electron transport chain, also respiratory chain, describes a chain of four complexes (I, II, III and IV) on the inner membrane of the mitochondrion that transfer electrons from electron donors to electron acceptors by the means of redox reactions. The enriched term in our study “respiratory chain complex” (GO:0098803) describes any protein complex that is part of the respiratory chain and is associated with 27 differentially expressed proteins in our colorectal cancer dataset. Almost all of the proteins turned out to be down-regulated in cancer (92.5%) with little correspondence to transcript level differential expression (38.9%). Further cross-referencing of those proteins to GO sub-terms revealed that 20 of them (74%) are actually associated with Respiratory Chain Complex I (NADH: ubiquinone oxidoreductase or NADH dehydrogenase).

Table 7: Differentially expressed proteins in COADREAD, associated with "respiratory chain complex", divided by regulation and differential expression at transcriptome level

Category	Proteins
down-regulated & diff. exp. at transcriptome level	18
down-regulated & not diff. exp. at transcriptome level	7
up-regulated & diff. exp. at transcriptome level	1
up-regulated & not diff. exp. at transcriptome level	1

Intriguingly, there is not much accumulated scientific evidence, supporting our findings of strong down-regulation of proteins and genes involved in respiratory chain complexes. On the contrary, one study into expression of mitochondrial genes, including ones encoding complexes I and III of the electron transport chain, states that expression of the NADH dehydrogenase 1 (ND1) and NADH dehydrogenase 6 (ND6) encoding genes is significantly elevated in cancer compared to normal tissue and even higher in late-stage cancer compared to early stage [53]. A study aiming singularly at mitochondrially encoded NADH dehydrogenase 2 (ND2) expression analysis also provides evidence for significantly higher expression of the protein in cancer, with a steady increase in disease stages I to IV [54].

Nevertheless, we have been able to identify the respiratory chain as a cellular component highly enriched in differential expression at the protein level. Further investigation into the function of the proteins associated to it may provide insight on the reason for the discrepancy with existing results.

3.6.3 Molecular functions

3.6.3.1 Catalytic activity

We identified “catalytic activity” as the “molecular function” GO term with highest uniqueness (0.915) in our colorectal cancer dataset (Supplementary Table S10). This term refers to all enzymes, being wholly or largely protein or in some cases RNA, which catalyze a cellular biochemical reaction. Because of the broadness of the term, it was found to be enriched by 569 out of the 1119 (50.8%) differentially expressed proteins in the dataset. The ratio of down-regulated and up-regulated proteins associated to the term is relatively equal (277 to 292) with the same amount of proteins being down- or up-regulated and also differentially expressed at the transcriptome level (227) (Table 8). Mostly, the differentially expressed proteins regulated in both directions have relatively weak reflection in differential expression at the gene level (18.0% for down-regulated and 22.2% for up-regulated), indicating that this term is mostly enriched at the protein level, further supporting the hypothesis that changes in protein level are not necessarily reflected by mRNA expression changes, especially when considering enzymes.

Table 8: Differentially expressed proteins in COADREAD, involved in "catalytic activity", divided by regulation and differential expression at transcriptome level

Category	Proteins
down-regulated & diff. exp. at transcriptome level	227
down-regulated & not diff. exp. at transcriptome level	50
up-regulated & diff. exp. at transcriptome level	227
up-regulated & not diff. exp. at transcriptome level	65

3.6.3.2 Oxidoreductase activity

Oxidoreductase activity (GO:0016491) generally refers to enzymes that catalyze a redox reaction. This term is closely connected to the cellular component term “respiratory chain complex” discussed in 3.6.2.3, as all the reactions in the electron transport chain are redox reactions. Moreover, all of the proteins associated to “respiratory chain complex” are also found to be connected to this term. Of the 132 proteins assigned to “oxidoreductase activity”, the majority have shown down-regulation (64.8%) which is partially supported by down-regulation at the transcriptome level (23.5%) (Table 9).

Table 9: Differentially expressed proteins in COADREAD, associated with "oxidoreductase activity", divided by regulation and differential expression at transcriptome level

Category	Proteins
down-regulated & diff. exp. at transcriptome level	65
down-regulated & not diff. exp. at transcriptome level	20
up-regulated & diff. exp. at transcriptome level	32
up-regulated & not diff. exp. at transcriptome level	15

Some differential gene expression studies targeting groups or families of oxidoreductases, however, indicate up-regulation of enzyme-coding genes associated with the term [55], which again is in contradiction to our findings. Nevertheless, these studies were mostly carried out based only on gene level expression, which might be a possible reason for this inconsistency. On the other hand, a study focusing solely on xanthine oxidoreductase (XOR) has concluded, that XOR expression is lower in more than 60% of the tumor samples, compared to normal tissue [56]. A possible explanation for the conflicting results from gene and protein differential expression may be mRNA-protein differential expression discordance regarding oxidoreductase activity.

