Computational Approaches to Assessing Clinical Relevance
Of
Pre-clinical Cancer Models

Vladimir Uzun

The University of Sheffield
Faculty of Science
Department of Molecular Biology and Biotechnology

November 30, 2018

Abbreviations

AUC Area under curve

BRCA Breast invasive carcinoma

CCL Cancer cell line

CCLE Cancer Cell Line Encclopaedia

GP Gaussian process

MDA Mean decrease of accuracy

OOB Out-of-bag

PCA Principal component analysis

PDTX Patient-derived tumour xenograft

ROC Receiver operating characteristic

SVM Support vector machine

TCGA The Cancer Genome Atlas

TPM Transcripts per million

UCEC Uterine corpus endometrial carcinoma

Contents

List of Figures

List of Tables

Abstract

Preclinical cancer models, such as tumour-derived cell lines and animal models, are essential in cancer research. Consistently used as a platform to investigate mechanism of action, they can identify potential biomarkers prior to clinical trials where similar exploration is more complicated and expensive. However, whilst cell lines are the most used preclinical model, their applicability in certain settings is questioned because of the difficulty of aligning the appropriate cell lines with a clinically relevant disease segment. I developed a methodology for systematic cancer cell line scoring based on patient sample subtypes and analysis of the causative elements of the subtype differentiation in cancer.

Machine learning classifiers I tailored to multi-omics nature of cancer have been highly accurate in predicting the subtype of new patient samples. Applying those models to cancer cell lines reslted in a clinically based cancer cell line relevance score. The majority of cell line scores were in line with the literature, but there were several misclassified cells.

Exploring the causative elements of the underlying biology, I confirmed the oncogenic nature of the features driving the classification. Additionally, through differential expression analysis, the nature of some of the misclassified breast cancer cell lines was elucidated–they were poorly representative of their receptor-positive type despite having HER2 receptor expressed. One of those cell lines, JIMT-1, has been shown to be resistant to HER2-targeted treatment, thus making the misclassification of my model more clinically relevant than the receptor statuses of the cell line itself.

Through several distance metrics I have expanded on the binary nature of the classifying methods and identified more and less suitable cell lines not just by their score, but also by how close they are to the patient samples.

The core aspects of my methodology have been implemented as an online tool, a Shiny application, in order to allow others to leverage my methods and findings.

1 Introduction

Cancer is a diverse disease due to variety of tissues it can affect, mechanisms of action, prognoses and treatments. However there are several abnormalities present in all cancer cases, commonly termed as hallmarks of cancer (Hanahan and Weinberg, 2011): sustained proliferative signaling, evading growth suppression, resistance to cell death, replicative immortality, enhanced angiogenesis and metastases.

Cancer is one of the leading causes of death worldwide and on the rise (by the year 2035, yearly cancer cases in the United Kingdom are expected to increase by 42% from 360 000 to 513 000 (Smittenaar et al., 2016)), making it a priority in biomedical research. However, multiple disease causing mechanisms and histopathological heterogeneity mean that developing treatments is a highly complex problem (Gazdar et al., 2010).

One of the main goals of oncology-related research is the development of anti-cancer drugs. Before a potential drug can be used as a therapy it must go through several stages of research to ensure its effectiveness, safe dosage and side effects are elucidated. With 7% combined likelihood of approval, phase 2 and 3 drug failure rates in oncology are higher than those in other medical disciplines (Hay et al., 2014). Not only is this problematic for the search of an effective treatment for patients, but also carries a high financial and logistical price. Typically, phase 2 and phase 3 trials involve hundreds of participants and have a combined average cost of around 60 million USD (Sertkaya and Birkenbach, 2011). This means failed drug trials cost 56 million USD per year. One proposed cause of the high failure rate is inefficiency of current preclinical models in modelling cancer pathology (Yang et al., 2013). Since phase 2 and phase 3 trials involve patients and are preceded by preclinical model testing, unsuitable preclinical models could make an inefficient treatment seem promising.

Preclinical research in oncology is the irreplaceable first step in studying a promising novel compound. With less resources and no human subject involvement, preclinical models are the litmus screening test for the future of a potential new drug.

Models such as cancer cell lines (CCL) and patient-derived tumour xenografts (PDTX) are used as an approximation of the malignancy in a patient. This enables drug properties such as toxicity, appropriate dosage and cancer inhibition to be assessed. Whilst preclinical studies have less ethical, financial and logistical constraints than clinical studies, the availability of models that accurately reflect the patient population is crucial and a number of challenges exist such as overcoming the difference in clinical significance compared to the original tumour (Combest et al., 2012) or the need for evaluation of clinical significance. Nonetheless, effective design and execution of preclinical studies informs clinical trials and supports personalised healthcare hypotheses thus ensuring the right therapies are brought to the patients.

Progress in oncology rests on preclinical models that accurately reflect the patient population and their evaluation. Thus selecting the most appropriate model for a study is crucial.

1.1 Pre-clinical models

1.1.1 Cell lines

A cell line is a group of cells isolated from the organism which manage to grow *in vitro* and avoid cellular senescence. A cancerous tissue of origin makes a cell line a cancer cell line. Availability and low-cost make cancer cell lines virtually ubiquitous in early research stages.

Prior use of cell lines has led to the discovery of many of the current cancer therapeutics. For example, temozolomide, an astrocytoma and melanoma drug, was developed using MAWI and K562 cells (Newlands et al., 1997). In addition, they have been used to further understand current drugs, for example, to identify for which types of cancer a certain drug is effective (Niu and Wang, 2015) or if a known drug can be repurposed for new use in a cancer type (Verbaanderd et al., 2017). Beneficial results of cell line use can be attributed to several of their advantageous properties. Overall, cell lines have a clinically relevant genomic profile similar to the corresponding cancer (Gazdar et al., 2010) and are less resource-intense than the alternatives. Many cell lines are available as well as corresponding genomic and metadata (Barretina et al., 2012) which also contributes to the reproducibility of research.

However, they still hold some significant drawbacks:

In vitro culture entails a lack of the microenvironment which a cancer tissue would experience in an organism. Cancer microenvironment has been shown to have a significant role in development and survivability of some cancers (Krause et al., 2010). In such cases, cell lines could be poorly representative of their matching original malignancies.

Furthermore, study of mechanisms of drug delivery and drug toxicity is severely limited as there is no vascularity or non-target cells to suffer collateral damage.

Due to being isolated cancer cells which have managed to survive in vitro, genes associated with survival are up-regulated in cancer cell lines compared to cancer samples (Gillet et al., 2011). Thus, drugs targeting those mechanisms would be more effective on cell lines, but not the original cancer.

Such genetic composition, in addition to long post-isolation time-spans, might also explain why some different cell lines resemble each other more than their tissue of origin (Gillet et al., 2011) therefore, likely, having a significant genetic drift from the original tumour.

Maybe the biggest limitation is the cellular heterogeniety of cancer (McGranahan and Swanton, 2017). Being a mono-culture, cancer cell lines cannot capture heterogeneous tissues and are, thus, poor models for highly heterogeneous disease samples.

Even excepting these limitations, a cell lines may not always be what an investigator thinks it is. There have been numerous cell line misidentifications (within and between cancer types and species) in the past, the extensive list of which has been assembled by the International Cell Line Authentication Committee (Capes-Davis and Freshne, 2012). Another issue is contamination of stocks of one cell line with cells from another, especially with HeLa cells, which has been noted for numerous cell lines (Gazdar et al., 2010).

1.1.2 Patient-derived tumour xenografts

Patient-derived tumour xenograft (PDTX) models consist of a slice of tumour tissue isolated from a patient and directly engrafted into an animal, most commonly a mouse. The mouse needs to be immunocompromised in order for the xenograft to hold, otherwise its immune system would respond to foreign tissue. Implantation can be done subcutaneously (under the skin) or orthotopically (at the same body site where the origin tumour was situated). The xenograft can be passaged through multiple organisms before subjected to analysis. Since the tumour implanted in a mouse is directly taken from the patient, it is superior to cell lines in representing the patient tumour and augmenting personalized medicine in oncology. Because it is in a living organism, it is possible to study tumour-stromal crosstalk and stromal processes such as angiogenesis. Even though it is not the ideal model, a microenvironment does exist. Another benefit of being in vivo is the ability to study metastatic development of the cancer. Because cancer is of a heterogeneous nature, a tissue model can be more accurate than a cell line model.

With animal models, genetically modifying the host organism (usually a mouse) is an option for refining the study of the genetic basis of drug resistance and tumour susceptibility. Such methods can also help estimate the toxicity resulting from inhibitory effects of a drug by inhibiting the corresponding genes through genetic manipulation (Combest et al., 2012). Even so, they have a far higher resource and time requirement than cell lines. Furthermore, the patient stroma will eventually be replaced by mouse stroma after early passage, therefore the microenvironment might inadequately represent the human one. This is further exacerbated if the xenograft is not implanted orthotopically. Not only do the species-specific features mismatch, but so do the microenvironment-related ones because the different tumour site might mean particularly dissimilar microenvironments. Furthemore, because of the compromised immune system of the animal host, the higher growth rate of a xenograft than the original malignancy can make antiproliferative drugs seem more effective than they actually are (Combest et al., 2012).

Drugs which interact with the immune system should not be tested in PDTX models since animals need to be immunocompromised for the implant not to get rejected (Gazdar et al., 2010).

1.1.3 Other models

A number of other preclinical models are available. Cell line xenografts cover cell lines' lack of microenvironment. Genetically engineering mouse models enable more accurate disease modelling while also providing a functional immune system. Organoids and organs-on-a-chip bring more multi-cellular complexities without animal tissue involvement, but are still rather novel. Although better at some aspects than cell lines, all of these models are much more resource demanding and do not have systematic analyses on a level of cell lines (Barretina et al., 2012).

1.2 Cancer subtyping

Cancer can arise in different tissues within a general site of original, and through different mechanisms. This leads to cancers existing as a range of different subtypes - malignancies which are part of the same type, but behave differently because of relatively small, but important, genomic differences. Subtypes might have separate developmental trajectories and respond to different treatments. Knowing which subtype a cancer corresponds to, aside from the general site of the disease, is key in finding the most appropriate therapy for the patient.

For example, hormone-dependent cancers can be treated with drugs that target receptors of the hormone in question. Breast cancers can be categorized according to the different hormone receptors they have (ER, PR, HER2). Survival rates and treatment options of the cancer vary depending on the subtype (Parise and Caggiano, 2014). Drugs that target the mentioned receptors (e.g. tamoxifen for the ER, trastuzumab for the HER2 receptor) and thus affect that specific cancer subtype have been developed. Further subtyping of established cancer subtypes has additionally improved the understanding of the disease development and prognosis

(Sabatier et al., 2011).

Thanks to subtype investigation, more is known about the molecular mechanisms of cancer and treatment options. Cancer subtyping can be done based on different information about the specific cancer cohort or sample, such as its size, differentiation, results of laboratory testing or genomic profile. Due to development of sequencing technologies, the use of the latter has seen significant increase in recent years.

The genetic profile of a cancer can influence many of its mechanisms, and, consequently, the response to a treatment. Thus, in order to make the selected treatment as effective as possible, it is necessary to distinguish cancer by genomic subtypes.

