CHALMERS
UNIVERSITY OF TECHNOLOGY

Improved in vitro model selection for cancer using molecular data

Master's thesis in Biotechnology

EMMA CHARLOTTE ANDERSSON
MARÍA TRINIDAD MAC-LEAN BALLIVIÁN

DEPARTMENT OF BIOLOGY AND BIOLOGICAL ENGINEERING
DIVISION OF SYSTEMS AND SYNTHETIC BIOLOGY

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2020
www.chalmers.se

MASTER’S THESIS 2020

Improved in vitro model selection for cancer using molecular data

EMMA CHARLOTTE ANDERSSON
MARÍA TRINIDAD MAC-LEAN BALLIVIÁN

Department of Biology and Biological Engineering
Division of Systems and Synthetic Biology
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2020

Improved in vitro model selection for cancer using molecular data

EMMA CHARLOTTE ANDERSSON
MARÍA TRINIDAD MAC-LEAN BALLIVIÁN

Supervisor: Babak Alaei-Mahabadi. Data Sciences & Quantitative Biology, AstraZeneca
Examiner: Aleksej Zelezniak. Department of Biology and Biological Engineering, Systems and Synthetic Biology, Chalmers

Department of Biology and Biological Engineering
Division of Systems and Synthetic Biology
Chalmers University of Technology
SE-412 96 Gothenburg, Sweden
Telephone +46 31 772 1000

Gothenburg, Sweden 2020

Abstract

Cancer cell lines are essential components in the process of cancer drug research due to their function as early stage models for primary tumors. However, selection of the optimal cell line model can be complicated and therefore many drug projects fail because the chosen cell line does not give sufficient results, whilst another model might. Furthermore, many of the candidate cell lines for drug studies are not always good models, only convenient because of their availability and high growth rate. This does, in many cases, result in waste of resources and incapability of drawing good conclusions within a research project. Thus, there is an apparent need for a smarter cell line selection.

In this project, a cell line selection workflow was carried out based on comparison of genomic copy number and transcriptomic data from cell lines and primary tumors. Tumor samples were grouped after tumor subtypes, to make the sub-populations more homogeneous, thereby enabling selection of cell lines resembling only tumor samples similar to the tumor phenotype of interest. The comparison was carried out, first using correlative analysis within copy numbers, then by expression correlation. Thereafter, the search for optimal cell line models was continued by investigating specific traits for the given primary tumor, using gene signature expression.

To demonstrate cell line selection using subtypes, we present an example; the breast cancer tumor subtype Her2 with DNA damage response deficiency. For this subtype, we found 10 cell lines among the top candidates from both the transcriptomic and the genomic correlative analysis, whereafter 5 were selected based the gene signature expression analysis of DNA damage response deficiency. The cell lines found as the best models for breast cancer Her2 tumors of this specific phenotype, were UACC893, KPL1, SW1990, BT474 and HCC1419. Most of these cell lines originate from breast cancer tissue, except for SW1990 which is derived from a pancreatic adenocarcinoma tumor.

Contents

1 Introduction

1.1 The Problem

For a cancer drug to be approved and used as an official treatment, the discovered compound has to be tested in a series of studies, ranging from in vitro to in vivo. Cancer cell lines play an important role in pre-clinical cancer research, as they are used to model tumors as well as to assess the effect of drug compounds [1]. However, there is one problem within this part of the drug development pipeline. Many drug projects fail due to that the cell lines are not good models of the tumor phenotype of interest [2]. In this project, we are attempting to address this issue by using molecular data to select cancer cell lines which are highly representative for a given tumor type.

1.2 Cancer Cell Lines as In Vitro Models

To succeed in the pre-clinical part of cancer research and move on to the next phase, the selection of cell lines is crucial. If the cell line does not translate well enough to the disease phenotype, the cell line may not simulate what actual effect the drug would have on the in vivo tumor cells. Typical cell lines used in pre-clinical research today are used due to their availability and high growth rate, but are not necessarily good models for studying the potential drug. As an example, considering expression and various genomic alterations, S. Domcke et al. (2013) found several rarely used cell lines to be good models of the high grade serous ovarian carcinoma tumor subtype, whilst other ovarian cell lines were found not to be representative of the tumor subtype [3].

Furthermore, one may assume that a good cell line model, highly representative of the tumor phenotype of interest, for a given research project, should have the same origin as the considered tumor type. However, this is not always the case. In a study by K. Yu et al. (2019), the expression profiles of pancreatic tumor samples and cancer cell lines were compared, resulting in many pancreatic cell lines showing a lower correlation to the tumor type than cell lines with other origins [2]. These results demonstrate the need for improvement in the selection of cell lines in order to deselect bad cell line models and focus on the most suitable ones.

1.3 Cancer is a Disease of the Genome

1.3.1 Hallmarks of Cancer

To become a cancer cell, a normal cell has to obtain a number of new characteristics. It is these characteristics that define a cancer cell, a result of many random genomic alterations. This process is comparable to an evolution of a cancer in the body. The risk of a cell to obtain the genomic alterations necessary for becoming a tumor cell is very small. Consequently, genomic instability is required for the progression of a tumor. Below, the synopsis of the hallmarks of cancer, as interpreted from the work of Hanahan and Weinberg (2011)[4], is listed.

Self-Sufficiency in Growth Signals: The number of cells in a normal tissue is carefully controlled by growth signals. The paracrine network consisting of these growth signals is not well known as it is difficult to study the release and destinations of growth signals in a tissue. Whilst the signaling over entire tissues is poorly understood, the pathways involved in the cell cycle of tumor cells are better studied. Cancer cells can gain autonomy of their proliferative process in a number of ways. Some of them involve production of growth signals and their respective receptors, stimulating the surrounding normal cells to send growth factors or increase the amount of cell surface receptors. An important approach to weaken excessive growth response in normal cells is by negative feedback loops. In this sense, cancer cells have progressed to evade safety mechanisms, to allow an increase in the number of cells whilst avoiding the safety protocols installed to prevent such an event.

Evading Growth Suppressors: There are many systems to down regulate growth, which cancer cells must evade. Two good examples of typical tumor suppressor proteins are retinoblastoma associated (RB) and TP53 proteins. These proteins operate in two complementary parts of a cell cycle regulatory system. Dependent on signals from extra- and intracellular sources, RB proteins allow the cell to continue its proliferative cycle or not. TP53 has a similar role, but receives only intracellular signals from abnormality or stress sensors. These signals convey, for example, levels of DNA damage or certain building blocks and metabolites. If the levels are suboptimal, TP53 can put the cell cycle to a halt, only allowing it to continue when conditions are back to normal again. However, if the conditions are too distressing it will activate apoptosis. Thus, it is common for these systems to be deficient in cancer cells.

Resisting Apoptosis: Apoptosis is triggered as a response to numerous stress factors, such as excessive expression of oncogenes and DNA damage, many of which cancer cells experience during their progression. Apoptosis is controlled by two groups of proteins: upstream regulators, receiving and processing extra- and intracellular apoptosis signals, and downstream effectors, that execute the apoptosis program if activated. A common genetic alteration in cancer cells is loss of the TP53 tumor suppressor. Another example to avoid cell death is by increasing the production of antiapoptotic factors.

Limitless Replication: In normal tissues, the number of replications are limited, whereafter the cells enter the nonproliferative state of senescence or a crisis state, which leads to death. In normal cells, the length of the telomeres suggest the number of divisions left for the cell line, since the telomeres are shortened after each division, until, finally, the chromosome ends are left unprotected. As opposed to normal cells, telomerase is highly abundant in cancer cells, adding telomeric repeats to the telomere ends, making them long enough for the cell to avoid entering the crisis state.

Angiogenesis Induction: For a tumor to grow, it needs the support of vasculature, to supply the cells with nutrients and oxygen as well as to discard waste products. In adults, angiogenesis is only activated for certain events, such as healing processes. Therefore, excessive growth is disabled. Nonetheless, most tumors manage to activate angiogenesis during early stages of their progression, enabling tumor growth.

Tissue Invasion and Metastasis: For cancer to spread and establish secondary tumors, the cells need to be capable of invading other tissues and grow there. In this process, adhesion molecules are often involved. Some adhesion molecules help the tissue maintain its shape, thereby aiding in keeping it quiescent. These are down regulated in carcinomas. Other adhesion molecules are up regulated due to their roles in assisting cell migration. Normally these compounds are active during inflammation or embryonal development. The process of colonizing distant tissues includes invasion into the nearby vascular system and the escape of it into another site of the body. Thereafter, the invading cell divides into a small group of cells, growing into a macroscopic tumor.

1.3.2 Oncogenes and Tumor Suppressor Genes

There are two types of genes that are essential to obtain the properties of cancer cells listed above and thus, the progression of a tumor. A functional mutation of any of them is required to generate cancer.

Tumor progression can be initiated in a cell by random chance or be caused by a cancer agent, such as radiation. Tumor evolution is generated by a somatic gene mutation that is believed to be acquired in most human cancers . The gene mutation gives rise to a growing disorder in the cells, making them multiply faster than a normal cell. This happens when proto-oncogenes are activated into oncogenes. Oncogenes promote the cell cycle, thus, when promoted they will accelerate cellular division [5].

Another set of genes key to initiation of cancer are the tumor suppressor genes. Tumor suppressor genes are present in normal cells, where their function is to stop the cell cycle and check that everything is in order before allowing the cell to proceed to the next step. One factor that is examined is the DNA duplication process. If there is DNA damage, the tumor suppressor will force the cell cycle to stop until the damage is repaired. Should it be impossible to repair, the gene activates apoptosis. An alteration of a tumor suppressor gene can make it lose its function, enabling the cell to proceed with the cell cycle regardless of DNA damage. This will accelerate cell division and promote DNA damage, leading to tumor development [5].