3.6.3.3 Ion binding

When a protein interacts with ions, charged atoms or groups of charged atoms selectively, it is generally indicative of the term ion binding (GO:0043167). This term is characterized by high frequency in the human Gene Ontology (35.7%) and high uniqueness (0.91) in the dataset. Furthermore, it was found to be supported by 288 differentially expressed proteins, with the majority of them being up-regulated (153) and differentially expressed at the gene level (226) (Table 10).

Table 10: Differentially expressed proteins in COADREAD, associated with "ion binding", divided by regulation and differential expression at transcriptome level

Category	Proteins
down-regulated & diff. exp. at transcriptome level	109
down-regulated & not diff. exp. at transcriptome level	26
up-regulated & diff. exp. at transcriptome level	117
up-regulated & not diff. exp. at transcriptome level	36

While this is a common term, as many proteins are known to possess ion-binding domains, a lot of large-scale proteomic studies have been able to associate it with cancer. Although Bizama et al. reported a strict low-abundance, down-regulation tendency in regard to gastric cancer [57], most available evidence points towards overexpression of proteins associated with the term in cancer [58][59][60], which is in compliance with our findings.

3.7 Variability of Protein Complexes from the CORUM Database

After extracting all protein complexes with significant variability of 0.2 or higher for each cancer type separately, they were ordered by their ratio of differential expression, highest to lowest. Then we looked at the complexes emerging at the top of the lists and identified those of significance for each cancer type and then discussed them in the section below.

3.7.1 Variable Protein Complexes in COADREAD

We were able to identify 83 variable complexes from the colorectal dataset in total (Supplementary Table S11). While the average variability ratio among those turned out to be relatively high (~43%) and a high homogeneity of direction of regulation among the members of same complexes was observed, we turned our attention towards the protein complexes with the highest proportion of differentially expressed members. We observed that a lot of the complexes at the top of the list are, to no surprise, actually subunits of the Respiratory Chain Complex I (NADH-Q oxidoreductase), which is in compliance with our findings described in 3.6.2.3 and 3.6.3.2. More accurately, one of the two protein complexes with 100% differential expression of quantified members in the list is exactly the "incomplete intermediate subunit" of Respiratory Chain Complex I. Because of the multiplicity of occurrences of this complex in the results and the close connection and sometimes equivalency of its subunits, in the following part we will be addressing the Respiratory Chain Complex I as a whole, rather than making distinctive remarks on separate subunits.

3.7.1.1 Respiratory Chain Complex I i.e. NADH-ubiquinone oxidoreductase

NADH-ubiquinone oxidoreductase is the first and largest protein complex in the electron transport chain (or respiratory chain complexes), found at the inner mitochondrial membrane. As described in 3.6.2.3, we were able to identify the "respiratory chain complex" cellular component GO term as significantly enriched in the COADRED protein differential expression dataset. In the following, however, we are focusing specifically on the implications of differential expression of all members of Complex I of the chain.

Complex I is known to have a total of 45 members, usually divided into a conserved "core" group of 14 subunits and a further 31 "supernumerary" subunits situated around the core [61][62]. Because of the L-shape of the Complex I assembly, the core subunits can furthermore be divided into "peripheral arm" group, encoded into the nuclear genome (NDUFS1, NDUFV1, NDUFV2, NDUFS2, NDUFS3, NDUFS7, NDUFS8) and "membrane arm" group, which is product of mitochondrial DNA (ND1, ND2, ND3, ND4, ND4L, ND5, ND6) [61]. While mutations in genes encoding the core subunits of Complex I are reported to be factoring in multiple disease conditions, including diabetes and cancer, reports regarding the latter are rather inconsistent. Even though research efforts indicate Complex I subunits may be regarded as tumor suppressors [63], there is also evidence for mutations of certain genes coding for core Complex I subunits, that promote colorectal [64][65] and even breast cancer [65][66]. It is noteworthy, however, that these studies are referring to mitochondrial genes, coding for the membrane arm of the complex. Unfortunately, the corresponding protein products were not measured

and present in the colorectal protein expression dataset in our study. Thus, we are not able to provide findings in regard to core subunits, part of the membrane arm of the complex.

The remaining seven core proteins, making up the peripheral arm of the complex, are observed to undergo differential expression in colorectal cancer, with the exception of NDUFS3, which barely fails to satisfy the requirements to be considered differentially expressed (log2-fold-change = -0.834 and p-value = 0.00164). Furthermore, all six differentially expressed proteins indicate underexpression in tumor tissue. A comprehensive study into the function of NADH-ubiquinone oxidoreductase in mitochondria-defective cancer cells states its importance for induction of aerobic glycolysis, also known as Warburg effect, which is the predominant energy production mechanism of cancer cells, and thus its importance for sustaining tumor growth [67]. It is also believed that the complex is needed in its intact form in order to carry out its crucial function in tumor cells, which is clearly indicating the opposite of our results.