A 2000 patient cohort of breast cancer samples had their genome and transcriptome analysed along with clinical outcomes (Curtis et al., 2012). Through integrative clustering, ten molecular subgroups were identified, carrying different clinical prognoses. Additionally, potential new cancer driving events were identified by examining copy number alteration effect on outlying expression.

The Cancer Genome Atlas (TCGA) has been a target of many -omics analyses with its over 10 000 samples spread across 34 cancer types (Weinstein et al., 2013). For example, molecular profiling of lung adenocarcinoma (Collisson et al., 2014). After DNA, mRNA and miRNA sequencing, copy number, methylation and protein analyses were carried out with clustering and GISTIC, a tool that detects cancer-driving genes targeted by somatic CNAs (copy number aberrations) (Mermel et al., 2011). Most commonly found mutations (KRAS, EGFR, BRAF) were already previously identified as potential driver mutations (Collisson et al., 2014). However, in the samples with the absence of those mutations, new mutations were found (MET, NF1, RIT1).

Such cancer subtyping studies have resulted in a large body of publicly available data which helps the confirmation of previously suspected mechanisms as well as discovery of new ones. Having such data publicly available allows for a better starting point for secondary analyses and should be further exploited.

Cancer subtyping using genomic profiles continues to be vital in furthering understanding of the disease. Applying a similar approach to cell lines, I hope to help improve the use of preclinical models by providing a systematic assessment of cancer cell lines from patient sample genomic data using machine learning methods.

1.3 Machine learning

The majority of algorithms used fall within the machine learning umbrella. Machine learning is a computer science discipline whose fundamental goal is creating models that can make predictions or uncover underlying patterns based on previously gathered data. The data represent a collection of data points or samples, each consisting of features which can be used to evaluate the data point. An example of a data point could be a set of mutations in a biological sample, pixel colours in an image or values of shares on a stock market in a certain time period.

If the data points have an associated, varying class label (e.g. genome with the presence of a certain phenotypic condition) or evaluation score (e.g. probability of a medical event), it can be used by machine learning algorithms to learn to assign that label or score to new, unassigned samples. This is the goal of *supervised learning* - predicting class assignment (for example, if an image contains a cat) or continuous score (for example, what the values of market shares will be) of new data points based on the features and known assignments of old, known data. The former is called classification (it puts samples into discrete classes) and the latter regression analysis (it associates samples with a continuous score, typically by fitting a mathematical function). Features and the corresponding evaluation tune the chosen computational model. The goal is to assign the correct class or score to new data points for which that information is lacking. With enough data and an appropriately chosen model, score/class of unobserved, similar in origin, data points can be predicted with a high accuracy. Accuracy, as measured by various metrics (F_1, AUC, specificity, sensitivity, . . .), is typically tested on the data points from the original data set which have not been used for model training.

Being able to predict an outcome of a preclinical model experiment, such as drug testing on a cell line (Menden et al., 2013), before doing it can enhance the decision making process behind it. Having predictions of drug responses across a panel of cell lines, for a specific drug, improves the cell line selection for the experiment and potentially reduces the number of different cell lines needed, thus, conserving valuable resources.

When no class labels or evaluation scores are available or of interest, *unsupervised learning* algorithms such as hierarchical clustering or principal component analysis (PCA) can still be used to explore potentially useful relationships between subgroups and/or samples. For example, for data points made up of mutation profiles of clinical samples of some cancer type, we might want to see if certain mutations co-occur or if there exist distinct subgroups of that cancer type.

Various clustering methods are commonly used in cancer genomics because of their ability to identify subtypes, discover new classes and visualise multi-dimensional data: hierarchical clustering (Koboldt et al., 2012), non-negative matrix factorisation clustering (Sia et al., 2013), t-distributed stochastic neighbour embedding (Jamieson et al., 2010) and PCA (Schmidt et al., 2008) being some of the most popular.

Semi-supervised learning, a hybrid of supervised and unsupervised learning, is typically used when there are data points for which labelling makes sense, but only a small subpopulation of them actually have labels. It has seen some success by performing equally or better than the more traditional types of machine learning (Park et al., 2014).

Machine learning in oncology

An assortment of machine learning techniques have been employed for cancer progression prediction and cancer subtyping.

Support vector machines (SVM) models have been particularly popular - they have been successfully applied in breast cancer recurrence estimation using clinical and pathology features (Kim et al., 2012), prediction of age at diagnosis using SNPs in multiple myeloma (Waddell et al., 2005), discovering anti-cancer peptides (Hajisharifi et al., 2014) and finding most responsible factors in recurrence of cervical cancer (Tseng et al., 2014).

Artificial neural-networks are amongst the earliest machine learning methods used with clinical and cancer data. They have accurately distinguished, based on ultrasound data, between renal cell carcinoma and non-cancerous renal cysts (Maclin et al., 1991). Other notable uses are lung cancer survival prediction (Chen et al., 2014) from gene expression data and mapping out oral cancer progression, alongside support vector machine (Chang et al., 2013).

Bayesian networks, a graph-based probabilistic method, have found successful applications in colon carcinoma survival modelling (Stojadinovic et al., 2011), oral cancer recurrence prediction through genomic, clinical and imaging features (Exarchos et al., 2012) and breast cancer microarray-based prognosis (Gevaert et al., 2006).

Tree-based models (models made by combining decsion trees), although easier to understand than most, have been as effective in numerous cancer analysis tasks. The ease of interpretation comes from the structure of a tree-model being made of many branches based on a feature value (e.g. is the expression level of a certain gene greater than the threshold or not?). A lot of other models do not have decision-making encoded as clearly internally (e.g. Gaussian process model determine values based on a correlation matrix that is not as intuitive to disentangle).

A decision tree using an entropy-based information gain splitting algorithm performed at least as well as an artificial neural network when predicting breast cancer survivability from clinical and environmental features (Delen et al., 2005).

Random forest, an ensemble method based on decision trees, projected a clinically relevant risk-adjusted survival rate following lymphadenectomy in esophageal cancer, based on clinical and behavioural features, in order to determine the isolated effect of lymphadenectomy on patient survival (Rizk et al., 2010). Using microarray gene expression, it identified biomarkers in leukaemia, colon and prostate cancer (Ram et al., 2017). Recently, random forest and a variation of the model were successful in predicting drug sensitivity in cancer cell lines (Rahman et al., 2017).

Gaussian process models were used for selection of potential gene biomarkers in prostate cancer from microarray data (Chu et al., 2005), for definition of functional volumes in radiation oncology based on radiation PET scanning cancer data (Shepherd and Owenius, 2012) and, among many other examples, estimation of survival time in renal clear cell carcinoma through multiple -omics data (Molstad et al., 2019).

Various types of clustering methods are a common tool in genomics analyses with variations of integrative clustering being the most common ones.

Integrative clustering has been used with joint copy number - gene expression data to separate unique lung cancer subtype profiles (Shen et al., 2009) and to create finer subdivisions of subtypes in the breast cancer luminal A subtype (Aure et al., 2017). Especially with TCGA dataset analyses, non-negative matrix factorisation, hierarchical or integrative clustering and their variations are typical in trying to further granulate the subtype landscape (Getz et al., 2013).

An improvement over traditional hierarchical clustering in creating predictive survival rate clusters based on gene profiles in mesothelioma cancer has been made with semi-supervised recursively partitioned mixture models (Koestler et al., 2010).

Using a breast cancer survivability dataset semi-supervised learning co-training method achieved high accuracy despite the lack of labelled data in the dataset (Kim and Shin, 2013).

A novel, graph-based, semi-supervised learning algorithm improved, compared to previously used methods, recurrence prediction by 25% in breast, colorectal and colon cancer (Park et al., 2014).

For dimensionality reduction purposes, aside from the ubiquitous PCA, t-SNE (typically used for cell subpopulation representation in single-cell genomics (Saadatpour et al., 2015)) and multi-dimensional scaling (MDS) have seen successful use (Abdelmoula et al., 2016).

1.3.1 Description of selected algorithms and their application

The following machine learning techniques are the algoirithms used for creation and analysis of the predictive computational models.

Random forest

A random forest model (Breiman, 2001) is an ensemble model - a model which combines multiple other models, in this case it is a collection of many single decision tree models. It can deal with high-dimensional low-sample size data easily (Lavecchia, 2015) and it was shown to be able to detect less extensive changes in the variables, leading to better predictions (Sarica et al., 2017). However, it struggles, compared with other models, with multi-class classification problems (Pranckevičius and Marcinkevičius, 2017) and has less probabilistic interpretation than models like Gaussian process-based ones (Rasmussen and Williams, 2005).

Each decision tree is a tree-like acyclical structure with a root node, leaves and intermediate nodes. When a pre-constructed decision tree is applied to an appropriate but unseen sample (one with same names of features as the training set), it decides in each node, starting from the root, what the next (daughter) node is based on a feature and a threshold specific to the node. Class label assigned to the sample is dependant on which leaf node this procedure ends in.

Generally, decision trees are built by selecting the most discriminative features until all the training samples have a properly assigned class in a leaf node. This is usually achieved through choosing, for each node, the feature which, when used as a splitting feature in the node, results in the most unbalanced class representation (f_{14} would be chosen over f_{13} for the splitting variable for the node in (Figure 1)). If a node ends up having only samples of the same class, no further tree development is necessary, from that node. In a random forest, each decision tree is trained on a different subset of samples (i.e. bootstrapping) and a different subset of features. This diverse collection of decision trees is used to determine the class of a new sample. Every decision tree assigns the class (or class probability) to the sample and the average of all those is the final assigned class.

Because of the variable subset of features included in a single tree, overfitting is much less of an issue than in other machine learning models and a more diverse set of potentially significant variables are enabled to play a role.

Another important characteristic of this strategy is the internal estimation of the error (out-of-bag error estimation): Since only a subset of samples are used to build a single tree, the samples which were not can then be used as an internal test set for that tree. This removes the need for a separate test set and allows for a bigger

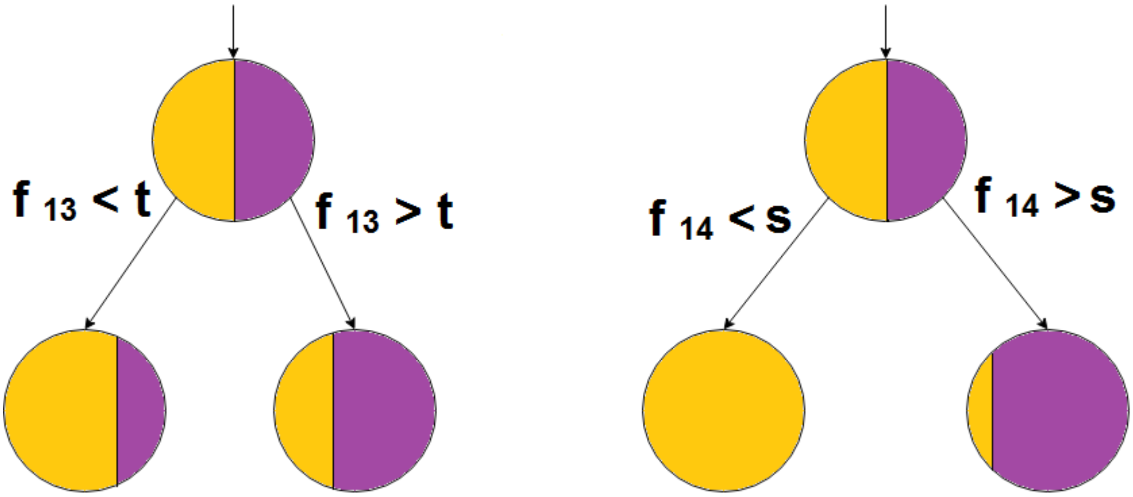


Figure 1: Two features as possible splitters are examined for the same node. The parent node is at the top with equal class distribution (same number of samples for both classes). Distributions of samples in the daughter nodes are dependent on which features was chosen as a splitting feature in the parent node. Typically Gini impurtiy or information gain metrics are used to compare the outcomes of different feature choices. Both strive to have a homogenous population in a node and, thus, both would opt for f_{14} over f_{13}. t and s are threshold values which, for the feature they are used with, split the population in the best way.

training data set. However, as this option is unique to the random forest, I still employed the standard training - test set strategy for the random forest model.