1.4 Availability of Data

The transcriptomic and genomic primary tumor data was obtained from The Cancer Genome Atlas (TCGA). It contains the information of more than 20,000 primary cancer samples including 18,000 genes. Furthermore, these primary cancer samples originate from 33 different cancer types [6].

The cell line data, also including transcriptomics and genomics, comes from the Broad Institute Cancer Cell Line Encyclopedia (CCLE). It is generated from studies of around 1,000 human samples with information of approximately 18,000 genes. The cell lines are derived from 36 different cancer types [7].

Because the databases have a large amount of data on both a genomic and a transcriptomic level, it presents an excellent opportunity for a project such as this one. Furthermore, information about the cancer origin and subtypes of the samples is available, which, together with the large amount of samples, enables a thorough comparison of the primary tumors and cell lines.

2 Background

2.1 Cell Lines and Primary Tumors

A primary tumor is the the tumor located at the anatomical site where the tumor progression began. It can grow and expand to other sites of the body, forming secondary tumors. The secondary tumors will thus be of the same type as their primary tumors [8], adapting to the new environment. Primary tumors can be studied in vitro with established cell lines, derived from the primary tumor and further cultured in the lab, or with the same primary tumor cells. As human cancers are heterogeneous, there are different types of tumors, such as breast, ovarian, bladder, among others, that have different biological features. Furthermore, each cancer type can have several subtypes, often differing from each other substantially.

Cell lines are established in vitro cell cultures, used for research such as cancer biology and studies on preclinical drugs. They are originally samples from tumors, which have been cultured in vitro. For any in vitro cancer research, a decision has to be made about which cell line should be used. The optimal scenario would be to use a cell line which is highly representative of the tumor phenotype of interest to the study. However, it can be difficult to determine what cell line are best models for a study. The reason is that what properties are most important depend on the objective of the study at hand.

The optimal case for modeling primary tumors would be if one could simply select a cell line with the same cancer origin as they would present similar characteristics, but the reality is not as straight forward as this. The properties of cell lines will not necessarily behave as the primary tumor with the same cancer origin. There are different causes, such as cross-contamination of cell lines, samples mistakenly taken from a metastatic tissue that leads to mislabeling of the cell line and that adapting to changes in a new in vitro environment could present genetic drifting. All these characteristics can cause the cell lines to differ from the primary tumors [9]. There are even some cases where cell lines from another tumor type origin could be more representative for a given cancer biology analysis [10].

2.2 Data

2.2.1 Transcriptomic Data

The transcriptome is the RNA expressed from the genome, being the first product of genome expression. This expression changes when the cell is under different conditions or developmental stages. Transcriptomics is an essential area in cancer research because mRNA levels differ significantly between cell types, giving an insight into cell characteristics, specific cancer types or even cancer subtypes. Moreover, transcription is linked to protein production, and thus the functional components of the cell [11].

2.2.2 Genomic Data

Normally, every gene has two copies, one on each chromosome. However, in cancer genomes, loss or gain of DNA segments is a common occurrence. This kind of genomic alteration is called copy number alteration (CNA), and can result in both a deletion or amplification of genes [12]. When comparing primary tumors and cancer cell lines, CNA profiles can be a significant factor to consider, as it might not be represented in the expression whilst being of importance to the protein production. This is because the ratio of number of transcripts to number of proteins translated differs between genes. Since there is a high amount of copy number (CN) data available, and this can serve as a good complement to the expression analysis, it can be useful as an additional factor to deselect bad cell line models for a given research project. For some tumor subtypes, this factor can have a greater relevance than for others, giving an indication about if it should be weighted more or less.

2.2.3 aCGH

The copy number data used in this project was estimated by Microarray-based Comparative Genomic Hybridization (aCGH), where an array is used, with a probe of a known sequence in each well. In aCGH, a reference genome is added to allow an estimation of CN in the test genome relative to the test genome. Both genomes are tagged with fluorescent dye, the reference genome in red and the test genome in green, whereafter they are fragmented and allowed to bind to the probes. If the test genome has a higher CN of segment X than the reference, more green-coloured DNA will bind to the probes matching segment X. In this instance, a green light will be detected from the well. In the case of less CN in the test genome, more red-coloured DNA will bind to the corresponding probes and the well will show a red light. Lastly, when there is an equal amount of CN in both genomes, the light detected will be yellow [13].

2.3 Batch Effect

To analyze the data that is present in TCGA and CCLE it is important to know where the RNA-seq data comes from to account for the possible batch effects. For example, data is often generated in different time points, machines, flow-cells and sequencing platforms, among other things. These factors should be corrected for, reducing biases from the results.

2.4 Dimensionality Reduction

Large data sets are difficult to visualize, but there are several tools that can help give a better picture of this kind of data.

A popular method is t-distributed Stochastic Neighbor Embedding (t-SNE), which uses a non-linear dimensionality reduction algorithm. It seeks for a structure in the data set by using a probability distribution, making a "*random walk on neighborhood graph*" [14]. This technique has been applied in different studies for analyzing transcriptomic data as it captures both local and global structure due to its non-linear approach. This enables interpretation of complex polynomials of the characteristics of a data set and a good visualization of the data. [2][15].

Another preferred method to visualize high dimensional data is Principal Component Analysis (PCA). Unlike t-SNE, PCA is a fast linear mathematical technique enabling dimension reduction of a data set [16]. The output of a PCA are principal components, which represent a linear combination of the features of the data, where the first principal component contains the features with highest variance, and so on.

The principal components provide information of the features making the data cluster into different groups. PCA and t-SNE can be used together to improve the approach for data visualization. As an example, after running a PCA one can select the features with higher variance, followed by using the t-SNE method with the selected variance.

2.5 Gene Set Scoring

When working with cancer transcriptomic data, it is possible to score different molecular signatures (gene sets) that can represent specific phenotypes. These gene sets can be obtained from a widely used database called Molecular Signatures Database (MSigDB) [17][18], in order to perform gene set enrichment analysis. The scoring of different gene sets allows a visualization of which groups of genes are up or down regulated within different samples of the data set.

There are different approaches to score a gene set in a sample based on expression level. Some methods give scores independently of other samples, such as *singscore* (single sample scoring) and ssGSEA (single sample gene set enrichment analysis). Others give scores dependently of the other samples, such as GSVA (gene set variation analysis), PLAGE (pathway level analysis of gene expression) and z-scores.

ssGSEA, PLAGE and z-scores are unsupervised single sample enrichment analyses, that score the gene sets and each sample of the data set [19]. The GSVA method accounts for all samples within the data. GSVA is an unsupervised method that uses a non-parametric weighting function called Kernel to estimate a distribution of a data set, for example, the gene expression level across all samples. This leads to a common scale in the expression profile [20]. The advantage of this method is that it accounts for the noise of the data and dimension reduction. However, it is not recommended to use in data sets with less than 10 samples [19].

Foroutan et al. (2018) implemented the *singscore* method [21], and presented it as simple, fast and able to produce stable and reproducible scores. This is a single sample ranked based gene set scoring method which does not depend on the amount and formation of the samples enclosed by the data. The singscore method has been used for transcriptomic data to predict mutations in cancer, detecting unique characteristic patterns [22].

2.6 Purity Correction

In all tumour samples, there is a risk of contamination of non-cancerous cells. The more contamination of normal cells, the less pure the tumor sample will be. This will, of course, alter the data. A purity estimate of a sample is an approximation of how much of that sample has actual tumour origin, based on analysis of the same samples. These estimates are difficult to calculate for cancers present in the blood, and so, there is no purity data for Acute Myeloid Leukemia (LAML) or Diffuse Large B-Cell Lymphoma (DLBC). However, the purity estimates for the remaining tumour types would hypothetically enable a removal of non-tumour related genes from every dataset, making the data more accurate. Theoretically, non-tumour related transcriptome present in a tumour sample would arise as a result of infiltration from the immune system cells. [23], [24] Nonetheless, in this project, there were some doubts as to how beneficial the usage of purity estimates are and how well they represent reality. The questions asked for confirmation of the relevance of a correction for purity were: How accurate are the purity estimates, and are the genes removed as a result of them part of immune system pathways? Accordingly, a correlation value was calculated, using the purity estimates, for every gene over all tumour types, as well as for separate tumours. Thereafter, the pathways associated with the genes, which were negatively correlated to the purity estimates of each tumour type, were studied. If the purity estimates were accurate, one would expect all the negatively correlated genes to be part of the immune system pathways.

2.7 Workflow

In this project, we have analysed and compared cancer cell lines and primary tumors in two layers of data; genomic and transcriptomic. In our comparison we had two approaches, the first one being of a holistic nature and the second a more detailed one, investigating functional relationships. The holistic comparison was carried out to deselect any bad cell line models, by correlative analysis of both expression and gene copy numbers. Using the cell lines with the highest correlation from the transcriptomic and genomic analyses, the functional relationships were examined in terms of gene signature expression, to find the good cell line models. The workflow described is shown in figure 1.

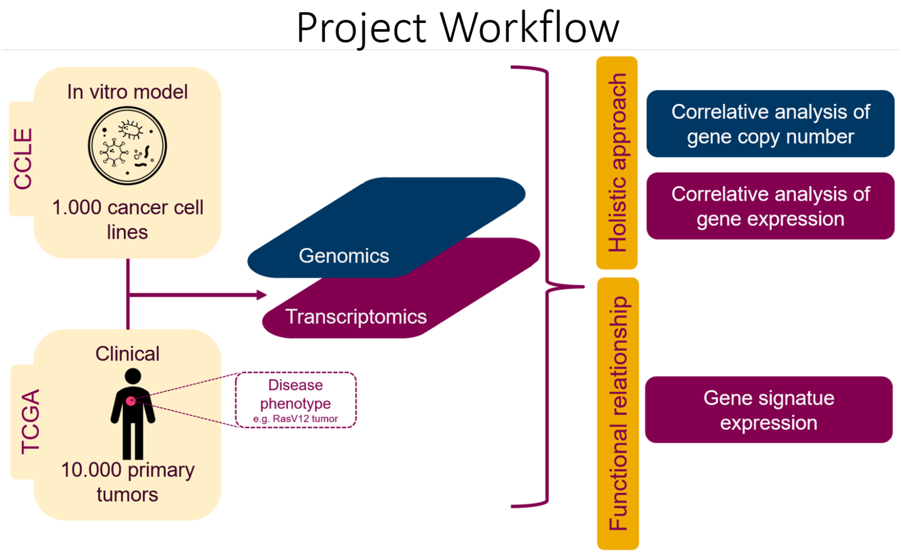


Figure 1: Scheme of project workflow, picturing that the TCGA and CCLE data ere analysed in two layers: genomics and transcriptomics. The analysis is conducted in two parts. First a holistic approach, where only the most highly correlated cell lines are selected, whereafter the functional relationship is investigated in terms of gene signature expression, to select the best cell models.