3.7.1.2 APOL1 Complex B (APOL1, APOA1, HPR, FN1, IGHM)

The other protein complex with 100% differential expression of quantified members in COADREAD is the APOL1 Complex B. Apolipoprotein L1 (APOL1) is primarily known as a minor constituent of High-density lipoproteins (HDL) [68]. In an epidemiological research by Weckerle et al. [69], aiming at characterizing the circulating APOL1 complexes in African Americans, the serum protein was bound to two different groups of constituents, which were named APOL1 Complex A, composed of APOL1, APOA1, HPR and complement C3, and APOL1 Complex B, comprising APOL1, APOA1, HPR, FN1 and IGHM.

The quantified and simultaneously differentially expressed members in our study are namely APOA1, HPR and FN1. Apolipoprotein A-1 (APOA1) is the major component of HDL in plasma. Intriguingly, APOA1 overexpression has been identified as biomarker for several cancer types [70][71]. On the other hand, an experimental study used ginsenoside Rp1 (G-Rp1) for APOA1 up-regulation and reported proliferation inhibition and enhanced apoptosis of colon cancer cell lines [72], suggesting higher APOA1 levels act as tumour suppressor, which complements the findings of our study, where APOA1 manifests down-regulation in cancer. The discrepancy indicates that APOA1 may not act as the same biomarker for different types of carcinoma. Haptoglobin-related protein (HPR) is a serine protease homolog, associated with APOL1-containing HDL. It is known to bind hemoglobin and is believed to take part in the clearance of cell-free hemoglobin and thus allow heme iron recycling [73]. HPR has also been suggested as biomarker for several cancer types, such as breast cancer [74] and malignant lymphoma [75], however, there is no established connection between HPR overexpression, observed in our study, and colorectal cancer. Fibronectin (FN1) is a glycoprotein mainly found at the extracellular matrix that is observed to usually bind to transmembrane integrins [76]. In our findings Fibronectin shows overexpression. FN1 has already been implicated in cancer progression. More specifically, high expression it has been found to stimulate lung carcinoma cell growth by activating the Akt/mammalian target of rapamycin/S6 kinase pathway and inactivating LKB1/AMP-activated protein kinase signaling pathway [77]. On another note, Pujuguet et al. observed, that two different isoforms of the protein, namely ED-A+ and ED-B+, characteristic of cellular FN1, rather than plasmatic one, undergo overexpression in human

colorectal carcinomas [78]. In our contribution, although we are not provided with isoform information, we suggest that up-regulated plasmatic FN1 expression may also be a biomarker for colorectal adenocarcinoma.

In the context of variability of APOL1 complex B, the study for which it was assembled states that the compositions and possibly the conformation of APOL1 complex A, as well as APOL1 complex B vary between different renal-risk variant genotypes. Our findings indicate that the same observation about APOL1 complex B might hold for healthy and tumor tissue as well.

3.7.2 Variable Protein Complexes in BRCA

Using the breast cancer protein expression dataset, we were able to identify only three protein complexes with more than 50% quantified members, from which at least 1/5 also differentially expressed. All three complexes consist of five members, from which three or four quantified. The average differential expression ratio of those quantified complexes amounts to 38.8% (Supplementary Table S12).

3.7.2.1 PCNA-MutS-alpha-MutL-alpha-DNA Complex

As detected by Hidaka et al. [79] in N-methyl-N-nitrosourea treated cells, this complex consists of MutS-alpha (MSH2 and MSH6), MutL-alpha (MLH1 and PMS2) and Proliferating Cell Nuclear Antigen (PCNA) on damaged DNA. In their study the inhibition of PCNA activity was linked to reduced PCNA-MutS-alpha-MutL-alpha-DNA complex levels as well as reduced caspase-3 activity, which is a distinctive feature of apoptosis induction. These measurements indicate that the activity of this PCNA-bound complex is associated to the induction of cell apoptosis.

The BRCA protein expression dataset contained measurements of the PCNA, MSH2 and MSH6 subunits, from which PCNA and MSH2 were found to be differentially expressed and down-regulated. MSH2 is a DNA mismatch repair protein which together with MSH6 can form the MutS-alpha heterodimer mismatch repair complex [96]. It is known that restrictions in expression of DNA repair genes and proteins are common to various cancer types as DNA damage is considered the primary cause for cancer [80]. Furthermore, there is sufficient evidence, linking exactly MSH2 to breast cancer. Westenend et al. reported loss of heterozygosity of the MSH2 gene with a missing wild-type MSH2 allele in a 49-year-old breast cancer patient [81], which strongly implied the involvement of MSH2 in this tumor development. On another note, a study comparing the risk for developing cancer as consequence of mutation in various DNA-mismatch-repair observed that a MSH2 mutation actually presents a higher risk of cancer when compared to MLH1 mutation [82], further validating the importance of MSH2 in cancer development.