Feature importance can be obtained from the out-of-bag (OOB) samples by calculating, for a feature m and tree t, how many correct classifications are there in the OOB estimation for t, subtracting from it the number of correct classifications when m is randomly permutated in OOB samples, and then taking the average of it across all the trees.

Gaussian processes classification

A Gaussian process (GP) is stochastic process (a collection of random variables along a continuous dimension, usually time) where each random variable of the process has a Gaussian (normal) distribution (Rasmussen and Williams, 2005). It is similar to multivariate Gaussian distribution except that the set of potential variables is infinite. Each random variable represents a potential sample with their values being class labels or regression values to model and their probability function made up from the features in the input data.

It has seen recent increase in use as a machine learning tool (Rasmussen and Williams, 2005). Some notable uses include aiding detection in radiation oncology (Shepherd and Owenius, 2012), white matter fibre clustering (Wassermann et al.,

2010) and finding subpopulations in single-cell RNA data (Buettner et al., 2015). Because it uses kernels (functions which project the data, which might be hard to separate (Figure 2), into a higher-dimensional space, making sample separation easier (Figure 3)), it can deal more easily with problems that are not linearly separable or modellable (Hofmann et al., 2008). The output is fully probabilistic (Gunn, 1998), making it more usable in some later analyses (e.g. estimating confidence intervals or combining with a different probabilistic method). Compared to other methods, especially support vector machines, a method which shares some of the advantages with GP, the standard implementation of a GP model has greater computational complexity (Hensman et al., 2013) thus requiring more resources.

Gaussian process regression and classification try to model the goal function (predictive function in the case of regression and separating function for classification) by using the covariance between random variables (points in the process) to estimate the values of new points. The closer the random variables are, the more they affect the value of each other. The exact relationship is dependent on the choice of the covariance function, their distance and the length-scale (how close they need to be to have a significant effect).

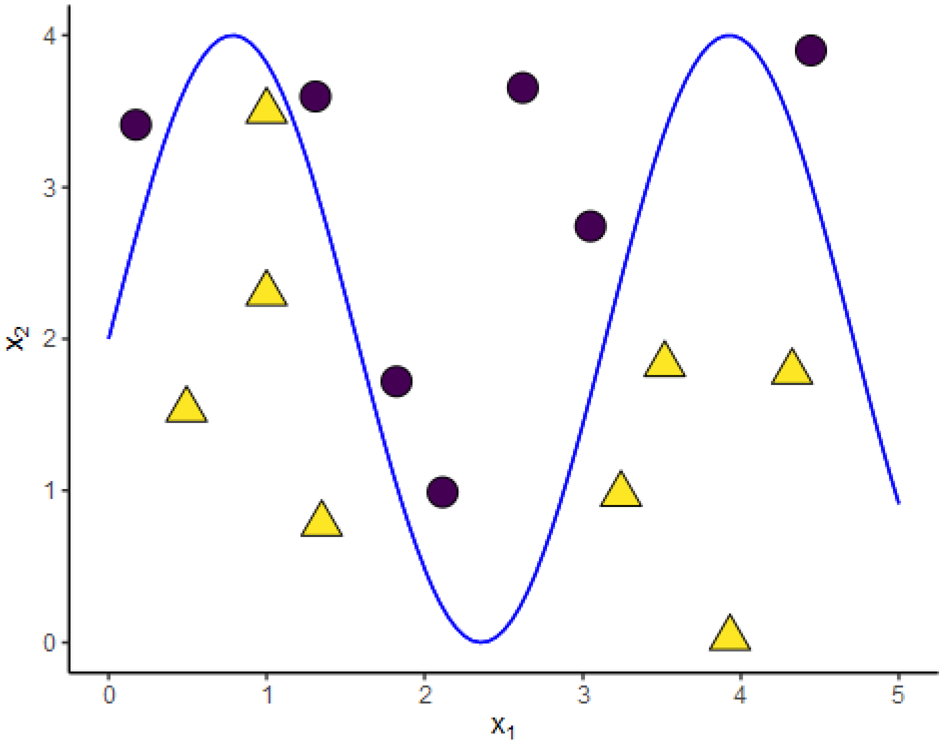


Figure 2: Sometimes the sample subtypes are not easy to separate in the original feature space (as defined by the obtained features prior to any transformations). In this example, sample classes (denoted by colour/shape) we wish the model to differentiate are distributed in a way which would make only a non-linear separator fully accurate

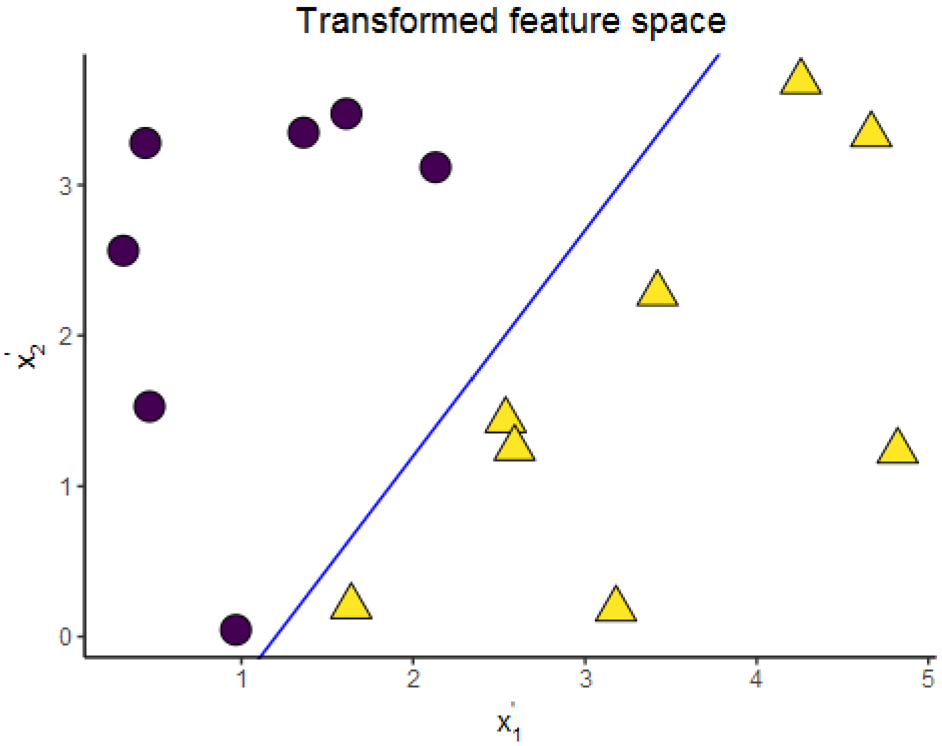


Figure 3: Applying a kernel (mathematical transformation) on the original features (x_1, x_2) can simplify the separation. While the new features ($x_1^{\prime}$, $x_2^{\prime}$) might be not directly interpretable (since they are transformations of the original features), the separation of classes is possible with a linear classifier.

Support Vector Machine

The support vector machine (SVM) model is a classifier which creates a class separation hyper-plane and maximizes its distance to the points closest to it (support vectors) in order to create as wide separation between classes as possible (Figure 4).(Cortes and Vapnik, 1995)

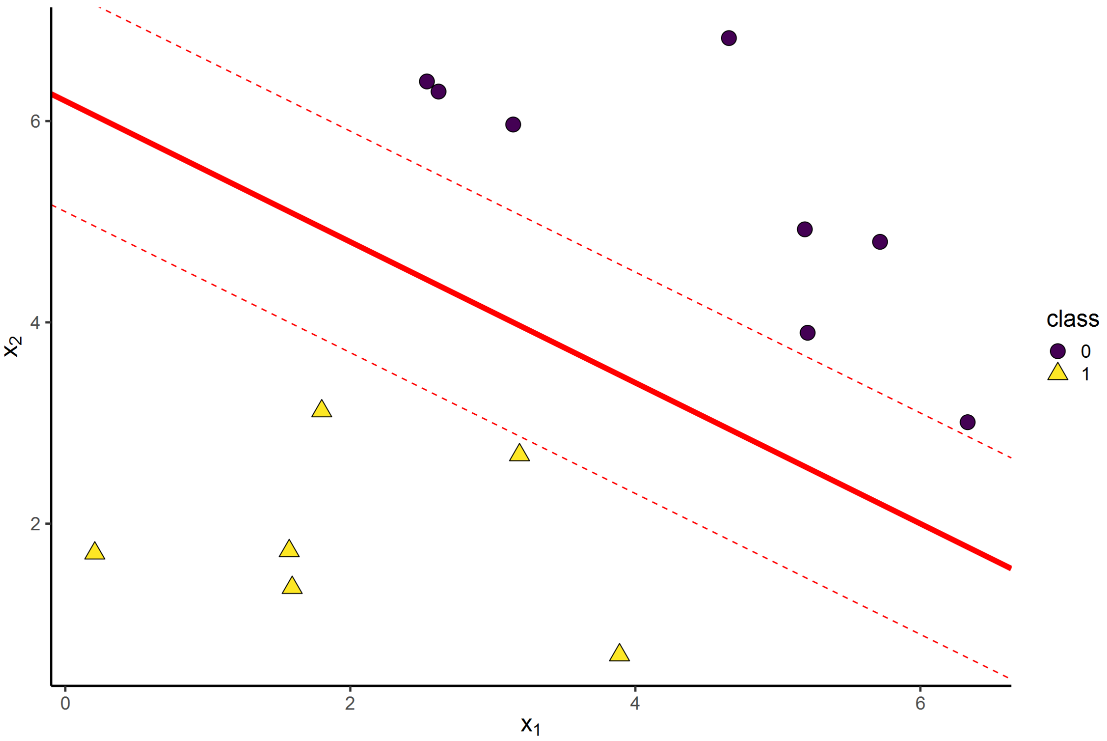


Figure 4: An example of a separating plane an SVM model might come up with for the separation of differently coloured/shaped samples. Any separator that does not cross the dashed lines would be successful with the test set in question, but the middle full line is the one that is furthest apart from both subgroups. SVM models do not just aim to seprate the subtypes, but to also make that separation as big as possible.