3 Aim

The aim of this project is to select the most suitable cell line model for a specific cancer subtype. The first part is based on comparison on a genomic and transcriptomic level. This involves analysis of TCGA and CCLE data, primary tumors and cell lines respectively, and evaluation of the differences and similarities within both layers of data. The second part incorporates analysis of specific gene signature expression, more specifically analysis of characteristic gene sets to enable identification of those cell lines that are most representative for a specific disease phenotype.

4 Method

4.1 Data Preparation

4.1.1 Expression Data

The raw expression data can be found in Google Cloud Pilot RNA-Sequencing for CCLE and TCGA, which is available here. This data was quantified using Kallisto, a program presented by N. Bray et al., that is based on the idea of *pseudoalignment*. It is a fast and accurate method to record the amount of RNA-seq without the necessity of alignment [25].

The counts of this data were normalized and the batch effect was corrected for as the data came from different sequencing platforms. Both normalization and batch correction of the TCGA and CCLE data have been solved by K. Yu et al. where the normalization was executed by making use of the upper quartile method. The batch effect correction was done using ComBat. In this study, 22 overlapping cancer types from TCGA and CCLE were used for the TCGA data. The tumor types included bladder, breast, cholangiocarcinoma, colon, lymphoid neoplasm, esophageal, glioblastoma multiforme, head and neck, kidney, leukemia, brain, liver, lung, lung squamous cell, mesothelioma, ovarian, pancreas, prostate, skin, stomach, thyroid and uterine corpus endometrial [2]. For this project, the normalized and corrected data from the research of K. Yu et al was used. It can be found by following this link.

To make TCGA and CCLE more comparable after the normalization and the batch correction, this data was submitted to a correction of data type using ComBat.

In order to investigate more homogeneous subpopulations of the primary tumor data, subtypes within tumor types were investigated. For this, the PanCancerAtlas_subtypes table, containing molecular TCGA subtypes, was used. The table can be found in the Bioconductor package TCGAbiolinks.

4.1.2 Copy Number Data

The copy number data for both CCLE and TCGA was downloaded from the cBioPortal. The data was converted from copy number estimations per DNA segment into copynumber estimation per gene, using the Bioconductor package CNTools. To convert the data into CN by gene a gene map was created to use as reference, by applying another Bioconductor package; biomaRt, according to the biomaRt users guide [26], in order to refer to the positions of the genes of interest [27].

4.2 Dimensionality Reduction

To visualize transcriptomic data two methods were used, t-SNE and PCA. For this, the normalized and batch corrected data of TCGA and CCLE from the paper of K. Yu et al [2] was used. For both methods, the 5.000 most variable genes of TCGA data were selected by applying Interquartile Range (IQR).

In t-SNE, the function Rtsne() was applied to analyze TCGA and CCLE separately and together. In the plots, the data points were colored by their primary tumor origin.

In the PCA method the prcomp() function was used to retrieve the principal components. This was performed for TCGA and CCLE separately as well as together. Furthermore, each loading of the first three principal components (PCs) was analyzed. For every one of the PCs, the 50 loads with the highest value were selected together with the loads with minimum values. For each selection, the associated genes were extracted and the pathways were analyzed in Enrichr [28][29].

4.3 Gene Set Scoring

Different gene sets were analyzed, among them DNA Damage and 9 gene sets from the cancer hallmarks (E2F targets, G2M checkpoint, MYC targets V2, allograft rejection, kras signaling up, oxidative phosphorylation, heme metabolism, UV response up and WNT beta catenin signaling) taken from the Molecular Signatures Database (MSigDB) [18]. These 9 gene sets from the hallmarks of cancer were chosen by finding 3 pathways enriched in CCLE (E2F targets, G2M checkpoint, MYC targets V2), 3 pathways enriched in TCGA (allograft rejection, kras signaling up, oxidative phosphorylation) and 3 pathways with similar enrichment between CCLE and TCGA (heme metabolism, UV response up and WNT beta catenin signaling). These gene sets were based on a study where several hallmark gene sets were analyzed and differentiated by enrichment of TCGA and CCLE [2] allowing a comparison to the results of this project. The gene sets were analyzed with transcriptomic data, using three different methods, GSVA, ssGSEA and singscore. Finally, the most appropriate method was selected.

GSVA and ssGSEA were implemented with the GSVA package within *R/Bioconductor*. To use it, TCGA and CCLE expression data were put together in one matrix and the gene sets were settled as a list. Together with this, the method was specified.

The third gene set scoring method analyzed, was singscore. This method was implemented using the singscore package within *R/Bioconductor*. The function multiScore() [30] was applied, which calculates singscores [21] by using the transcriptomic data of TCGA and CCLE, and the gene sets of cancer hallmarks and DNA damage. To use this method, the expression data of TCGA and CCLE were merged into one matrix. Before using the main function, this matrix was ranked using rankGenes(), whereafter the ranked data and the gene sets were settled with simpleScore().

4.4 Correlation Calculation

The correlation was calculated using the Spearman method, where two different approaches were used for CN and expression analysis respectively. For both calculations, only the top 5,000 differentially expressed genes were selected for the tumor type in question, using the upper quartile method, followed by a selection of the genes present in both primary tumor and cell line data. The correlation values were investigated separately for a given tumor type or subtype.

4.4.1 Two Methods to Calculate Correlation

Two ways of finding the correlation values for each cell line were investigated and compared. **Method 1** involved calculating the correlation between each cell line sample and every tumor sample before using the mean correlation value for each cell line. This method was used for the expression analysis. For the CN analysis **Method 2** was used. Here, a mean CNA profile for all tumor samples was created by calculating the mean CN for every gene across all samples. Thereafter, the correlation was calculated between the mean CNA profile and every cell line.

4.5 Purity Correction

To investigate the usefulness of a correction for purity in the data, correlation values were calculated for each gene, using Spearman's rank test. Given the purity estimates of each sample, correlation values could be calculated for every gene, and the genes giving a negative correlation were selected. The limits for correlation value and adjusted p-value were set to -0.6 and 0.01 respectively. Regarding the gene rho values for all tumor types separately, only the 60 genes with the lowest rho were selected. Two tables of genes were generated. One where the correlation values had been calculated for every tumor type separately, and another where the correlation values had been computed over all the samples, LAML and DBCL samples excluded. The purity estimates, together with the purity correction code, were obtained from the github repository of K. Yu et al, found here. To assess which pathways the genes were present in and their significance, Enrichr was used. The resulting pathways were thus produced using the hypergeometric test by Enrichr, where they were also ordered after significance.

5 Results and Discussion

5.1 Visualization of the Transcriptomic Data

Because the raw expression data is not normalized it is not comparable as it is. Furthermore, there are other variables that have not been considered, such as batch effects due to samples being run on different days and with different equipment. Thus, to have consistency within all data, a few factors have to be corrected for. Yu-Lin K. Chang and his team performed a transcriptomic analysis of cell lines as models of primary tumors, where TCGA and CCLE data were treated with normalization by counts as well as batch correction before the actual analysis [2]. For this project the normalization and batch correction was used according to that study. First, as a sanity check of the data, a t-SNE and PCA plot were created in order to visualize it.

In the t-SNE plot, figure 2, one can see that primary tumor samples, where different colors represent the anatomical origin of the sample, are well clustered. Thus, the samples from TCGA are well annotated.

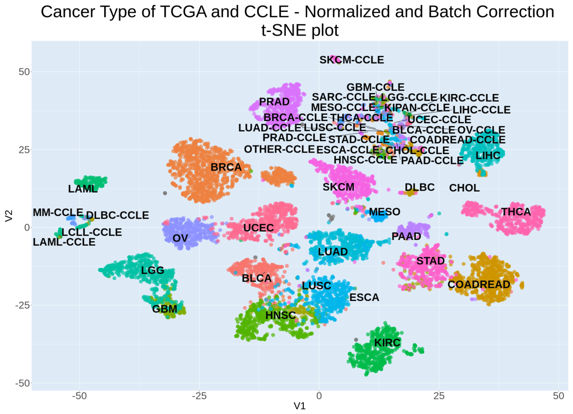


Figure 2: t-SNE visualization of the data with normalization by length and batch correction. Different colors represent a different cancer origin of each sample.

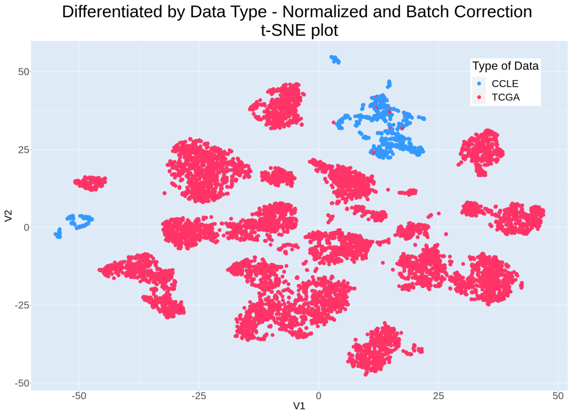


Figure 3: t-SNE visualization of the data with normalization by length and batch correction. The blue dots represent CCLE samples and TCGA samples are colored red.