In conclusion, we were able to identify the connection between the PCNA-MutS-alpha-MutL-alpha-DNA complex and breast cancer in our dataset and consolidate the importance of DNA-repair mechanism malfunctions for the development of tumor tissue.

3.7.2.2 MSH2/6-BLM-p53-RAD51 Complex

The initial complex in hand consists of BLM, an ATP-dependent DNA helicase known to suppress inappropriate homologous recombination ; p53, a tumor suppressor protein, which blocks proliferation of cells with misrepaired DNA, and RAD51- an enzyme that assists in repair of DNA double strand breaks. As inferred by its constituents, this complex is a part of the DNA mismatch repair pathway. However, Yang, Qin et al. observed that the MutS-alpha heterodimer forms a complex with BLM–p53–RAD51 as a result of damaged DNA forks during double-stranded break repair [83]. By binding to the initial complex, MutS-alpha appeared to enhance BLM's activity, which is mediating branch migration of Holliday junctions and thus suppressing hyper-recombination [84]. An increase in activity by a factor of four, as reported, is considerably noteworthy, which would imply that a MSH2 deficit, as observed in our breast cancer dataset, would hinder forming of the MSH2/6-BLM-p53-RAD51 complex and in turn amplify hyper-recombination- a property frequently attributed to cancer [85].

We once again report a fault in a DNA-repairing protein complex with imputations matching the cancer condition. Further analysis into this complex' activity, however, would reveal its significance in both formations regarding cancer.

3.7.2.3 MAP2K1-BRAF-RAF1-YWHAE-KSR1 Complex

From this signaling complex (GO:0023052) we were able to identify RAF1, BRAF, YWHAE and MAP2K1 as quantified in the BRCA protein expression dataset and RAF1 as differentially expressed and up-regulated. RAF1 is a RAF proto-oncogene serine/threonine-protein kinase with well-documented constitutive overexpression patterns in breast cancer [86][87][88].As observed in the subunit interaction network of the complex on the CORUM website, RAF1 is the only constituent of the complex, which interacts directly with all other subunits, underlining its importance for the complex.

While the significance of this complex for breast cancer development remains unknown, one can speculate that a complex, part of a signal transduction mechanism, with overrepresented proto-oncogene kinase subunit may be of significant importance for gaining insight regarding cancer-specific signaling pathways.

4 Conclusion

In this study we were able to identify multiple genes and their protein products that undergo differential expression with implications in either Breast Invasive Carcinoma or Colorectal Adenocarcinoma. Furthermore, we described cellular components and molecular functions enriched in differential expression in those cancer types. We also analyzed variable protein complexes with under- or overrepresented subunits extracted from our datasets.

We identified 3 547 genes and 42 proteins as differentially expressed and 3 protein complexes as variable in breast cancer, owing to the size of the RPPA protein expression dataset, as well as 12 616 genes, 1119 proteins and 83 variable protein complexes in the colorectal dataset. Five (5) of the 42 differentially expressed proteins (11.9%) were also contained in the BRCA mRNA expression dataset with their corresponding gene expression, and 883 of the 1119 (78.9%) in the COADREAD dataset. Because each protein expression dataset uses different type of expression measurements, also non-identical to mRNA expression values, we could not calculate or visualize the correlation between expressions of gene-protein pairs.

In the BRCA dataset we suggested differentially expressed proteins associated with phosphorylation as biomarker for breast cancer. Furthermore, we observed a connection between the cellular glucose homeostasis pathway and differential expression between normal and breast tumor tissue. We underlined the importance of the MYC gene and the protein it encodes for breast cancer as well as observed down-regulation at both mRNA and protein levels, which contradicts what most research into MYC expression in cancer has previously reported.