While SVM is a linear classifier, through the use of kernels (in a similar manner to the GP model), it can create a linear separation in a higher dimensional space (created from the original data through kernel functions) which is effectively non-linear in the starting feature space, thus allowing for more complex classes to be correctly separated and achieve better accuracy (Hofmann et al., 2008). The way separation plane problem is set-up mathematically leads to optimisation of a convex function (meaning there is no local maxima) in the model training phase, so it has a simpler/faster optimisation process than most other methods which have to deal with the issue of being stuck in a local maximum (Gunn, 1998). The downside of SVM models is that they are less interpretable (Lavecchia, 2015) than most (modelling for a good separator plane says little about the samples it separates) and consequently have an output which is harder to use in further analyses without additional transformations.

A hyper-plane can be defined through points $\vec{x}$, weights $\vec{w}$ and offset b as $\vec{w} \times \vec{x} - b = 0$. If y is a binary class variable with values 1 and -1, then the SVM requirement of points being on the appropriate side of the hyper-plane is $y(\vec{w}\vec{x} - b) > 1$

Generally, the finding of an appropriate hyper-plane separator can be formulated as trying to minimise the following expression(Equation 1) with respect to $\vec{w}$.

$$[\frac{1}{n}\sum_{i=1}^{n} max(0, 1 - y_i(\vec{w}\vec{x}) - b)] + \lambda\|\vec{w}\| \tag{1}$$

The first addend is the penalty for the point being on the wrong side of the plane, summed over all the points. The second addend is penalty for the separation size (the bigger the $\|\vec{w}\|$, the smaller the distance between the groups) with the choice of λ determining the magnitude of the penalty.

The minimisation is done through an optimisation algorithm, typically gradient descent.

K-means clustering

A common approach in computational analysis of big biological datasets has been clustering (Curtis et al., 2012). Since some clustering methods are quick to test while representing a standard application of machine learning techniques, I have employed a k-means clustering algorithm as a comparison model to the main ones.

The k-means algorithm partitions a set of data points into K clusters aiming to separate underlying subtypes. K, the number of clusters, is provided at the start of the algorithm. K centroids, cluster centers, are randomly initiated. The data points are then assigned to the cluster of their nearest centroid, typically determined through Euclidean distance. New centroids are then calculated, as means of the data points assigned to their cluster. Data point assignment and centroid calculation steps are repeated until there is no change in the structure of the clusters.

In order to decide on the number of clusters, silhouette value (Rousseeuw, 1987) for evaluation of clustering methods was used. k-means grouping with different number of clusters(k) was then assessed using the silhouette measure.

Silhouette value or index (defined as *s(i)*, Equation 2) evaluates an instance of clustering by comparing intra-cluster similarity (how similar samples that are grouped in the same cluster are, later defined as *a(i)*, Equation 3) with inter-cluster similarity (how similar, i.e. different, is a sample from one cluster to samples from other clusters, later defined as *b(i)*, Equation 4). Every sample has its own silhouette value derived from similarities (most commonly distances) to other samples. Silhouette index of a cluster is the average value of all the silhouette values of the samples belonging to that cluster. Averaging the indices of the clusters gives the evaluation of the entire clustering procedure. Silhouette values range from -1 (worst) to 1 (best). Exact formulas used for the calculation of silhouette value in this case are shown below. i and j stand for samples within clusters, C and *C'* represent clusters.

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{2}$$

$$a(i \in C) = \frac{\sum_{j \in C, j \neq i} |i - j|}{|C|} \tag{3}$$

$$b(i \in C) = min_{C' \neq C} \frac{\sum_{j \in C'} |i - j|}{|C'|} \tag{4}$$

Principal component analysis

PCA is a data transformation technique which reshapes the coordinate system of the data in a way that makes the new data features (axes of the coordinate system) capture the most variance in the data (Figure 5).

It is often used for dimensionality reduction and data preparation before other analyses, but it also provides information about the effect of the original features on the variance through their weights in the linear combinations which make the new features.

Typically, PCA starts by centering and rescaling the data through mean subtraction and standard deviation division for each of the features. This step ensures that some variables do not get a higher importance simply because they have larger values. For example if there was a variable with a range from 1000 to 1001 across thousands of variables, we would not want for it to have higher impact on determining the PCA result compared to a variable with a range 0 to 1. Afterwards the covariance (Equation 5) is calculated for each feature pair to make a covariance matrix (Equation 6). Eigenvectors of the covariance matrix (v) and their corresponding eigenvalues (λ) can be found by solving the Equation 7 Ordered by their eigenvalues, the eigenvectors are then used as a rotation matrix which transforms the original data into principal components (Equation 8).

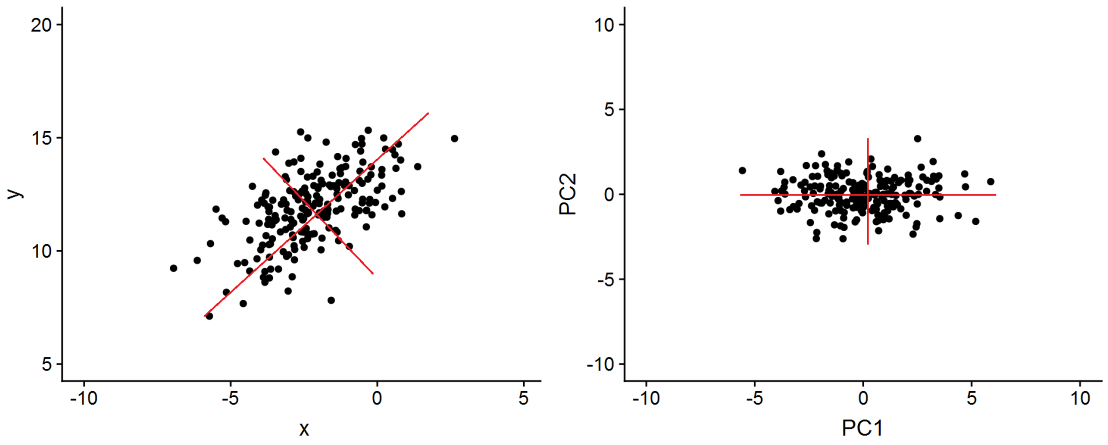


Figure 5: An example of PCA transformation on 2-dimensional data. x and y are original features of the samples shown as black dots. The directions of biggest variance are shown with straight red lines. PC1 and PC2 are principal components resulting from PCA transformation of the original data. PCA identifies the spread of variance in the data and generates prinicipal components - new features which capture the most possible variance. The first principal component captures the most variance, the second principal component captures the second most variance and so on. This way features with low variance or high correlation with other features end up in the background enabling easier identification of interesting patterns or simplifying further analysis.

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{(n-1)} \tag{5}$$

$$C^{nxn} = (\forall i, j \in N)\, cov(Feature_i, Feature_j) \tag{6}$$

$$(C - \lambda I) \times v = 0 \tag{7}$$

$$TransformedData = v^t \times OrginialData \tag{8}$$

1.3.2 Metrics used for model evaluation and comparison

F_1 score and area under receiver operating characteristic curve are common performance measures for binary classifiers with continuous class assignment. Since most of the methods fit that description, difference in F_1 score and AUC (area under curve) are used as an accuracy comparison metric. Other methods (principal component analysis, k-means) are not inherently compatible with those metrics, but were adjusted for their use.

Area under curve is area (integral) under an ROC (receiver operating characteristic) curve (Hastie et al., 2009). ROC curve is a classifier diagnostic contrasting true positive (also called sensitivity or recall) (Equation 10) with false positive rate (Equation 9) for different values of class threshold. For a binary classifier that outputs probability of a sample belonging to one class or the other (or other continuous variable), threshold value determines which probabilities assign sample to which class. The curve is constructed by calculating and plotting of true/false - positive rate pairs for all the threshold values. As the threshold changes from more to less conservative, more samples are classified as "positive" for the target class leading to a, generally, higher true positive and false positive rate. ROC curve helps with threshold selection by visualising changes in true positive and false positive rates across all thresholds. The bigger the area under ROC curve, the higher true positive rate for lower values of false positive rate is and, thus, the better the classifier is at prediction.

Three examples of ROC curves and their AUC are shown in Figure 6. A perfect classifier would have a 100% true positive rate for a threshold which makes the false positive rate 0% with the area underneath the curve being 1, whereas a classifier that assigns classes at random (e.g. through a coin-flip) would have the true positive and false positive rates match each other for any threshold selection making the area under the curve 0.5.

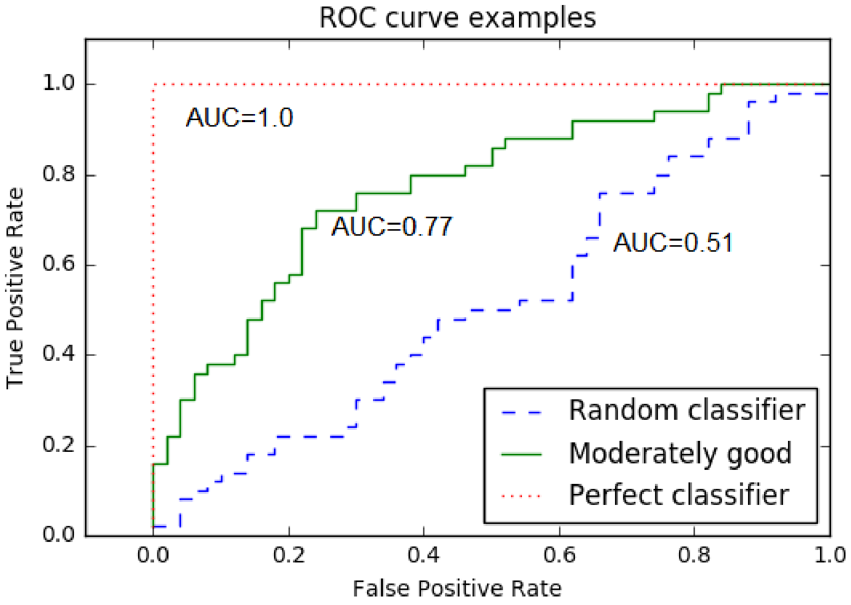


Figure 6: ROC curves and area under curve values for a random, moderately accurate and a perfect classifier.

F$_1$ score (Equation 13) is a special case of an F-score measure (Equation 12) that combines a binary classifier's precision(Equation 11) and sensitivity(Equation 10) to give an accuracy score (Hastie et al., 2009). Parameter β determines the contribution of sensitivity as opposed to precision. With $\beta = 1$, they both contribute equally. Higher value places more weight on sensitivity while lower one does so for precision. F_1 score and AUC are both commonly used classifier performance measures which both use sensitivity (true positive rate) in their calculations. F_1 score focusing on precision gives importance to minimising false positives as a fraction of all positively assigned samples while AUC with the false positive rate helps to minimise false positives as a fraction of all samples that are negative. Sample imbalance in the favour of true positives (versus false positives) makes precision alone less effective, while imbalance in the favour of true negatives (versus false positives) does so for the false positive rate.

While these measures are related to each other, it is important to keep in mind that optimising for one of the measures does not lead to the best case for the other one (Davis and Goadrich, 2006).