Figure 3 is the same plot as figure 2 but differentiated by data type. Here, it is possible to see that cell lines, colored in blue, are clustered in two locations outside of the TCGA sample clusters. This shows that there is a large difference between TCGA and CCLE.

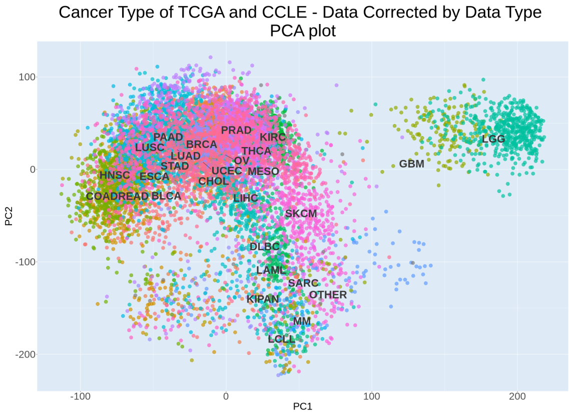


Figure 4: PCA visualization of the data with normalization by length and batch correction. The colors represent cancer origin of each sample.

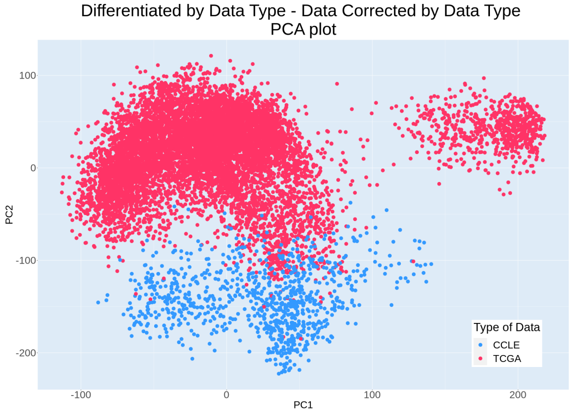


Figure 5: PCA visualization of the data with normalization by length and batch correction. Blue dots represent the CCLE samples and red represent the TCGA samples.

A PCA plot was also created to visualize the expression data. In figure 4 one can observe a significant separation in principal component 1, in the x-axis. The primary tumor samples coming from LGG (Brain Lower Grade Glioma) and GBM (Glioblastoma Multiforme) are clustered to the right whereas the rest are clustered to the left. Both of these types of cancers are coming from the brain. To check the annotation of this separation, the 50 genes with the highest values in principal component 1 were analyzed in Enrichr [28][29], a website platform to annotate functions to selected genes.

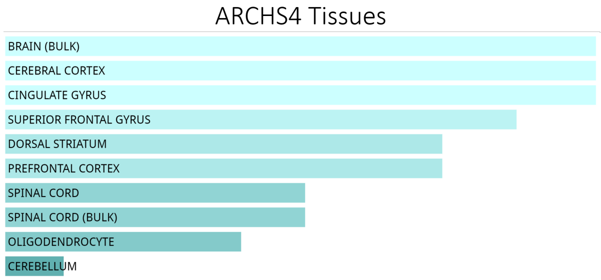


Figure 6: Enrichment analysis of LGG and GBB genes in **ARCHS4 Tissues**

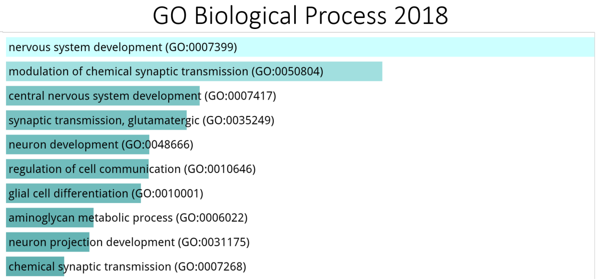


Figure 7: Enrichment analysis of LGG and GBB genes in **GO Biological Process**

In this platform, two features were seen. Figure 6 shows the **ARCHS4 Tissues**, which provides the information of in which tissues the 50 genes are most highly expressed. Among the tissues with higher expression, it is possible to observe brain, cerebral cortex, cingulate gyrus, superios frontal gyrus and dorsal striatum. The origin of all these tissues are part of the brain.

Figure 7 shows the **GO Biological Process 2018** which provides information about in which biological processes the genes are most enriched. Some of the processes are nervous system development, modulation of chemical synaptic transmission and neuron development. These are also brain related.

The results show that the genes that are clustering LGG and GBB primary tumor samples to the right in the principal component 1 (figure 4) are consistent as they are brain associated. This gives a further corroboration of that the data is well annotated. Moreover, it conveys that the expression of these genes are partly responsible for the generation of the variance within PC1.

5.2 Data Correction

Observing figures 3 and 5 above, where the data type differentiation is visualized using t-SNE and PCA, one can presume that there is a strong global difference between TCGA and CCLE data. Motivated by this, a data type correction was made.

After a data type correction of the TCGA and CCLE data, the variance within the data was visualized again. The results are shown by t-SNE (figures 8 and 9) and PCA plots (figures 10 and 11).

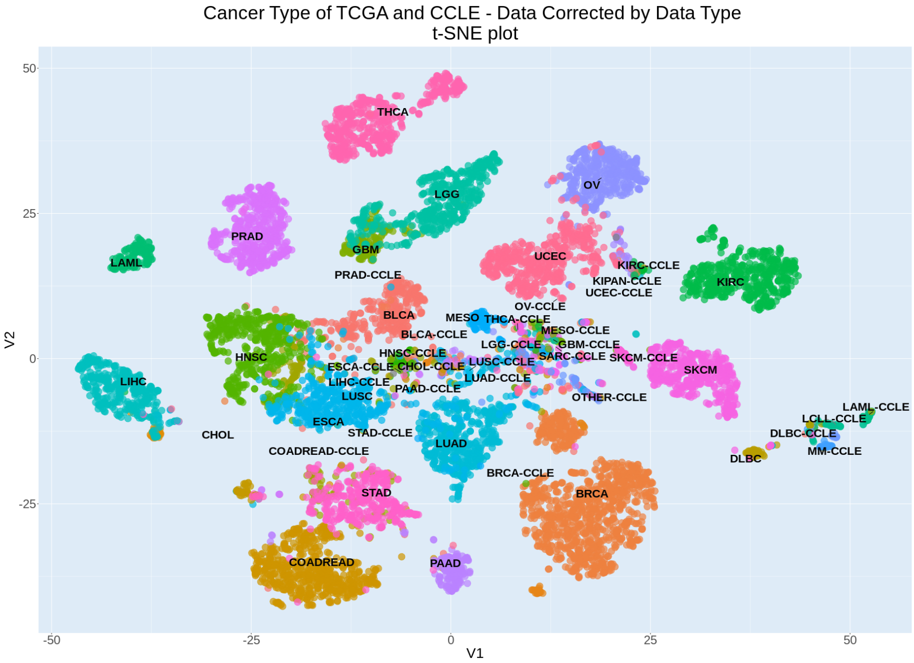


Figure 8: t-SNE visualization of the data after correcting by data type. Different colors represent different cancer origins. The name of each cancer origin is placed by its group mean value in the x- and y-direction on the plot.

In figure 8 the names of the cancer origin of primary tumor samples and cell lines are set at the mean value of all the samples from every group of a TCGA or CCLE cancer type. It is possible to distinguish primary tumors from cell lines because cell lines are accompanied by ”-CCLE” in the name. One can see that primary tumor samples and cell lines from the same origin are closer to each other compared to figure 2.

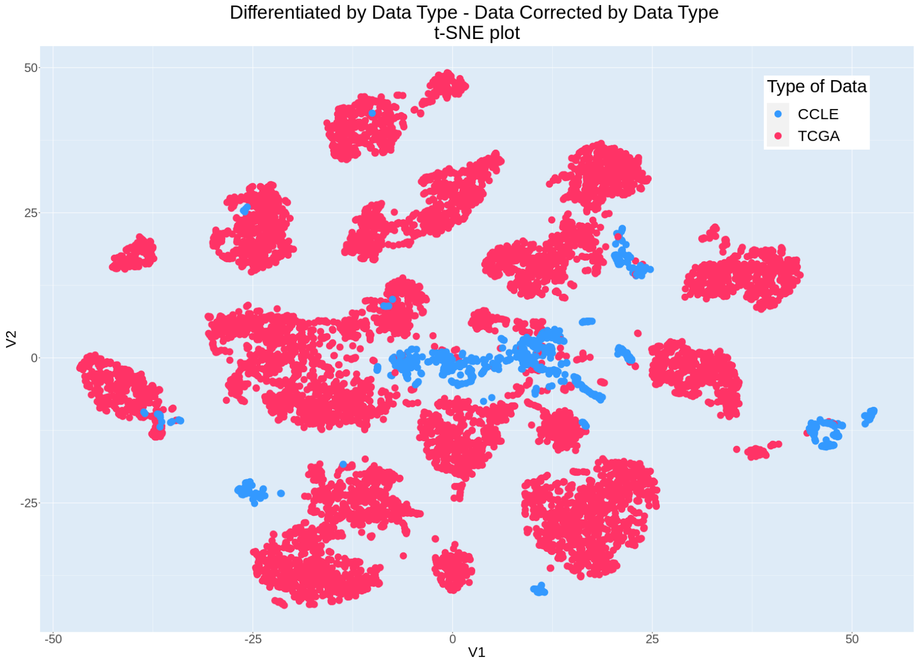


Figure 9: t-SNE visualization of the data after correcting by data type. Different colors represent the data type of each sample (TCGA in red and CCLE in blue).