On another note, we highlighted the xenobiotic stimulus response pathway as significantly enriched in underexpressed proteins in COADREAD, thus inferring its plausible importance for colorectal tumor suppression. Additionally, the nucleobase, nucleoside and nucleotide metabolism pathway also emerged as enriched in differential expression. Extracellular space, the vacuolar lumen and the electron transport chain were the cellular components of significance in regard to differential expression in COADREAD. Lastly, catalytic and oxidoreductase activity as well as ion binding were the molecular functions predominantly affected by differential expression, with oxidoreductase activity being downregulated, while the other two terms were associated with nearly identical number of under- and overexpressed proteins. From the COADREAD dataset we detected variability in the Respiratory Chain Complex I with strong underexpression in the peripheral arm of the complex as well as in APOL1 Complex B, where we were able to establish a previously undocumented relation between HPR protein overexpression and colorectal cancer. In BRCA we managed to identify three small protein variable complexes, namely PCNA-MutS-alpha-MutL-alpha-DNA complex and MSH2/6-BLM-p53-RAD51, both DNA-repair mechanisms with down-regulated subunits, indicative of tumor tissue condition, as well as the MAP2K1-BRAF-RAF1-YWHAE-KSR1 complex, part of a signaling pathway with overexpressed oncogene product RAF1.

In this project we have been able to show the importance of mRNA-protein differential expression correspondence, as well as the discrepancy of inferring one from the other. Furthermore, we identified for each cancer type specific pathways, cellular components and molecular functions enriched in differential expression. We additionally expanded our research to the protein complex level and highlighted variable protein complexes, which might gain valuable insight regarding cancer-specific complex alterations. The methodology described here also has the potential to be applied for various other cancer types, taken there are paired patient mRNA and protein expression datasets available for those.

5 References

[1] de Sousa Abreu, Raquel, et al. "Global signatures of protein and mRNA expression levels." Molecular BioSystems 5.12 (2009): 1512-1526.

[2] Östlund, Gabriel, and Erik LL Sonnhammer. “Quality criteria for finding genes with high mRNA–protein expression correlation and coexpression correlation.” Gene 497.2 (2012): 228-236.

[3] Koussounadis, Antonis, et al. "Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system." Scientific reports 5 (2015): 10775.

[4] Zhang, Bing, et al. “Proteogenomic characterization of human colon and rectal cancer.” Nature 513.7518 (2014): 382.

[5] Kosti, Idit, et al. “Cross-tissue analysis of gene and protein expression in normal and cancer tissues.” Scientific reports 6 (2016): 24799.

[6] Maier, Tobias, Marc Güell, and Luis Serrano. “Correlation of mRNA and protein in complex biological samples.” FEBS letters 583.24 (2009): 3966-3973.

[7] Kim, Min-Sik, et al. “A draft map of the human proteome.” Nature 509.7502 (2014): 575.

[8] Liang, Peng, and Arthur B. Pardee. “Analysing differential gene expression in cancer.” Nature Reviews Cancer 3.11 (2003): 869.

[9] Gov, Esra, and Kazim Yalcin Arga. “Differential co-expression analysis reveals a novel prognostic gene module in ovarian cancer.” Scientific reports 7.1 (2017): 4996.

[10] Ucal, Yasemin, et al. “Proteomic analysis reveals differential protein expression in variants of papillary thyroid carcinoma.” EuPA Open Proteomics (2017).

[11] Vincent, Krista Marie, Scott D. Findlay, and Lynne Marie Postovit. “Assessing breast cancer cell lines as tumour models by comparison of mRNA expression profiles.” Breast Cancer Research 17.1 (2015): 114.

[12] Garraway, Levi A., Jaap Verweij, and Karla V. Ballman. “Precision oncology: an overview.” J Clin Oncol 31.15 (2013): 1803-1805.

[13] Bode, Ann M., and Zigang Dong. “Precision oncology-the future of personalized cancer medicine?.” (2017): 2.

[14] Jones, Siân, et al. “Personalized genomic analyses for cancer mutation discovery and interpretation.” Science translational medicine 7.283 (2015): 283ra53-283ra53.

[15] Hoshida, Yujin, et al. "Gene expression in fixed tissues and outcome in hepatocellular carcinoma." New England Journal of Medicine 359.19 (2008): 1995-2004.

[16] Reis, Patricia P., et al. "A gene signature in histologically normal surgical margins is predictive of oral carcinoma recurrence." BMC cancer 11.1 (2011): 437.

[17] Weinstein, John N., et al. "The cancer genome atlas pan-cancer analysis project." Nature genetics 45.10 (2013): 1113.

[18] Whiteaker, Jeffrey R., et al. "CPTAC Assay Portal: a repository of targeted proteomic assays." Nature methods 11.7 (2014): 703.

[19] Christian Nordqvist. “Breast cancer: Symptoms, risk factors, and treatment.” Medical News Today (2017) (http://www.medicalnewstoday.com/articles/37136.php)

[20] Gaude, Edoardo, and Christian Frezza. “Tissue-specific and convergent metabolic transformation of cancer correlates with metastatic potential and patient survival.” Nature communications 7 (2016): 13041.

[21] Ellsworth, Rachel E., et al. “Differential gene expression in primary breast tumors associated with lymph node metastasis.” International Journal of Breast Cancer 2011 (2011).