Since the dataset used in this comparison section has been trimmed to have balanced class sizes, both metrics should perform similarly. However, since positives are the rarer and more fatal subtype (serous), precision, and thus, F_1 score has a slight favour in this specific case.

$$FalsePositiveRate = \frac{FalsePositives}{TrueNegatives + FalsePositives} \tag{9}$$

$$Senstivity = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{10}$$

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{11}$$

$$F_\beta = (1 + \beta^2) \times \frac{precision \times sensitivity}{(\beta^2 \times precision) + sensitivity} \tag{12}$$

$$F_1 = 2 \times \frac{precision \times sensitivity}{precision + sensitivity} \tag{13}$$

1.4 Previous efforts to evaluate the clinical relevance of cancer cell lines

Awareness of cancer cell lines as problematic models is not a new phenomenon. Several approaches already contributed to their better use.

There have been comparisons based on genomic profiles between cell lines and their respective cancer type (Qiu et al., 2016). Numerous cancer cell lines have been found to be misclassified (Capes-Davis et al., 2010). Unfortunately, identifying the misclassification and making changes to practice sometimes takes a while to have effect. For example a cell line previously thought to be of breast origin, MDA-MB-435, has been discovered to be a melanoma cell line (Ellison et al., 2002). However, it was continued to be treated as a breast cancer cell line for several more years (Rae et al., 2007). The International Cell Line Authentication Committee (Capes-Davis and Freshne, 2012) is an effort to catalogue all the known cell line misidentifications in hopes of reducing the use of inappropriate cell lines.

Typically, cell line misidentification findings and cancer type wide comparisons focus on a small cell line population and analysis driven by predetermined set of genetic markers for the cancer type in question. To my knowledge, there have been few systematic comparisons - ones that would take the entirety of the available genomic information as a base for comparison or a larger cancer and cell line cohort, especially with respect to known cancer subtype.

Ovarian cancer cell lines used as high grade ovarian serous cancer models were compared to the patient samples of that cancer subtype using TCGA and CCLE (Cancer Cell Line Encyclopaedia) data (Domcke et al., 2013). Each cancer cell line had a genomic signature constructed from the overall copy number and mutation profiles (average rates of alteration in the whole genome) and a few specific driver events (e.g. TP53 mutation status). How similar that signature was to the cancer type determined the cell line's suitability score, that is how similar it is to the original high grade ovarian serous cancer. A rarely used cell was discovered to be a highly suitable model while one of the commonly used cell lines was a poor match for the high grade ovarian serous cancer.

In a similar manner, a weight-based model was used to compare cancer cell-lines to tumours for six different cancer types (Sinha et al., 2015). The weight assigned to a feature depended on whether the feature is a known driver event for cancer generally, a specific type or not a driver at all. Despite finding moderate resemblance to the appropriate tumour type in most cell lines, several outlier cell lines that do not fit the overall patient sample profiles were identified through their method.

In both of these cases (Domcke et al., 2013) a number of specific cancer related features were used to assess the suitability of a cell line for studying a particular cancer type. This runs the risk that cell lines may become unrepresentative at loci not normally associated with that cancer type, but this not being picked up because of the predetermined relevant features. It is also unsuitable in cases where the key differences separating subtypes of interest are unknown. As can be seen in the case of endometrial cancer and its endometrioid and serous subtypes (Figure 7), although cell-lines can have a mutational profile vastly different from the tumours, there can still be better or worse choices with respect to overall similarity to a subtype. In the case of predetermined relevant features, cell lines which are suboptimal, but with vastly different appropriateness, could be deemed equally suitable because of the reliance on a small subset of features.

A recent cancer cell line evaluation methodology and tool (CELLector (Najgebauer et al., 2018)) brings fresh options for cancer cell line selection. Using a set of oncologically relevant genomic features, it builds a tree structure where each node corresponds to a combination of molecular alterations, resulting in subtyping of the chosen cancer cohort hierarchically. Cancer cell lines are mapped to the nodes/subpopulations with the same alterations signature.

An advantage of this method as opposed to the previously discussed ones is integrated tumour subtyping, application to all the cancer cohorts (as the subtyping is dependent on the list of all known genomic drivers, not cancer type specifics) and identifying tumour subpopulations lacking a representative cancer cell line model (generated nodes without a cancer cell line matching its alteration signature).

While useful and widely applicable, there are some potential drawbacks. The overall genomic cell line to tumour differences play no part as the method is focused on a predetermined set of features. This also makes it possible to miss out some important genomic features which have not been registered as known drivers.

1.5 Aims and structure

The goal of this project was to explore the suitability of typical machine learning methods in subtyping cancer cell lines based on patient sample subtypes and the cell line results obtained in the process. Additionally, the core parts of the developed methods were implemented in a tool to enable researchers to quickly score their cell lines for a chosen subtype.

With an automated model and score building tool, we can assign a clinical relevance score to cell lines for a specified cancer subtype context. This insight helps the cell line model selection and reduces the chances of using inappropriate models.

While score based on subtype is not always appropriate, the ability to assign any kind of subtyping and the lack of such an approach in other methods make it an advantage of this cell line evaluation scheme.

An additional strength lies in the use of machine learning models, allowing for a

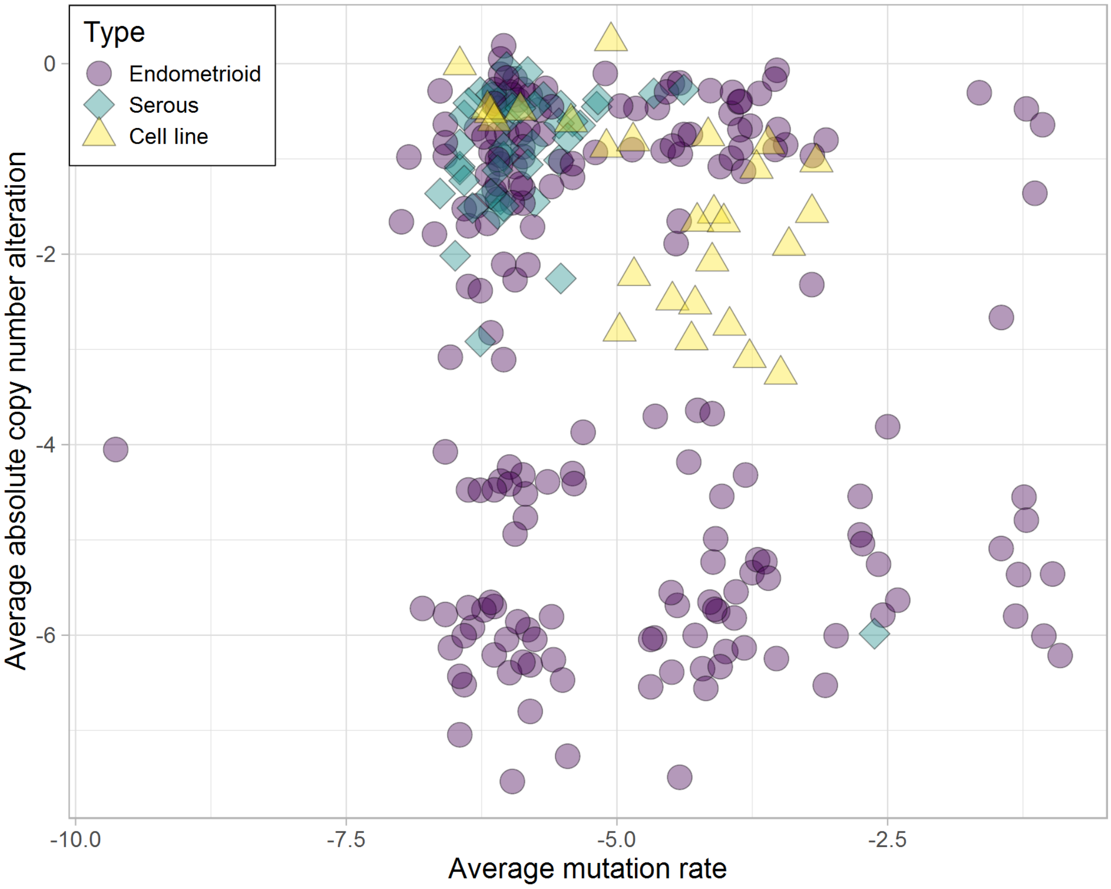


Figure 7: Mutation and copy number trends in endometrial cancer endometrioid and serous subtypes as well as cancer cell lines. Data was obtained from TCGA for cancer samples and CCLE for cell lines. Copy number data is from the GISTIC2 files. Mutation data was obtianed from the .maf files.

broader coverage of features and outsourcing the decision of feature importance to established machine learning techniques. By not focusing on predetermined features, chances of missing an important, but under-represented driver, are lessened. Machine learning methodology also provides a framework for determining feature importance as a result of the mathematical optimisation of the model.

Structure

In the following chapter, data, different methods, their combinations, the way they were used to build the computational models and overall structure of the model application will be described. Application of the models to the patient data test set is used to score and compare the different methods.

Application of the developed models to cancer cell lines, as well as their literature subtype review, is covered in chapter 3. Other methods are also used to gain more insight into the cell line landscape and their relationship to patient cancer samples.

Through differential expression and gene set analysis, chapter 4 explores the meaning and significance behind outlier cell line results, in addition to examining the driving genomic features in the constructed models.

Finally, the core methodology is implemented in an interactive application, allowing anyone to obtain a score for a chosen cell line and cancer subtype combination.

The code used for development and analysis can be found on https://github.com/vuzun/clscoring.

2 Development of a general cancer subtype classifier

2.1 Introduction

In this chapter a general purpose framework for creating classifiers is developed for classifying cancer subtype from TCGA patient data which can then be applied to similar data from cell lines in future chapters. I begin with a comparison of selected models (see previous chapter) by testing them on a single data set - endometriod copy number data. After selecting random forests as the model type to use, the three different data types are compared for their ability to inform subtyping models to distinguish the endometrial serous vs endometriod subtyping and receptor positive vs triple negative breast cancer typing problems. Finally methods for producing classfiers based on multiple data types is explored.

Prior to scoring the cell lines, the scoring algorithms were applied to the cancer test set - a subset of the cancer data which had not been used in the training process. If the algorithm is good at recognising underlying differentiating patterns, it will assign scores which are in line with the known subtype of the samples in the test set. The better the correspondence between the model's evaluation and the known subtypes of the test set, the better the accuracy of the algorithm.

Accuracy on the patient test set provides a way of assessing the sensibility of the general methodology before it is applied to cell lines. If the methods are not performing well on the patient test set, there is not much to be gained from applying them to cell lines. But if they perform well, they are differentiating between the cancer subtypes of interest in patient samples, making it sensible that cell lines' score they generate from cell line data will be tied to cancer subtypes in question.

The result is a framework for the creation of subtyping models and highly accurate models for classifying endometriod vs serous endometrial cancer (AUC=0.98) and receptor positve vs triple negative breast cancer (AUC=0.94) that can be applied to classify CCLE cell line data in the next chapter.

The general methodology for cancer cell line evaluation implemented here (Figure 8) is to first build and validate a classifier using patient cancer data and appropriate cancer subtypes as classes, and then apply the classifier on cancer cell line data to produce a cell line evaluation in the context of the chosen subtype. This way the method and the score of the cell lines are completely based on the patient data.