By creating the same t-SNE plot, but differentiated by type of sample, one can study the difference between primary tumor samples in red (TCGA) and cell lines in blue (CCLE). This is visualized in figure 9. It is possible to see that cell lines are more distributed in the graph after data type correction, when comparing to figure 3. In figure 9 the data is not as separated in groups of TCGA and CCLE data, in other words, the data has improved. Thus, the conclusion is that data type correction should be executed before further analysis.

The data type correction was also visualized using PCA, where the same results are seen; the cell line distribution among the TCGA samples is enhanced compared to before the data type correction, when CCLE and TCGA samples were clearly separated from each other. The PCA plots are presented in figures 10 and 11. Figure 11 shows that there is less difference between cell lines and primary tumors than before the data type correction, visualized in figure 5.

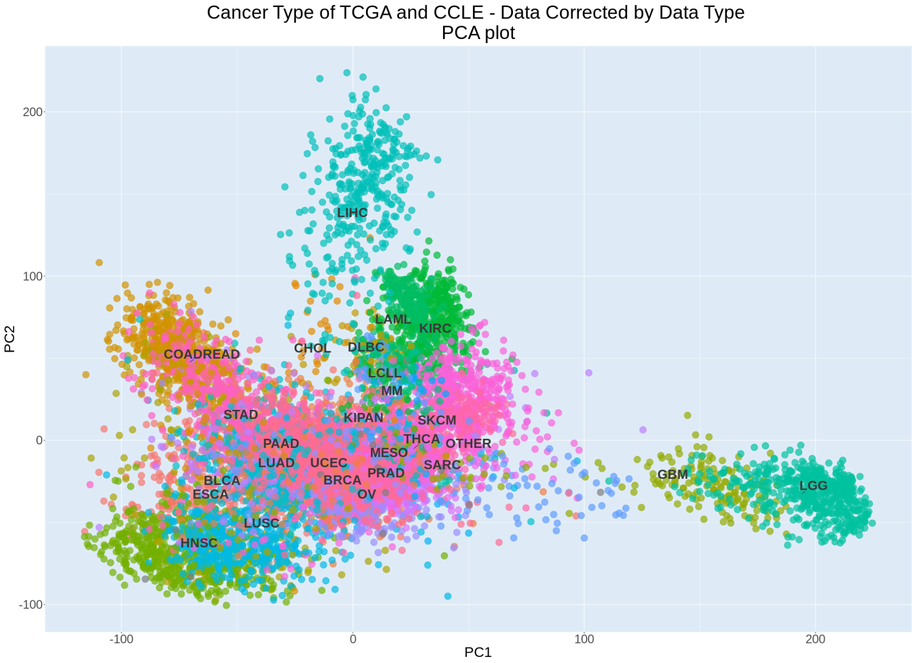


Figure 10: PCA plot of the data after correcting by data type. Different colors represent the cancer origin and the name of each cancer origin is placed by its group mean value in the x- and y-direction.

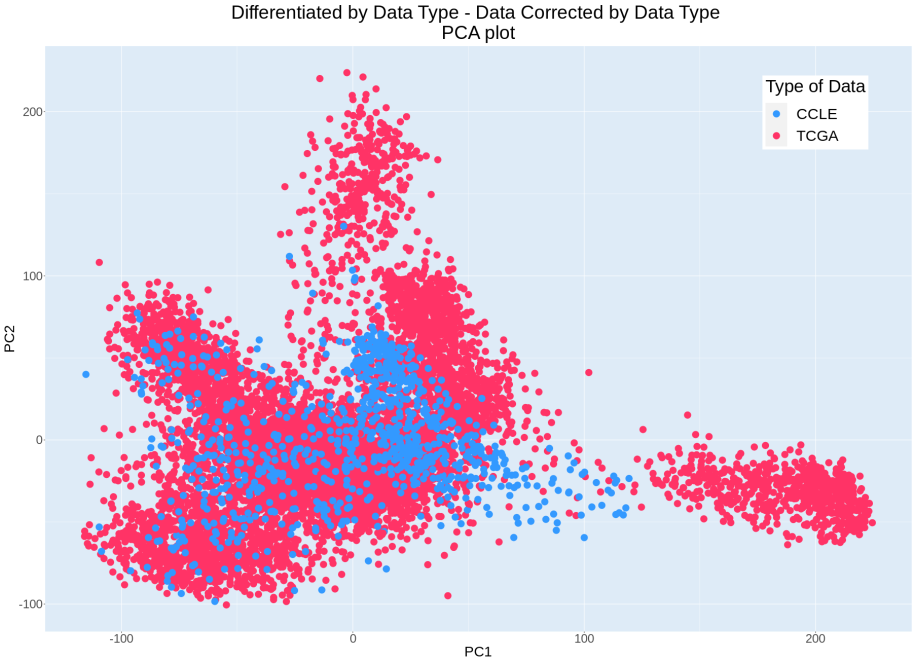


Figure 11: PCA plot of the data after correcting by data type. Different colors represent the type of data of each sample (TCGA or CCLE).

5.3 Breast Cancer Subtypes

The expression of the breast cancer subtypes were also visualized using t-SNE and PCA, and the results are presented below, in figures 12 and 13. Both plots reveal the heterogeneity existing between subtypes, which are represented by different colors in the plots. These plots show that expression within a tumor type can be heterogeneous. Thus, it is reasonable to assemble a more homogeneous population of tumor samples, a subtype, representing the tumor phenotype of interest, before going through with a comparative analysis with cell lines.

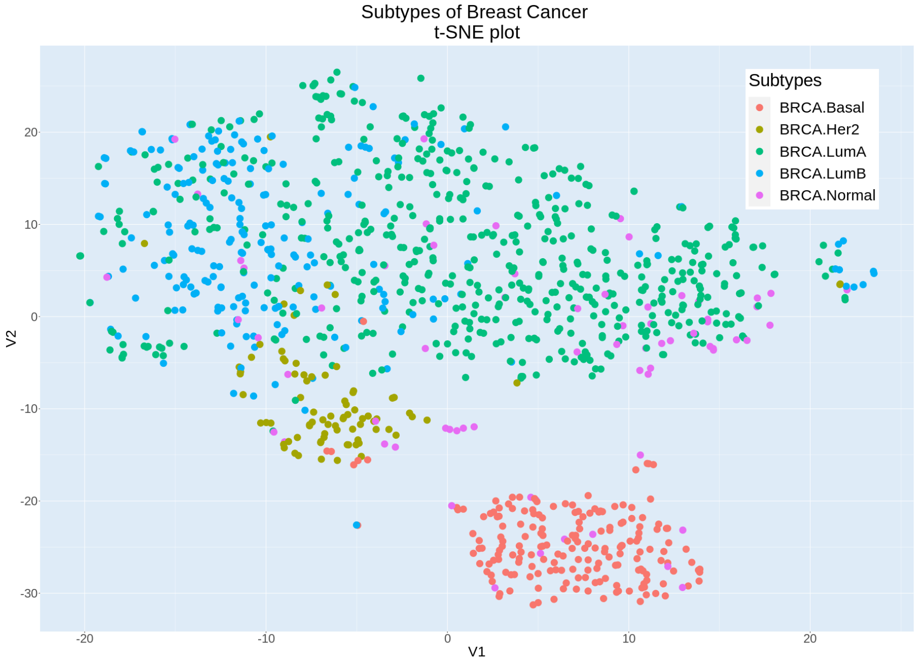


Figure 12: Visualization of the subtypes of breast cancer in a t-SNE plot. Different colors represent different subtypes.

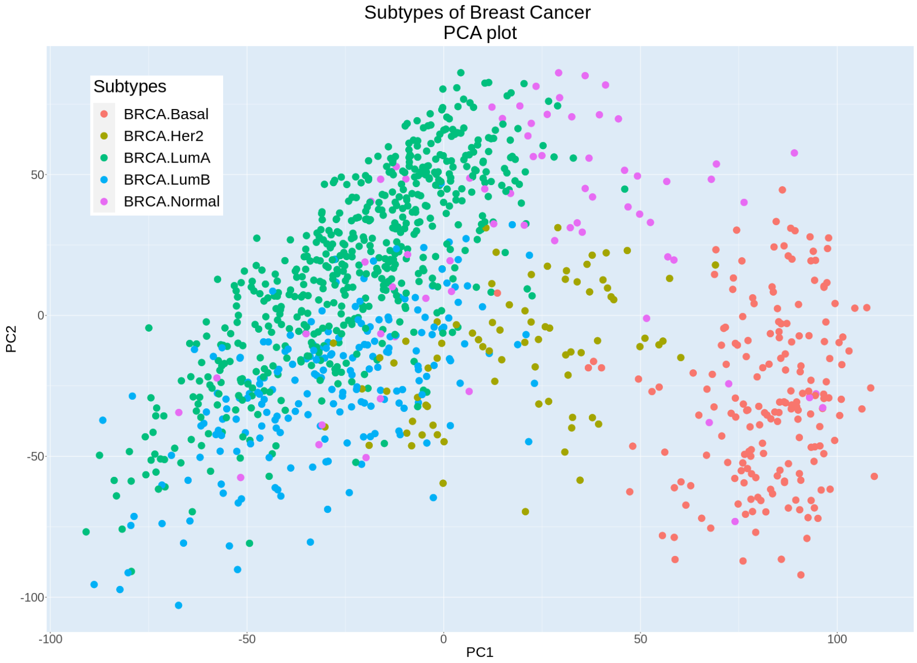


Figure 13: Visualization of the subtypes of breast cancer in a PCA plot. Different colors represent a different subtype

5.4 Correlation Analysis

5.4.1 Method Comparison

To compare the two methods for correlation calculation, the results from each approach were visualized in a boxplot, where the cell lines were grouped after their tumor origin and the two methods were shown in different colors. The corresponding plot for CN correlation methods is presented in figure 14.