[22] Thongwatchara, Phatcharaporn, et al. “Differential protein expression in primary breast cancer and matched axillary node metastasis.” Oncology reports 26.1 (2011): 185-191.

[23] Zhu, Jing, et al. "Empowering biologists with multi-omics data: colorectal cancer as a paradigm." Bioinformatics 31.9 (2014): 1436-1443.

[24] Ruepp, Andreas, et al. "CORUM: the comprehensive resource of mammalian protein complexes—2009." Nucleic acids research 38.suppl_1 (2009): D497-D501.

[25] Broad GDAC firehose, http://gdac.broadinstitute.org/

[26] Colaprico, Antonio, et al. "TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data." Nucleic acids research 44.8 (2015): e71-e71.

[27] Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." Bioinformatics 26.1 (2010): 139-140.

[28] Ding, Jiarui, et al. “Systematic analysis of somatic mutations impacting gene expression in 12 tumour types.” Nature communications 6 (2015): 8554.

[29] Geistlinger, Ludwig, Gergely Csaba, and Ralf Zimmer. "Bioconductor’s EnrichmentBrowser: seamless navigation through combined results of set-& network-based enrichment analysis." BMC bioinformatics 17.1 (2016): 45.

[30] Ritchie, Matthew E., et al. "limma powers differential expression analyses for RNA-sequencing and microarray studies." Nucleic acids research 43.7 (2015): e47-e47.

[31] Benjamini, Yoav, and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." Journal of the royal statistical society. Series B (Methodological) (1995): 289-300.

[32] Eichner, Johannes, et al. "RPPApipe: A pipeline for the analysis of reverse-phase protein array data." Biosystems 122 (2014): 19-24.

[33] Supek, Fran, et al. "REVIGO summarizes and visualizes long lists of gene ontology terms." PloS one 6.7 (2011): e21800.

[34] Ori, Alessandro, et al. "Spatiotemporal variation of mammalian protein complex stoichiometries." Genome biology 17.1 (2016): 47.

[35] Fanous, Ibrahim, and Patrick Dillon. "Paraneoplastic neurological complications of breast cancer." Experimental hematology & oncology 5.1 (2015): 29.

[36] Giat, Eitan, Michael Ehrenfeld, and Yehuda Shoenfeld. "Cancer and autoimmune diseases." Autoimmunity reviews (2017).

[37] Chen, I-Hsuan, et al. “Phosphoproteins in extracellular vesicles as candidate markers for breast cancer.” Proceedings of the National Academy of Sciences (2017): 201618088.

[38] Aquino, Gabriella, et al. "MYC chromosomal aberration in differential diagnosis between Burkitt and other aggressive lymphomas." Infectious agents and cancer 8.1 (2013): 37.

[39] Xu, Jinhua, Yinghua Chen, and Olufunmilayo I. Olopade. "MYC and breast cancer." Genes & cancer 1.6 (2010): 629-640.

[40] Gearhart, John, Evanthia E. Pashos, and Megana K. Prasad. "Pluripotency redux—advances in stem-cell research." New England Journal of Medicine 357.15 (2007): 1469-1472.

[41] Murphy, Daniel J., et al. "Distinct thresholds govern Myc's biological output in vivo." Cancer cell 14.6 (2008): 447-457.

[42] Dang, Chi V. "MYC on the path to cancer." Cell 149.1 (2012): 22-35.

[43] Fallah, Yassi, et al. "MYC-driven pathways in breast cancer subtypes." Biomolecules 7.3 (2017): 53.

[44] Wu, Jia-Ching, et al. "Polymethoxyflavones prevent benzo [a] pyrene/dextran sodium sulfate-induced colorectal carcinogenesis through modulating xenobiotic metabolism and ameliorate autophagic defect in ICR mice." International journal of cancer 142.8 (2018): 1689-1701.

[45] Beyerle, Jolantha, et al. "Biotransformation of xenobiotics in the human colon and rectum and its association with colorectal cancer." Drug metabolism reviews 47.2 (2015): 199-221.

[46] Kryczek, Ilona, et al. "IL-17+ regulatory T cells in the microenvironments of chronic inflammation and cancer." The Journal of Immunology 186.7 (2011): 4388-4395.

[47] Putila, Joseph J., and Nancy Lan Guo. "Association of arsenic exposure with lung cancer incidence rates in the United States." PloS one 6.10 (2011): e25886.

[48] Zuikov, S. A., et al. "Correlation of nucleotides and carbohydrates metabolism with pro-oxidant and antioxidant systems of erythrocytes depending on age in patients with colorectal cancer." Experimental oncology 36,№ 2 (2014): 117-120.