Although classifiers require the annotation of training examples into classes and determining their proper number, there are several advantages to this approach:

- Having defined classes allows the model to ignore potentially more present but less relevant features which can have confounding effects in clustering methods. Because clustering methods deal only with features (e.g. gene expression), high variance features are always highly influential in the generation of clusters. However, it can happen that some high variance features bear no clinical or molecular significance for the dataset analysed. In that case, they would be an irrelevant feature which is still driving the model building and affecting assignment of samples into clusters. But in the case of classification (or, similarly, other supervised learning), there is a biologically relevant class/label (e.g. disease subtype predictive of survival) assigned to samples which makes the model select the features which help distinguish between the classes. This way a high variance feature, which is not helping separate the classes/groups of interest gets disregarded (if it has high variance, but same distribution in the class-based groups, the classification algorithm will not select for it).
- Since the model-building algorithm determines the relevance of the features, it is highly applicable to a set in which key alterations are not known and, similarly, includes potentially relevant features which might have been otherwise neglected
- Through the significance assigned to features by the classifier, there is potential to identify new class drivers

Random forest, Gaussian process classification and support vector machines are the classification algorithms used here. Although they are classifiers they predict a class label with a probability which then gets rounded up to the closest class. This is advantageous because it allows for a better understanding of the classification results and expands the analysis of the accuracy of the results.

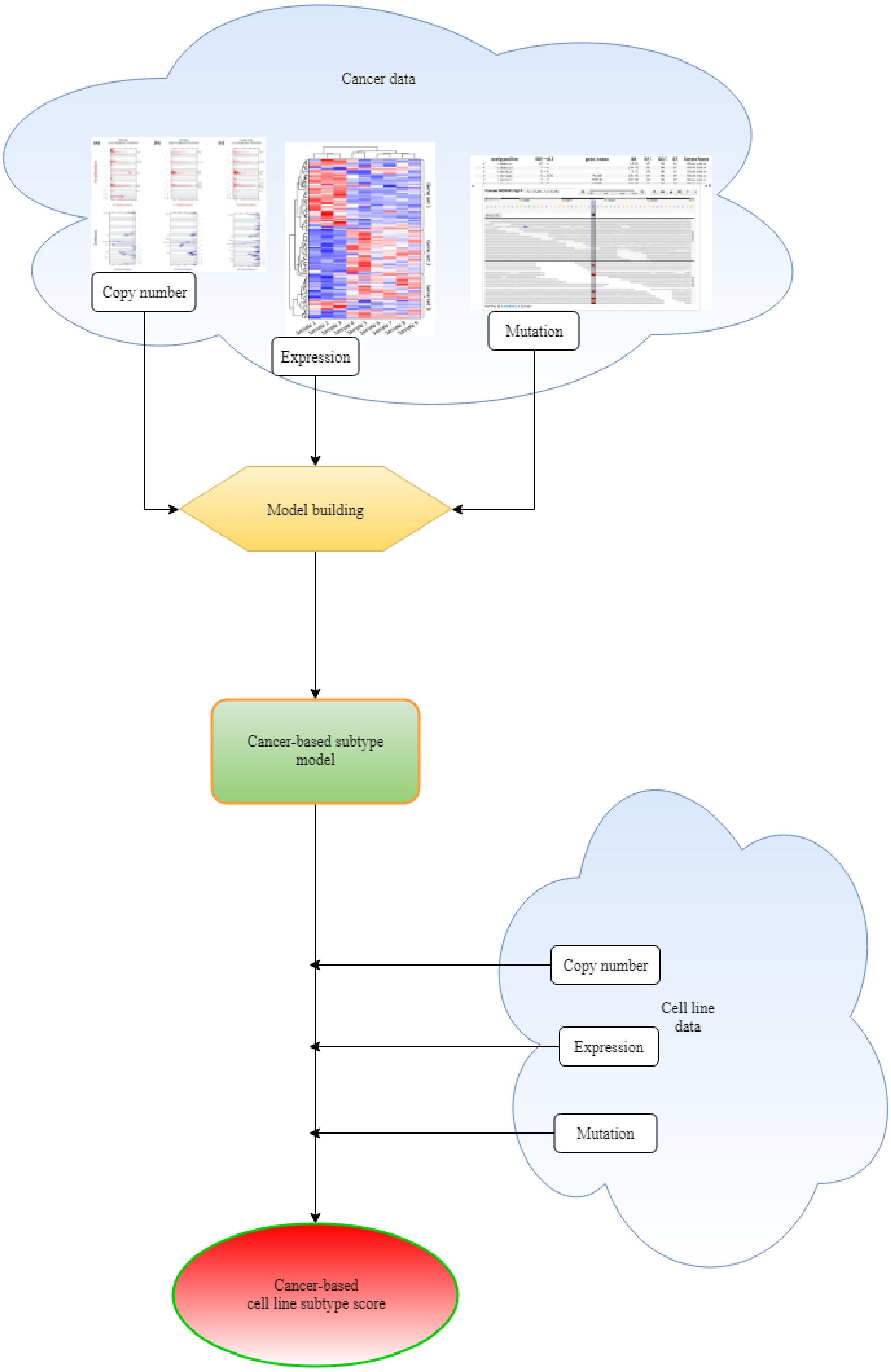


Figure 8: General methodology of the cancer cell line analyses. Patient cancer data is used to create a cancer subtype prediction model which is subsequently applied to cancer cell lines to obtain a cell line score based on patient data.

2.2 Methods

2.2.1 Datasets

Patient cancer samples and cancer cell lines were obtained from TCGA (Weinstein et al., 2013) and CCLE (Barretina et al., 2012), respectively.

I focused on endometrial carcinoma (UCEC) and breast cancer (BRCA) because of their large sample size and presence in the subtyping literature(Getz et al., 2013). TCGA contains 1098 samples from breast cancer patients and 560 of endometrial carcinoma.

Genomic properties used were expression, mutation and copy number in the form of RNA-seq count data and estimated transcript counts per million (TPM), mutation annotation format, and GISTIC2 (Mermel et al., 2011) copy number scores per gene. Classes for cancer samples were based on the clinical information for the samples (see below).

TCGA data was accessed through the firebrowse portal (http://firebrowse.org) for each of the cancer types used. GISTIC2 scores were taken from *"all_data_by_genes.txt"* in the "CopyNumber Gistic2" tab. Mutation annotation file, downloaded from the MAF archives on the same page, was processed using *maftools* (Mayakonda and Koeffler, 2016) R package to obtain a binary (cells being either 1 or 0) matrix of mutations per gene. RNA-seq RSEM normalised data was downloaded from mRNAseq archives and manually converted to TPM data. Histology column ("*CLI_histological_type*") of "Aggregate AnalysisFeatures" was used for assigning classes (endometrioid and serous) for endometrial cancer (UCEC). Breast cancer classes were based on the receptor information (progesterone, estrogen, HER2) in the "All_CDEs.txt" file from the "Clinical_Pick_Tier1" clinical archive. Samples with missing or ambiguous class assignment were excluded from the dataset.

CCLE was accessed through the official Broad Institute portal (https://portals.broadinstitute.org/ccle). Files used were copy number segmentation, mutation annotation file, RNA-seq RPKM expression and raw RNA-seq counts. The copy number segmentation file was processed with a GISTIC2 pipeline on the GenePattern (Reich et al., 2006) platform in order to obtain a data type compatible with the TCGA copy number data. After selecting GISTIC2, *.seg* file and a markers file (containing, in order, names and starting points from *.seg* file) were submitted for processing resulting in a GISTIC2 analysis files of the same structure as TCGA's.

RPKM gene expression data was converted to TPM by normalising the gene expression values across all the sample sizes within the dataset (R code of the conversion can be found at github.com/vuzun).

Patient sample size and feature numbers varied by cancer and data type (subsubsection 2.2.1). The same was the case for cell lines (Table 2.2.1). When multiple data types were used at the same time, the sample size was the size of the intersection of the samples for each data type by TCGA barcode.

For methods that were not able to handle the size of the data, features were ordered by their variance and only the top 20% were selected.

Gene copy number, processed expression (TPM) and mutation data was used to build models and score cell lines. Raw counts of cell lines were used in the later analysis of different cell line groups.

Data type	Sample size	Number of features
BRCA - copy number	1083	24776
UCEC - copy number	539	24776
BRCA - mutation	987	13391
UCEC - mutation	248	15200
BRCA - expression	594	10886
UCEC - expression	186	10886

Table 1: Sample size and features for different cancer type and genomic information combinations.

2.2.2 Sample and feature subsetting

For model development studies a training set of 50 endometrial and 50 serous samples was chosen at random from the complete TCGA dataset. The remaining samples were used as the test set (50 serous and 389 endometrioid).

Cell line data type	Sample size	Number of features
BRCA - copy number	28	32109
UCEC - copy number	59	32109
BRCA - mutation	51	1639
UCEC - mutation	27	1639
BRCA - expression	57	56318
UCEC - expression	28	56318

Table 2: Sample size and features for cancer cell lines by cancer type and genomic information combinations.

Model comparison studies used copy number data from TCGA. When testing the effect of feature reduction on classification accuracy and timing, subsets of 0.2, 0.4, 0.6 and 0.8 of all features were selected at random. For the final classifiers the 2000 features with the highest variance were selected.

A second dataset was also used. The same samples and training/test split were used, but with only 2000 features. The features were chosen from the initial 19006 based on their variance - features that differentiate between the classes will also have noticeable variance on the entire dataset. Methods are evaluated on this set for their accuracy.

2.2.3 Model implementations

Random Forest

The random forest algorithm used is from R package randomForest(Liaw and Wiener, 2002) installed from CRAN (https://cran.r-project.org/web/packages/randomForest/index.html).

The *randomForest* function which creates a random forest model was called with the following arguments - *formula* set to *class* ~ .; *data* set to a data frame of training set samples with columns as features and a class variable indicating the sample subtype; *importance*, *keep.forest* and *proximity* set to TRUE; *ntree* set to the number of trees in the forest; and *mtry* set to the number of features considered when creating daughter nodes.

The default number of trees (*ntree*) is 500. Generally, more trees are better, but after a certain threshold, there is no improvement in the prediction. By varying the number of trees from 10 to 1000 (Figure 9), the improvement seems to stagnate after around 500 trees. 800 seems to be ideal for the data set in question (copy number data or endometrial cancer) and gives a large margin of error while not changing the time of the execution much, so 800 was decided on as the parameter value for *ntree*. The default setting for *mtry* (model parameter determining the number of features to be tested for in a node of a tree) was used.

Function *predict()* is used to make predictions from the model by having the random forest model as the first argument, the new unclassified samples as the second argument and explicitly setting *type* argument to *"prob"*.