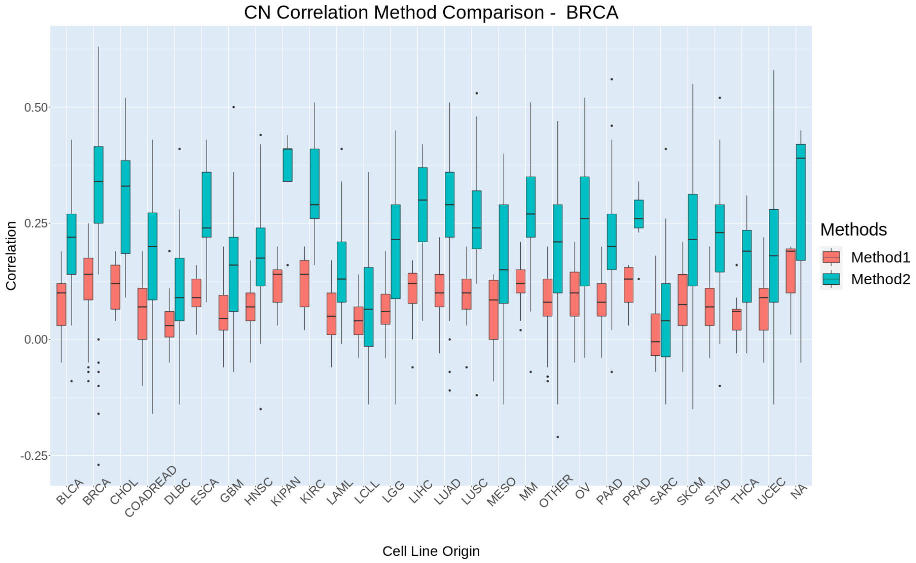


Figure 14: Comparison of correlation calculation methods regarding CN. The blue/green boxes show the results for method 2, where the correlation have been calculated between cell lines and the mean CNA profile of the tumor samples from BRCA. The red boxes show the results from method 1, where the mean correlation for each cell line over all BRCA tumor samples has been used.

The plot shows that when not applying the mean CNA profile approach, the correlation values are all under 0.25, which is too low. To view the selected cell line as a valid model, a higher correlation value is required to select a cell line. The correlation values generated from using the mean CNA profile are higher than when using method 1 and some cell lines obtain a correlation value above 0.5 from method 2. One can also see that both methods are following the same trend between cell line origin, implying that both methods should be accurate.

Method 2 was not used when calculating the expression correlation. Whilst method 1 could have been used in both analyses for consistency, method 1 gave correlation values high enough to be considered valid regarding expression data, as is displayed in figure 15. The figure also shows that both methods have a similar trend. Thus, method 1 was considered acceptable regarding the expression analysis. As the copy number data is more heterogeneous than the transcriptomic data, the usage of method 2 was especially important for the genomic analysis.

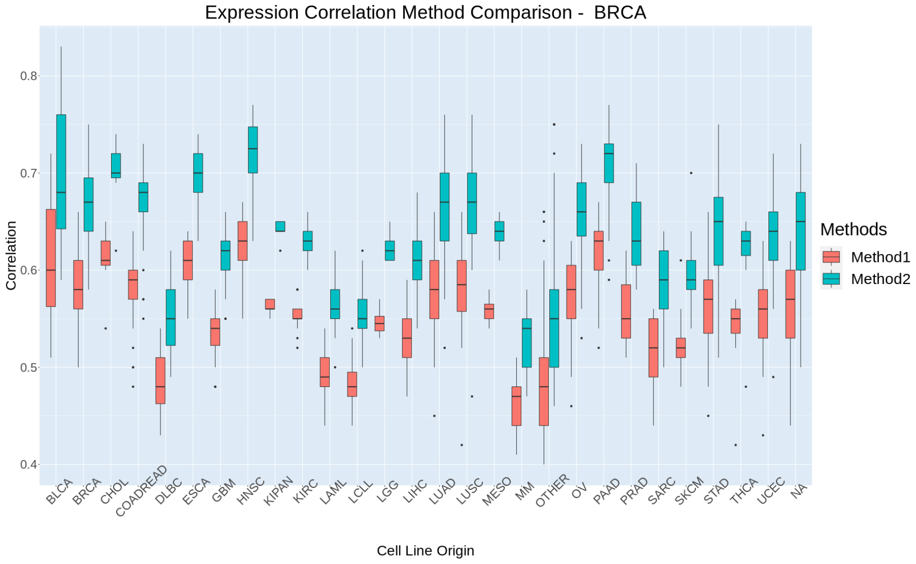


Figure 15: Comparison of correlation calculation methods regarding expression. The blue/green boxes show the results for method 2, where the correlation have been calculated between cell lines and the mean expression profile of the tumor samples from BRCA. The red boxes show the results from method 1, where the mean correlation for each cell line over all BRCA tumor samples has been used.

5.4.2 The Top Correlated Cell Lines and Their Tumor Origin

In order to discard the bad cell line models regarding expression, the cell lines with the highest correlation to each tumor subtype were selected. These cell lines would represent the best models, when only considering expression data. In figure 16 the top 30 correlated cell lines are shown for all BRCA subtypes. In the plot, each cell line is colored depending on its origin.

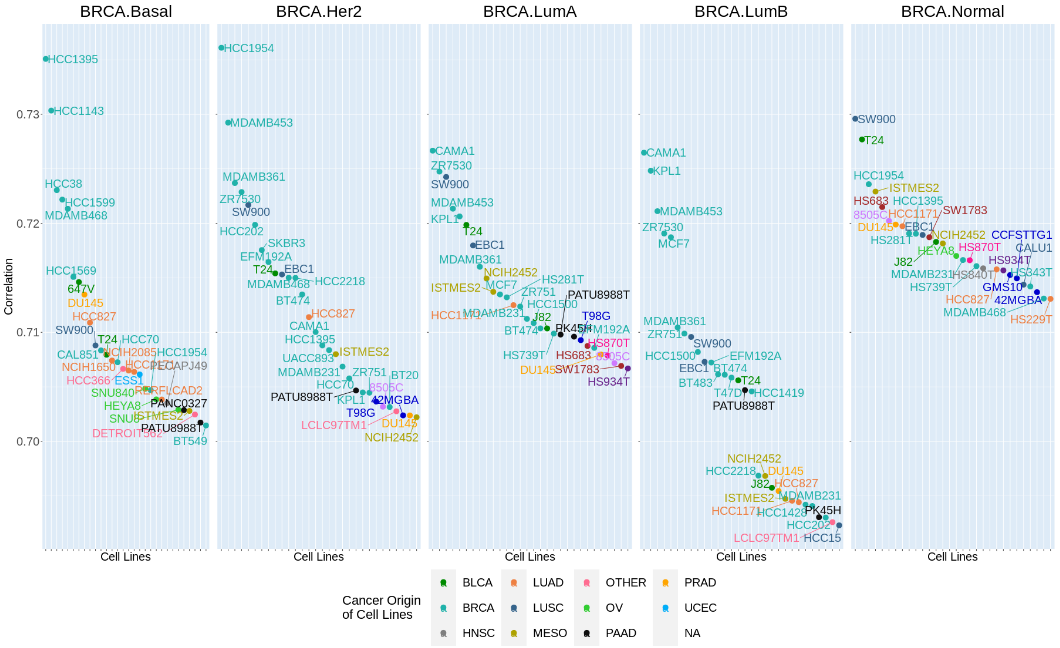


Figure 16: The 30 top correlated cell lines for BRCA primary tumor subtypes, based on expression.

To complement the expression correlation analysis, the CN correlation was calculated using method 2. Below, figure 17 shows the top 30 correlated cell lines for each BRCA subtype, based on CN.

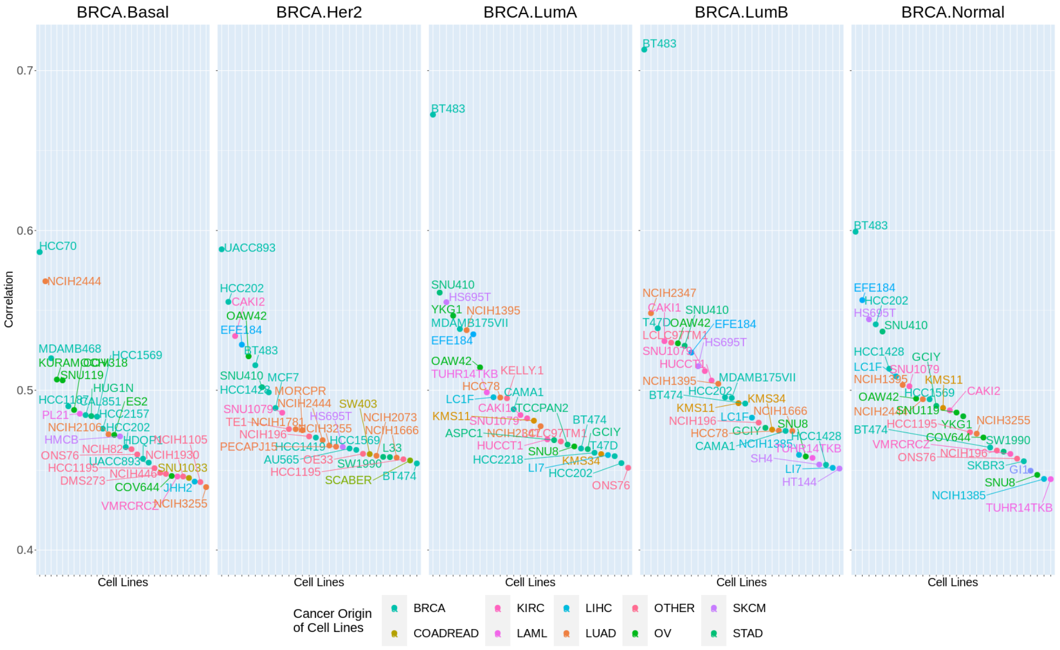


Figure 17: The 30 top correlated cell lines for BRCA primary tumor subtypes, based on CN.