[49] Ji, Hong, et al. "Proteome profiling of exosomes derived from human primary and metastatic colorectal cancer cells reveal differential expression of key metastatic factors and signal transduction components." Proteomics 13.10-11 (2013): 1672-1686.

[50] Loo, Jia Min, et al. "Extracellular metabolic energetics can promote cancer progression." Cell 160.3 (2015): 393-406.

[51] Mouillé, Béatrice, et al. "Inhibition of human colon carcinoma cell growth by ammonia: a non-cytotoxic process associated with polyamine synthesis reduction." Biochimica et Biophysica Acta (BBA)-General Subjects 1624.1-3 (2003): 88-97.

[52] Jagan, Ishaan, et al. "Rescue of glandular dysmorphogenesis in PTEN-deficient colorectal cancer epithelium by PPARγ-targeted therapy." Oncogene 32.10 (2013): 1305.

[53] Wallace, LaShanale, et al. "Expression of mitochondrial genes MT-ND1, MT-ND6, MT-CYB, MT-COI, MT-ATP6, and 12S/MT-RNR1 in colorectal adenopolyps." Tumor Biology 37.9 (2016): 12465-12475.

[54] Feng, Shi, et al. "Correlation between increased ND2 expression and demethylated displacement loop of mtDNA in colorectal cancer." Molecular medicine reports 6.1 (2012): 125-130.

[55] Kim, Youngho, et al. "Differential expression of the LOX family genes in human colorectal adenocarcinomas." Oncology reports 22.4 (2009): 799-804.

[56] Linder, Nina, et al. "Xanthine oxidoreductase–Clinical significance in colorectal cancer and in vitro expression of the protein in human colon cancer cells." European journal of cancer 45.4 (2009): 648-655.

[57] Bizama, Carolina, et al. "The low-abundance transcriptome reveals novel biomarkers, specific intracellular pathways and targetable genes associated with advanced gastric cancer." International journal of cancer 134.4 (2014): 755-764.

[58] Xiong, Wei, et al. "Microarray analysis of long non-coding RNA expression profile associated with 5-fluorouracil-based chemoradiation resistance in colorectal cancer cells." Asian Pac J Cancer Prev 16.8 (2015): 3395-3402.

[59] Xu, M., et al. "A microRNA expression signature as a predictor of survival for colon adenocarcinoma." Neoplasma 64.1 (2017): 56-64.

[60] Batist, Gerald, et al. "Overexpression of a novel anionic glutathione transferase in multidrug-resistant human breast cancer cells." Journal of Biological Chemistry 261.33 (1986): 15544-15549.

[61] Urra, Félix A., et al. "The mitochondrial complex (I) ty of cancer." Frontiers in oncology 7 (2017): 118.

[62] Letts, James A., and Leonid A. Sazanov. "Gaining mass: the structure of respiratory complex I—from bacterial towards mitochondrial versions." Current opinion in structural biology 33 (2015): 135-145.

[63] Iommarini, Luisa, et al. "Different mtDNA mutations modify tumor progression in dependence of the degree of respiratory complex I impairment." Human molecular genetics 23.6 (2013): 1453-1466.

[64] Akouchekian, Mansoureh, et al. "Analysis of mitochondrial ND1 gene in human colorectal cancer." Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences 16.1 (2011): 50.

[65] Chatterjee, A., E. Mambo, and D. Sidransky. "Mitochondrial DNA mutations in human cancer." Oncogene 25.34 (2006): 4663.

[66] Santidrian, Antonio F., et al. "Mitochondrial complex I activity and NAD+/NADH balance regulate breast cancer progression." The Journal of clinical investigation 123.3 (2013): 1068-1081.

[67] Calabrese, Claudia, et al. "Respiratory complex I is essential to induce a Warburg profile in mitochondria-defective tumor cells." Cancer & metabolism 1.1 (2013): 11.

[68] Cubedo, Judit, et al. "Apo L1 levels in HDL and cardiovascular event presentation in patients with familial hypercholesterolemia." Journal of lipid research (2016): jlr-P061598.

[69] Weckerle, Allison, et al. "Characterization of circulating APOL1 protein complexes in African Americans." Journal of lipid research 57.1 (2016): 120-130.

[70] Li, Hongjie, et al. "Identification of Apo-A1 as a biomarker for early diagnosis of bladder transitional cell carcinoma." Proteome science 9.1 (2011): 21.

[71] Clarke, Charlotte H., et al. "Proteomic biomarkers apolipoprotein A1, truncated transthyretin and connective tissue activating protein III enhance the sensitivity of CA125 for detecting early stage epithelial ovarian cancer." Gynecologic oncology 122.3 (2011): 548-553.