An example run:

```
RF_model <- randomForest(class ~ ., data=training_set,
                         importance = TRUE,
                         keep.forest = TRUE,
                         proximity = TRUE,
                         ntree = 800)

predictions <- predict(RF_model, newdata = testing_set,
                       type = "prob")
```

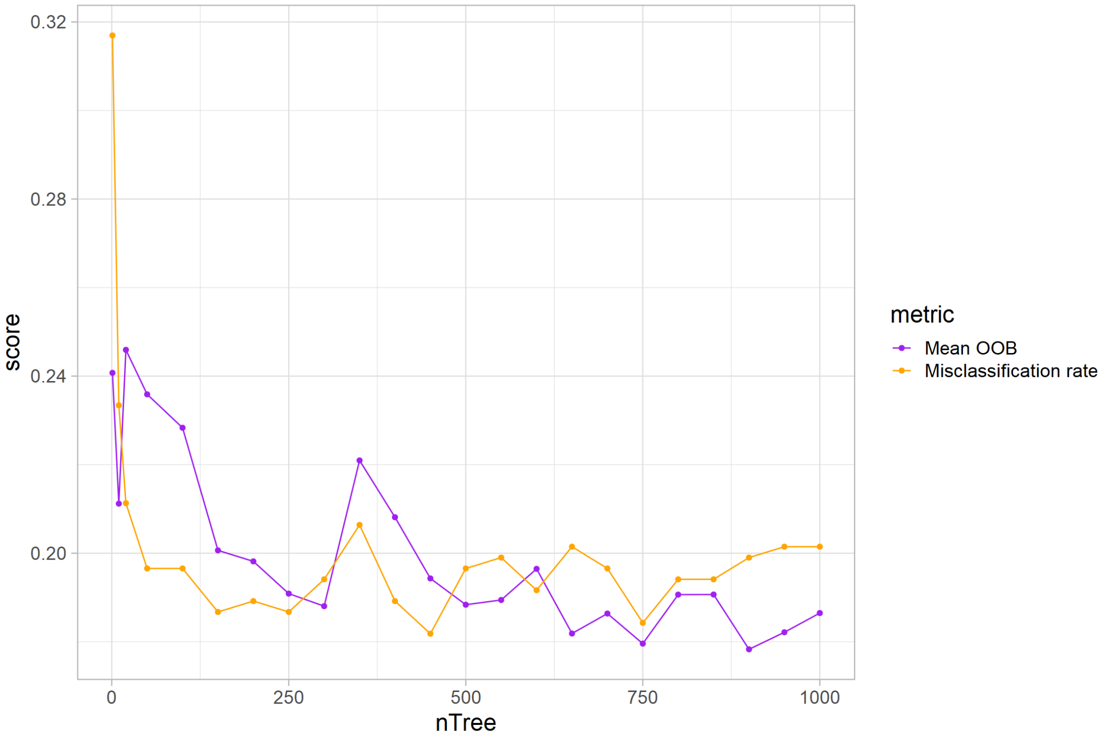


Figure 9: Error rates (OOB and misclassification rate) generally go down as the number of trees increases, but stagnate after a certain number. As the increased number of trees means more time and memory is necessary for the model to be trained, it is important to select the smallest number which leads to highest accuracy.

Gaussian Process classifier

Gaussian process classification model from the GPy Python library was used for all GP models. Function *GPy.models.GPClassification()* creates the model from two *numpy* arrays - sample features and the corresponding classes vector. The kernel function was the RBF (radial basis function) kernel.

Next, *.optimize()* method of the model with no arguments is called to train the model. Without specifying any values for arguments, the default values are used (1000 as the maximum number of iterations in the optimisation step, and limited-memory BFGS algorithm for the model optimiser).

Afterwards, with the argument being the array of sample features for which a prediction is desired, *.predict()* method of the model can be used to make predictions on the test set or new samples.

An example run:

```
import numpy as np
import GPy

GPC_model=GPy.models.GPClassification(training_samples,
                                      training_classes)
GPC_model.optimize()
test_set_prediction=GPC_model.predict(testing_samples)
```

Support Vector machine classifier

SVM model was created with the model-building function *svm.SVC()* and trained with the model method *.fit()* from the Python package *scikit-learn.*

The function *svm.SVC()*, with the argument *probability* set to *True*, is used to generate the model and then the method of the model *.fit()* with training set and class vector is used to optimise the model to the data of interest. The kernel used was the default *'rbf'* kernel.

K-means clustering

K-means algorithm used is part of the stats R package (R Core Team, 2018) which uses the Hartigan-Wong(Hartigan and Wong, 1979) version of the algorithm.

Inputs of the *kmeans* function (https://stat.ethz.ch/R-manual/R-devel/library/stats/html/kmeans.html) were data, as rows of samples, and the number of clusters to be made. Output is a k-means object, one of whose elements is cluster assignment of the samples which was used in further analysis.

Principle component analysis

PCA algorithm used is *prcomp* from the R stats package (R Core Team, 2018)

The input of the function, in all cases, was a data frame of samples and *scale.* and *center* set to *TRUE*. This was to make sure the data is scaled and centred before the transformation. The result is the *prcomp* list which contains standard deviation of the components, weights (rotations) which were used to transform the data and the transformed data.

2.2.4 Other

Speed of execution was measured through R's *tictoc* and Python's *time* libraries. Only the time spent on model-building and optimising was measured. Data preparation and loading times were not tracked.

All the method execution pertaining to this chapter was run on the high performance cluster at The University of Sheffield.

Cross-validation

In the cases of algorithms which have a training and a test set, 10-fold cross validation was conducted in order to ensure prediction stability. https://github.com/vuzun/RF

Firstly, the sample size was trimmed so that it would contain equal amount of samples from both classes. If there is significantly more of one class in the training set, the classifier will be more sensitive to that class error which can skew the predictions in the dominant class direction.

After accounting for proportionality of the classes, the set of all samples was partitioned into ten subsets, nine of which made the training set with the remaining one being the test set. This was performed ten times, with a different test set every time. Finally, the values assigned to the samples in the test set and the cell line samples were averaged out.

This way, every sample gets to be in the training and the test set the same number of times.

Calculation of AUC scores for classifiers

AUC scores were calculated using *performance()* function from ROCR library (Sing et al., 2005) for the R and scikit-learn's *roc_auc_score()* function (Pedregosa et al., 2011) for the Python code.

The ROC curve visualisations were done using matplotlib Python library (Hunter, 2007).

Calculation of F1 scores and AUCs for non-classifier methods

Random forest, Gaussian process and support vector machine are classifiers capable of generating a zero-to-one probability score describing their confidence in a sample's class.

Since PCA and k-means are not classifiers by design, the following approach was used to produce a pseudo-score in order to enable comparison with the classifying algorithms.

PCA of the training data was performed with the base R function *prcomp()* as described above and then combined with the patient sample subtype class to create subtype clusters in the principal component space. Assignment of new samples was done by the distance of their PCA transformation to the PCA transformation of the subtype cluster centers. Euclidean distance weighted by the principal components' variances (Equation 14) was scaled to match the probability assignment of the classifiers (0 to 1 probability of a sample being one class, and the sum of all class probabilities for a sample being 1) and inverted since lower distance indicated higher fit with the cluster, as opposed to the class score (Equation 15, Equation 16). This enables compatibility with ROC and F_1 statistics.

$$ClusterDistance^2 = \sum_i \frac{(PC_i - PC_i^{ClusterCenter})^2}{VarianceExplained_i} \tag{14}$$

$$scaledClusterDistance_A = \frac{ClusterDistance_A}{ClusterDistance_A + ClusterDistance_B} \tag{15}$$

$$scaledClusterDistance_B = \frac{ClusterDistance_B}{ClusterDistance_A + ClusterDistance_B} \tag{16}$$

A similar procedure was used with k-means. K-means clustering results in k clusters based on feature similarity. The number of clusters used was 2 (not only is it much more convenient for comparison with the classifiers, but it also has the highest Silhouette score (Figure 10)). Samples which are the majority subtype in the cluster were considered true positive/negative, while the others were considered false positive/negative. F_1 scores were calculated using this grouping. To facilitate ROC statistic compatibility, distance of a sample to the generated clusters was scaled to the classifiers' probability score and inverted.

While these adaptations are sensible considering what information the metrics use (closeness to cluster center is proportional to how representative the sample is of that cluster, similarly to how a class score can be used to say how similar a sample is to that class), they are not normally used and there is no proven validity of such adaptations.

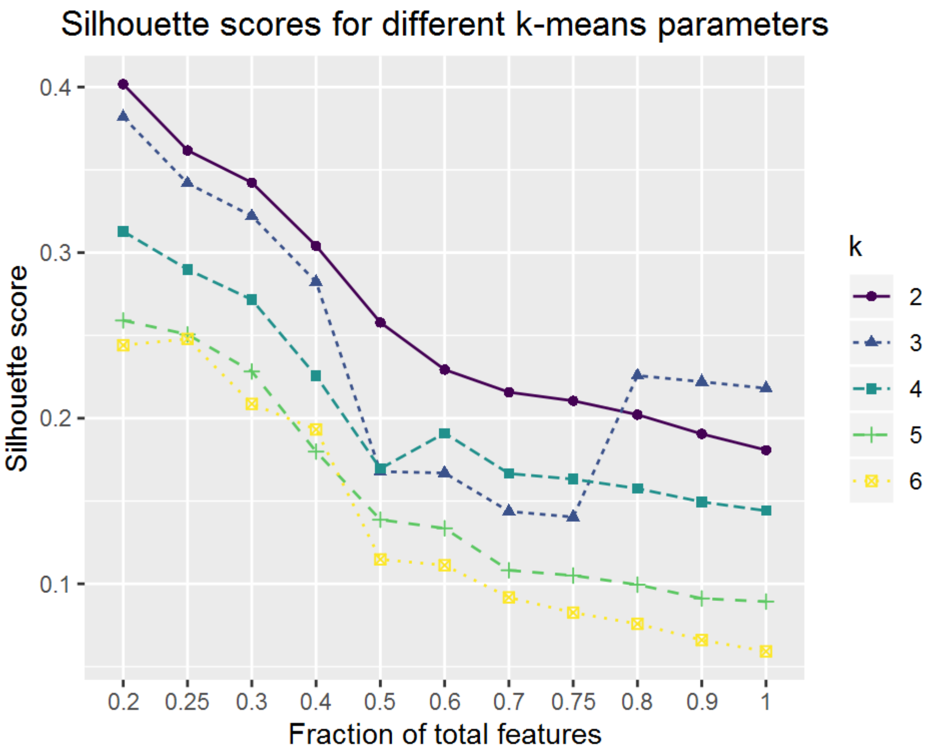


Figure 10: Silhouette scores (Rousseeuw, 1987) for different k-means settings. Different values of k (2-6) are shown through different colours and lines. Fraction of total features denotes the number of features used from all the available features. Best choice for the dataset seems to be to use lower number of features and $k = 2$. Silhouette score decreasing with the number of features added can be explained by additional features not bringing any new relevant information and, thus, reducing the impact of the important features.

2.3 Comparison of subtype analysis methods

Which machine learning or statistical method is the most suitable to use for an analysis is heavily dependent on the data it will be applied to. For example, linear models have the most interpretable solutions and a faster execution time, but if the underlying data subgroups are not linearly separable, it will perform poorly. If the data has a time component which plays an important role in modelling, specialised time series methods will be more suited for it than standard approaches (Längkvist et al., 2014). Artificial neural networks have seen great popularity lately, but they require large number of samples in order to be effective.

In this section, the methods covered in the methods section (random forest (Breiman, 2001), support vector machine (Cortes and Vapnik, 1995), Gaussian process (Rasmussen and Williams, 2005), principal component analysis (Pearson, 1901) and k-means (MacQueen, 1967)) will be compared, on the same data sets, through a set of comparison metrics.