Figures 16 and 17 show that many of the highest correlated cell lines are derived from breast tumors. This is an expected result since the cell lines should be similar to their origin tumor samples. However, one can also see that the results are not the same between subtypes and, indicated by what colors are visualized, the top correlated cell lines are not only derived from breast tumors. There can be many reasons for this outcome, some of which are discussed further below.

Moreover, when comparing figures 16 and 17, it is noticeable that the correlation values are in general higher for expression than for CN. The reason for this is that copy number data is more heterogeneous than expression data. There are many different copy number alterations that can lead to similar results in expression. This is also why the analysis should not be based on only expression or any factor alone, but on as many as can be included. As an example, whilst the expression is related to both genomic alterations and protein production, it can give misleading results for some cases, because it is not necessarily representing the actual situation. There are several different types of genomic alterations that can give the same expression and at the same time, there are many situations where the amount of expression would not correlate with the protein production from the same gene.

The correlation results have shown that no cell line's correlation value is very close to one. This is because cell lines are engineered to grow in the laboratory, thus giving them different properties to those of a tumor. Moreover, cell lines live in a different environment from in vivo tumors. The tumors are required to adjust to the surrounding cells in the body, the immune system and to the signals coming from the body. They need all the different hallmarks of cancer, mentioned in section 1.3.1, whilst cell lines do not. Therefore, a cell line can only be a better or worse model of a primary tumor phenotype than other cell lines, but never correlate completely to the tumor phenotype. Furthermore, cell lines from the same tumor type as the tumor phenotype can be have a very low correlation value, making it highly relevant to find the best cell lines.

5.5 Selection of Good Models

After exploring the holistic relationship between cell lines and primary tumors as well as carrying out a selection of good models from the genomic and transcriptomic levels, 10 cell lines are left. These cell lines are the ones in common between the 50 top correlated ones from the expression and CN correlation analysis.

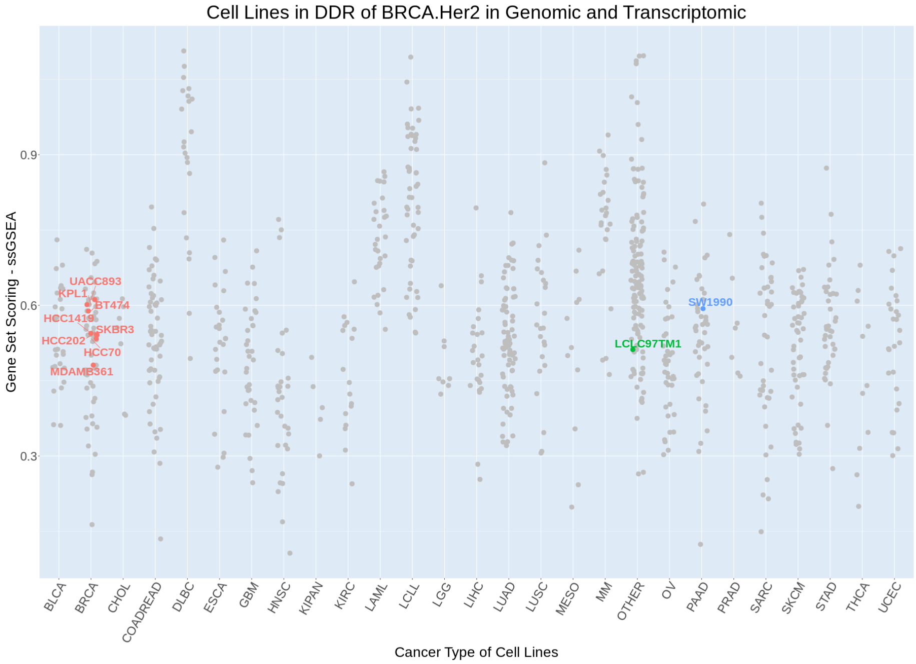
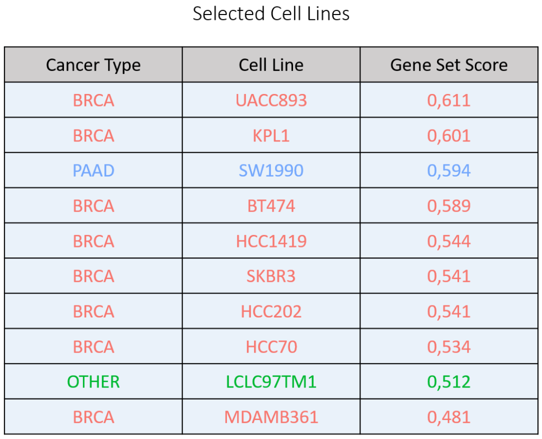


Figure 18: Gene set score of DNA damage response deficiency among all cell lines. The ones in color are the cell lines highly correlated in both the expression and the CN level.

In figure 18 two cell lines from another tumor origin than breast are shown. One of them (LCLC97TM1 in green) is annotated as ”OTHER” and another (SW1990 in blue) as ”PAAD” which is short for pancreatic adenocarcinoma. The cell line annotated as ”OTHER” is actually derived from lung cancer tissue, according to depmap portal. The reasons that one might find cell lines from other cancer types to be highly correlated to the phenotype of interest could be because of miss-classification, metastasis or genomic drifting. Thus, it could be that the cell line is actually from the same tissue origin as the primary tumor subtype, only it was miss-classified or mistakenly assumed to be collected from a primary tumor sample. Another option could be that the cell line actually is derived from another tumor type, and genomic alterations over time have lead it to become more similar to a breast tumor.


Selected Cell Lines

Cancer Type	Cell Line	Gene Set Score
BRCA	UACC893	0,611
BRCA	KPL1	0,601
PAAD	SW1990	0,594
BRCA	BT474	0,589
BRCA	HCC1419	0,544
BRCA	SKBR3	0,541
BRCA	HCC202	0,541
BRCA	HCC70	0,534
OTHER	LCLC97TM1	0,512
BRCA	MDAMB361	0,481

Table 1: The 10 selected top cell lines including information of their cancer type origin, names and gene set score.

The selected cell lines are shown again in table 1, where they are listed by gene set score of the DNA damage response gene set, from highest to lowest. For some tumor subtypes, of course, the functional relationship analysis might generate lower or higher gene set scores. Thus, if one were to set a limit to the score here, it might result in too many cell line candidates or too few. Therefore, the top 5 cell lines of this list are proposed as the best in vitro models for research on breast cancer subtype Her2 (UACC893, KPL1, SW1990, BT474 and HCC1419), with DNA damage response deficiency.

To obtain the result of a few cell lines common among the top correlated ones from both expression and CN, implies that the holistic approach may be sufficient as it is. The aim of the holistic approach is only to deselect the bad cell line models, in which considering the entire genome in two levels of analysis should be enough. The entire genome should be considered in this part of the analysis because the intention is not to consider any details, only the major dissimilarities making some cell lines unacceptable models. The expression profiles, although not conveying proteomic or genomic profiles perfectly, are relatively highly linked to both genomics and proteomics, which is why this is important in this stage. The copy numbers add another layer to the analysis on a whole genome-scale, yet, it might not always be considered as heavily, in those cases where it is of less importance to the research project in question.

5.6 Purity Correction Investigation

This part of the project was carried out to investigate the relevance of purity estimates and whether or not they would only remove genes not expressed in the tumor cells. A list of genes negatively correlated to the tumor purity was created. The gene list was analysed using Enrichr [28][29], and the result gave several lists of pathways of which the genes were part of. The list can be seen in figure 19.

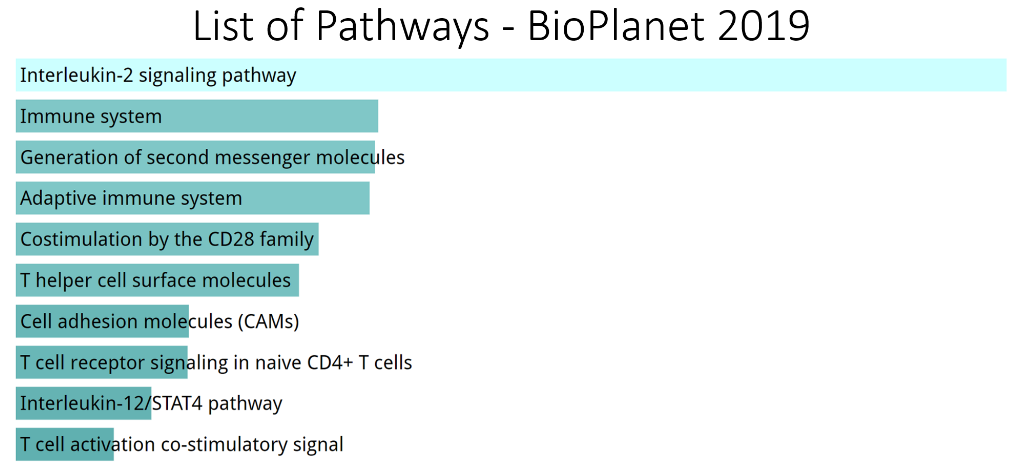


Figure 19: Enrichr generated list of pathways annotated to the genes most negatively correlated to tumor purity, for all tumor samples. The pathways are listed by number of genes contributing to each pathway, from highest to lowest.

All pathways in the list are part of the immune system which indicates that the purity estimates are accurate enough to not discard any tumor related genes and probably remove some noise from the data. Thus, the purity is something to take into consideration during future work, following this project.

6 Conclusion

When carrying out a comparative analysis between cell lines and the Her2 subtype of breast cancer tumors, 10 cell lines have been found to be among the 50 top correlated ones on both the genomic and transcriptomic level. Based on the gene signature expression, five of these could be selected as the potential best models for in vitro studies of the tumor subtype. These were UACC893, KPL1, SW1990, BT474 and HCC1419, where SW1990 has a pancreatic origin and the others are derived from breast tumors. The result of the BRCA subtype example indicates that the same analysis could be performed for many other tumor types and subtypes.