[72] Kim, Mi-Yeon, Byong Chul Yoo, and Jae Youl Cho. "Ginsenoside-Rp1-induced apolipoprotein A-1 expression in the LoVo human colon cancer cell line." Journal of ginseng research 38.4 (2014): 251-255.

[73] Nielsen, Marianne Jensby, et al. "Haptoglobin-related protein is a high-affinity hemoglobin-binding plasma protein." Blood 108.8 (2006): 2846-2849.

[74] Kuhajda, Francis P., Steven Piantadosi, and Gary R. Pasternack. "Haptoglobin-related protein (Hpr) epitopes in breast cancer as a predictor of recurrence of the disease." New England Journal of Medicine 321.10 (1989): 636-641.

[75] Epelbaum, Ron, et al. "Haptoglobin-related protein as a serum marker in malignant lymphoma." Pathology & Oncology Research 4.4 (1998): 271-276.

[76] Pankov, Roumen, and Kenneth M. Yamada. "Fibronectin at a glance." Journal of cell science 115.20 (2002): 3861-3863.

[77] Han, ShouWei, Fadlo R. Khuri, and Jesse Roman. "Fibronectin stimulates non–small cell lung carcinoma cell growth through activation of Akt/mammalian target of rapamycin/S6 kinase and inactivation of LKB1/AMP-activated protein kinase signal pathways." Cancer research 66.1 (2006): 315-323.

[78] Pujuguet, Philippe, et al. "Expression of fibronectin ED-A+ and ED-B+ isoforms by human and experimental colorectal cancer. Contribution of cancer cells and tumor-associated myofibroblasts." The American journal of pathology 148.2 (1996): 579.

[79] Hidaka, Masumi, et al. "PCNA–MutSα-mediated binding of MutLα to replicative DNA with mismatched bases to induce apoptosis in human cells." Nucleic acids research 33.17 (2005): 5703-5712.

[80] Bernstein, Carol, and Harris Bernstein. "Epigenetic Reduction of DNA Repair in Progression to Cancer." Advances in DNA Repair. InTech, 2015.

[81] Westenend, Pieter J., et al. "Breast cancer in an MSH2 gene mutation carrier." Human pathology 36.12 (2005): 1322-1326.

[82] Vasen, H. F. A., et al. "MSH2 mutation carriers are at higher risk of cancer than MLH1 mutation carriers: a study of hereditary nonpolyposis colorectal cancer families." Journal of Clinical Oncology 19.20 (2001): 4074-4080.

[83] Yang, Qin, et al. "The mismatch DNA repair heterodimer, hMSH2/6, regulates BLM helicase." Oncogene 23.21 (2004): 3749.

[84] Karow, Julia K., et al. "The Bloom's syndrome gene product promotes branch migration of holliday junctions." Proceedings of the National Academy of Sciences 97.12 (2000): 6504-6508.

[85] Rothstein, Rodney, and Serge Gangloff. "Hyper-recombination and Bloom's syndrome: microbes again provide clues about cancer." Genome research 5.5 (1995): 421-426.

[86] Davis, Julianne M., et al. "Raf-1 and Bcl-2 induce distinct and common pathways that contribute to breast cancer drug resistance." Clinical Cancer Research 9.3 (2003): 1161-1170.

[87] Callans, Linda S., et al. "Raf-1 protein expression in human breast cancer cells." Annals of surgical oncology 2.1 (1995): 38-42.

[88] El-Ashry, Dorraya, et al. "Constitutive Raf-1 kinase activity in breast cancer cells induces both estrogen-independent growth and apoptosis." Oncogene 15.4 (1997): 423.

[89] Skrzypek, Marek S., et al. "The Candida Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data." Nucleic acids research (2016): gkw924.

[90] Yu, Bing, et al. “Study of the expression and function of ACY1 in patients with colorectal cancer.” Oncology letters 13.4 (2017): 2459-2464.

[91] Dennis, Glynn, et al. "DAVID: database for annotation, visualization, and integrated discovery." Genome biology 4.9 (2003): R60.

[92] UniProt Consortium. "UniProt: a hub for protein information." Nucleic acids research 43.D1 (2014): D204-D212.

[93] Grossman, Robert L., et al. "Toward a shared vision for cancer genomic data." New England Journal of Medicine 375.12 (2016): 1109-1112.

[94] Ashburner, Michael, et al. "Gene Ontology: tool for the unification of biology." Nature genetics 25.1 (2000): 25.

[95] Zhou, Jianbiao, et al. "Tumor-derived exosomes in colorectal cancer progression and their clinical applications." Oncotarget 8.59 (2017): 100781.

[96] P43246 (MSH2_HUMAN). UniProtKB

[97] Gene Ontology Consortium. "Expansion of the Gene Ontology knowledgebase and resources." Nucleic acids research 45.D1 (2016): D331-D338.