As the results have demonstrated, traits can differ considerably between tumor subtypes, which implies the importance of selecting a homogeneous population of tumor samples to represent the tumor of interest. In this project, the population of samples was collected based on molecular subtypes, resulting in further selection of a small number of cell line models after correlation analysis. This shows that using subtypes could be an acceptable approach to choose which cell line to use. However, other means to select samples have not been dismissed. Consequently, a better way to group samples could exist, or the approach best suited could depend on the research objective in question. Additionally, it is relatively simple to arrange tumor samples into groups of molecular subtypes, since data and tools are available for this, yet there are some tumor types with no annotated subtypes, in which case the selection of a homogeneous sample group will require a different approach.

Along with choosing an optimal subpopulation of samples, a well defined target disease will improve the accuracy of the cell line model selection. As there are many factors that should be considered in an optimal search for a good cell line model, a high amount of input about the target enables weighting of factors more or less, finding cell lines with the most important traits and an improved conclusion of what cell lines should be the best models.

As discussed, sometimes the best model cell lines for a given tumor phenotype are not of the same origin as the tumor in question. In the example of this report, that of BRCA subtype Her2, the results showed two out of ten cell lines with another origin than breast tissue; lung and pancreas. The reasons for finding cell lines with other origins could be, as mentioned, miss-classification, genomic drifting or metastasis. However, the fact remains that cell lines annotated as derived from another anatomically located tumor, can be good models for the tumor phenotype of interest. This supports the importance of an improved cell line model selection, where all cell lines should be considered.

6.1 Future Work

This project workflow alone is not enough to select the optimal cell line model, as too few factors are considered. Every study is different and for a given research objective, different factors will weigh more or less. For example, in one project a certain signature expression could be of the highest importance for the cell model selection, whilst a certain genomic alteration could be more significant in another. Because of this, additional factors are required to complete the selection. As mentioned earlier, the CN and expression analyses could be enough to deselect any bad cell lines, whereas the search regarding functional relationship should be improved. Therefore a future workflow is proposed, where proteomic data is added as a layer of analysis, and the functional relationship analysis is broadened to include several additional factors. Additional factors to be included should be selected genomic events and fraction of genome altered, on the genomic level. Furthermore, specified target genomic alterations and/or target expression should be included. Finally, proteomic expression should be taken into consideration as well, since this may show a different trend from the transcriptomic expression. A scheme of the proposed workflow can be seen in figure 20.

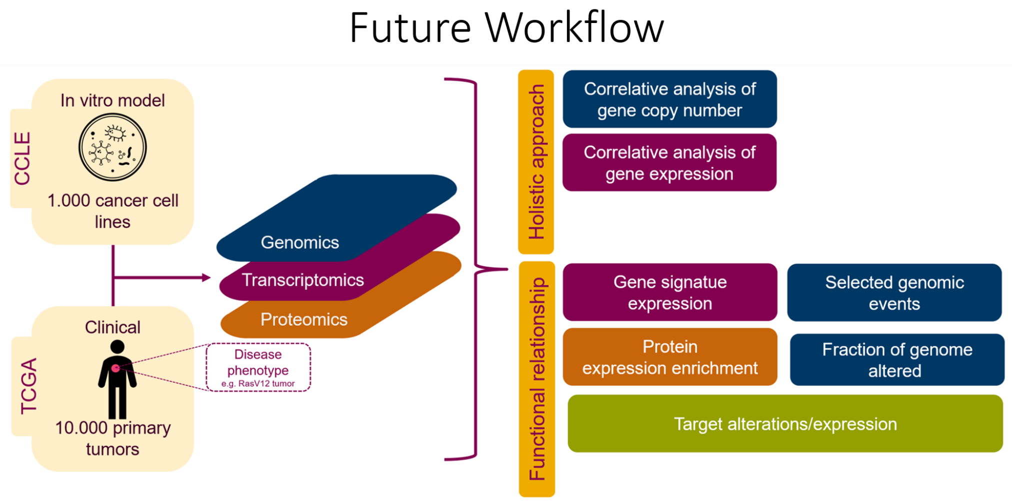


Figure 20: Scheme of a possible future workflow, where a third layer is added to the data analysis, proteomics. Furthermore, the functional relationship analysis is extended to consider several additional factors.

The expectations of a proteomic analysis is to review what conclusions could be drawn from the pattern that would be shown and to examine how to approach the proteomic analysis. The reason for the expectations being quite unspecific is that there many possible outcomes in protein production, which is dependent on factors such as the state of the cell, what other metabolites are present and the type of gene it is related to. For example, for some genes the mRNA to protein ration can be 1:1, whilst it can be significantly higher for others. Thus, the protein output can be difficult to predict. Furthermore, some proteins can be challenging to detect because of their chemical properties, or if a gene has been mutated the resulting protein could be lost in the detection instrument because of its changed properties.

Finally, for future studies in this topic, a validation of the cell line selection method will be required. A suggestion is to use available data from previous drug research projects, to compare which cell lines were chosen for in vitro studies and the results of the model selection process, to cell lines selected using the approach given in this project.

References

[1] Andrew Goodspeed, Laura M Heiser, Joe W Gray, and James C Costello. Tumor-derived cell lines as molecular models of cancer pharmacogenomics. *Molecular Cancer Research*, 14(1):3–13, 2016.

[2] K Yu, B Chen, D Aran, J Charalel, C Yau, DM Wolf, LJ van‘t Veer, AJ Butte, T Goldstein, and M Sirota. Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types. *Nature communications*, 10(1):1–11, 2019.

[3] Silvia Domcke, Rileen Sinha, Douglas A Levine, Chris Sander, and Nikolaus Schultz. Evaluating cell lines as tumour models by comparison of genomic profiles. *Nature communications*, 4(1):1–10, 2013.

[4] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.

[5] Richard LoCicero Kim Brown. How genes cause cancer health encyclopedia, university of rochester medical center, 2020.

[6] The Cancer Genome Atlas. National cancer institute, tcga. Accessed: 2020-06-03. url-https://www.cancer.gov/tcga, 2019.

[7] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A Margolin, Sungjoon Kim, Christopher J Wilson, Joseph Lehár, Gregory V Kryukov, Dmitriy Sonkin, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, 2012.

[8] National Cancer Institute. Nci dictionary of cancer terms. *Dysplasia*, 2016.

[9] Marina Salvadores, Francisco Fuster-Tormo, and Fran Supek. Matching cell lines with cancer type and subtype of origin via mutational, epigenomic and transcriptomic patterns. *bioRxiv*, page 809400, 2019.

[10] Karolina Janik, Marta Popeda, Joanna Peciak, Kamila Rosiak, Maciej Smolarz, Cezary Treda, Piotr Rieske, Ewelina Stoczynska-Fidelus, and Magdalena Ksiazkiewicz. Efficient and simple approach to in vitro culture of primary epithelial cancer cells. *Bioscience reports*, 36(6), 2016.

[11] Anuj Srivastava, Joshy George, and Radha KM Karuturi. Transcriptome analysis. 2019.

[12] Adam Shlien and David Malkin. Copy number variations and cancer. *Genome medicine*, 1(6):62, 2009.

[13] Aaron Theisen. Microarray-based comparative genomic hybridization (acgh). *Nature Education*, 1(1):45, 2008.

[14] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[15] Simon Koplev, Katie Lin, Anders B Dohlman, and Avi Ma'ayan. Integration of pan-cancer transcriptomics with rppa proteomics reveals mechanisms of epithelial-mesenchymal transition. *PLoS computational biology*, 14(1):e1005911, 2018.

[16] Lindsay I Smith. A tutorial on principal components analysis. Technical report, 2002.

[17] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[18] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell systems*, 1(6):417–425, 2015.

[19] Sonja Hänzelmann, Robert Castelo, and Justin Guinney. Gsva: gene set variation analysis for microarray and rna-seq data. *BMC bioinformatics*, 14(1):7, 2013.

[20] Adi L Tarca, Gaurav Bhatti, and Roberto Romero. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PloS one*, 8(11), 2013.

[21] Momeneh Foroutan, Dharmesh D Bhuva, Ruqian Lyu, Kristy Horan, Joseph Cursons, and Melissa J Davis. Single sample scoring of molecular phenotypes. *BMC bioinformatics*, 19(1):1–10, 2018.

[22] Dharmesh D Bhuva, Momeneh Foroutan, Yi Xie, Ruqian Lyu, Joseph Cursons, and Melissa J Davis. Using singscore to predict mutation status in acute myeloid leukemia from transcriptomic signatures. *F1000Research*, 8, 2019.

[23] Dvir Aran, Marina Sirota, and Atul J Butte. Systematic pan-cancer analysis of tumour purity. *Nature communications*, 6:8971, 2015.

[24] Francesca Finotello and Zlatko Trajanoski. Quantifying tumor-infiltrating immune cells from transcriptomics data. *Cancer Immunology, Immunotherapy*, 67(7):1031–1040, 2018.

[25] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic rna-seq quantification. *Nature biotechnology*, 34(5):525–527, 2016.

[26] S. Durnick et al. The biomart users guide. Technical report, 2020.

[27] Jianhua Zhang. How to use cntools. 2019.

[28] Edward Y Chen, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R Clark, and Avi Ma'ayan. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC bioinformatics*, 14(1):128, 2013.

[29] Maxim V Kuleshov, Matthew R Jones, Andrew D Rouillard, Nicolas F Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L Jenkins, Kathleen M Jagodnik, Alexander Lachmann, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, 44(W1):W90–W97, 2016.

[30] Singscore.R. R/multiscoregeneric.r. Accessed: 2020-03-27. url-https://rdrr.io/bioc/singscore/src/R/multiScoreGeneric.R, 2019.

DEPARTMENT OF BIOLOGY AND BIOLOGY ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se