To compare the effectiveness of the different classification approaches, models were trained using copy number data from 100 endometrial cancer patients, equally split between subtypes (endometrioid and serous). The remaining 439 samples are mostly endometrioid (50 serous and 389 endometrioid samples). The data was split in this manner to counteract the potential problems that can sometimes arise from having an unbalanced class proportions in the training data, while also leaving a significant amount of samples of both classes in the testing set. For example, if the training set consist of 99 samples of class A and 1 sample of class B, if it classifies all the samples as A, it will have a 99% accuracy, but it will still be a fairly poor classifier. By having classes equally represented (50% each) in the training set, it is made sure that the classifier treats errors on each of them equally. Support vector machine (Figure 11), Gaussian processes (Figure 12) and random forest (Figure 13) classifiers trained on this data all performance well, with AUCs exceeding 0.92 in all cases. Gaussian process classifier performed better than the other two with all the features, but only by 0.02.

Trimming features from the maximal size can help experiments run faster and ease portability. Effect of feature size on the accuracy of each of the methods was examined across a varying number of features, selected at random from 20 to 100 percent of the total feature number (19 006). The information about the scalability of the models can also help inform later decisions. All models still showed an improvement with more features (Figure 11, Figure 12, Figure 13), although the difference between using only 20% of features and all features was surprisingly small.

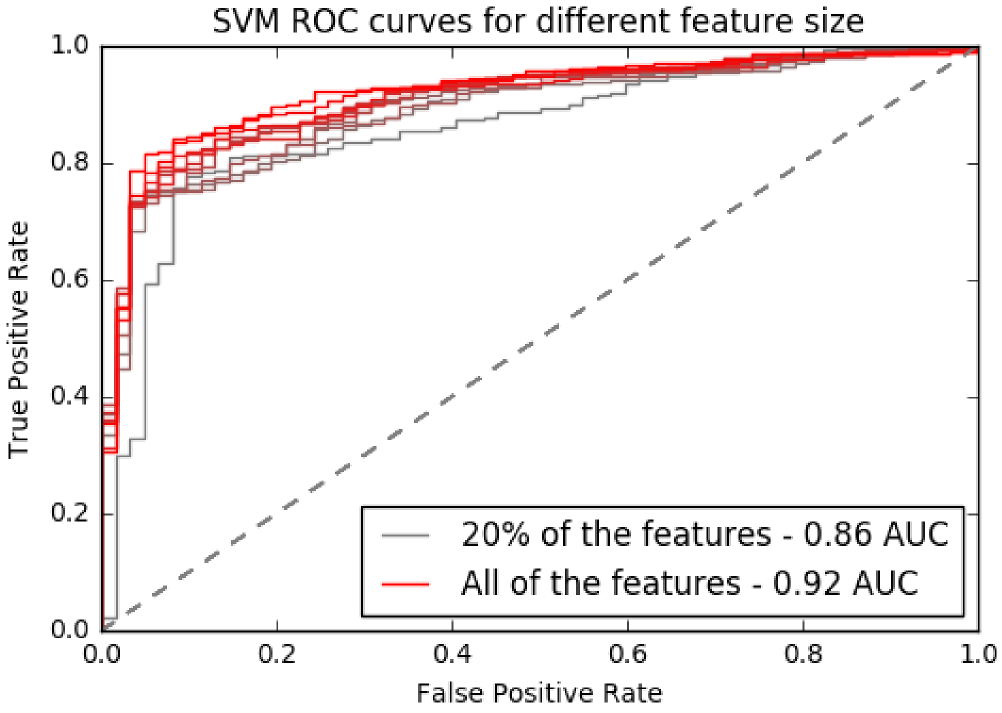


Figure 11: Support vector machine classifier performance across a range of feature numbers

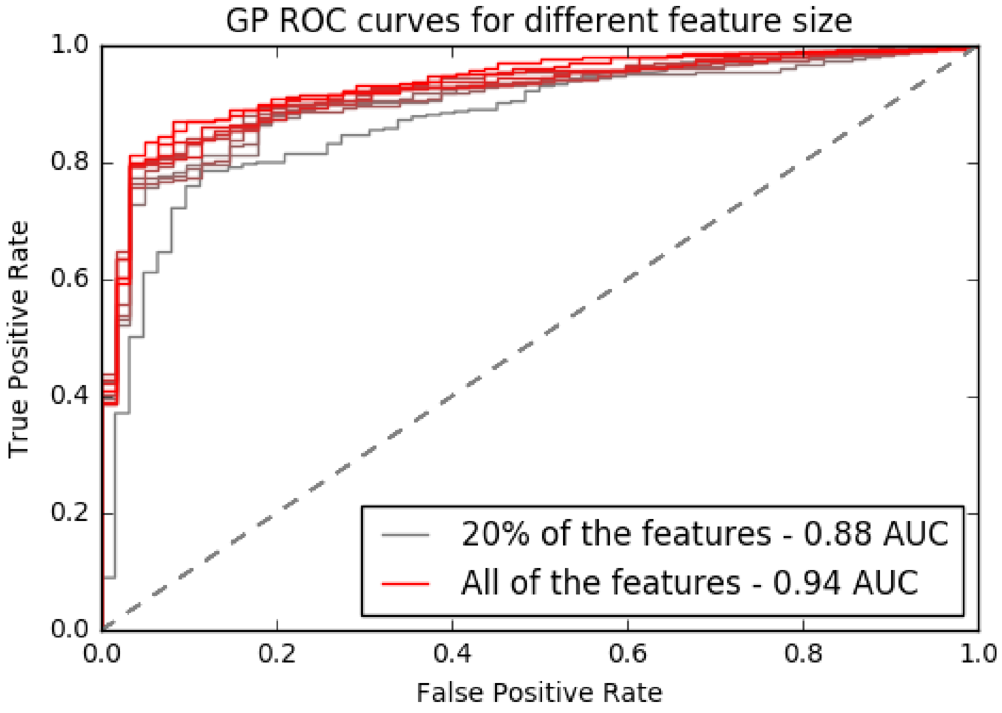


Figure 12: Gaussian process classifier performance across a range of feature numbers

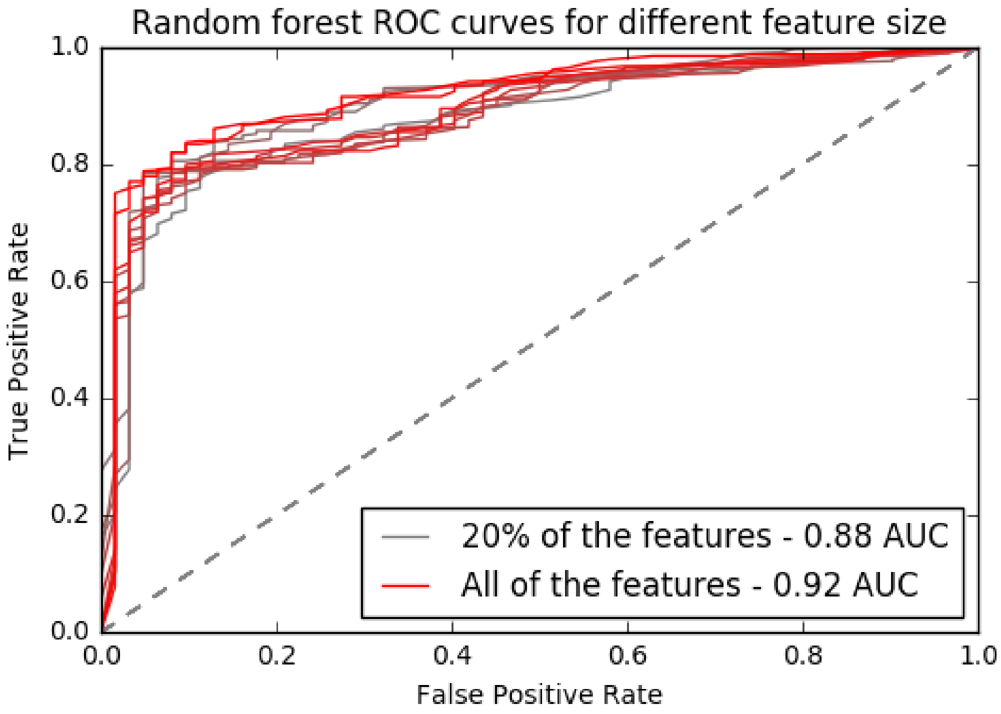


Figure 13: Random forest classifier performance across a range of feature numbers

K-means clustering and PCA are two non-classifier methods commonly used in understanding data sets. Models can be trained using existing data and new data points classified on the basis of their distance to other, already classified points. Using the distance to cluster approximation (see methods), a pseudo-ROC curve can be generated for PCA and k-means . This approach was used to test the usefulness of k-means(Figure 14) clustering and PCA (Figure 15) for examining the class of unseen data (e.g. cell lines). K-means performs poor at this task irrespective of the number of features it has access to. The PCA has rather good scores, but it should be kept in mind that the application of AUC to such analysis is unproven.

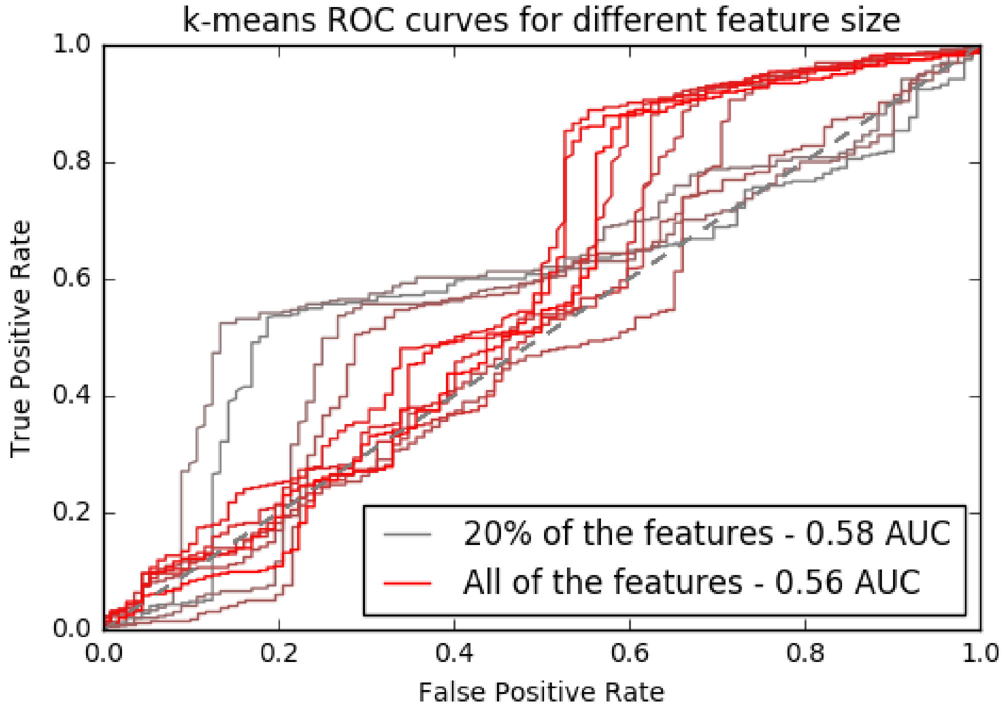


Figure 14: K-means clustering performance across a range of feature numbers based on a class pseudo-score adapted from the cluster center distance of the samples

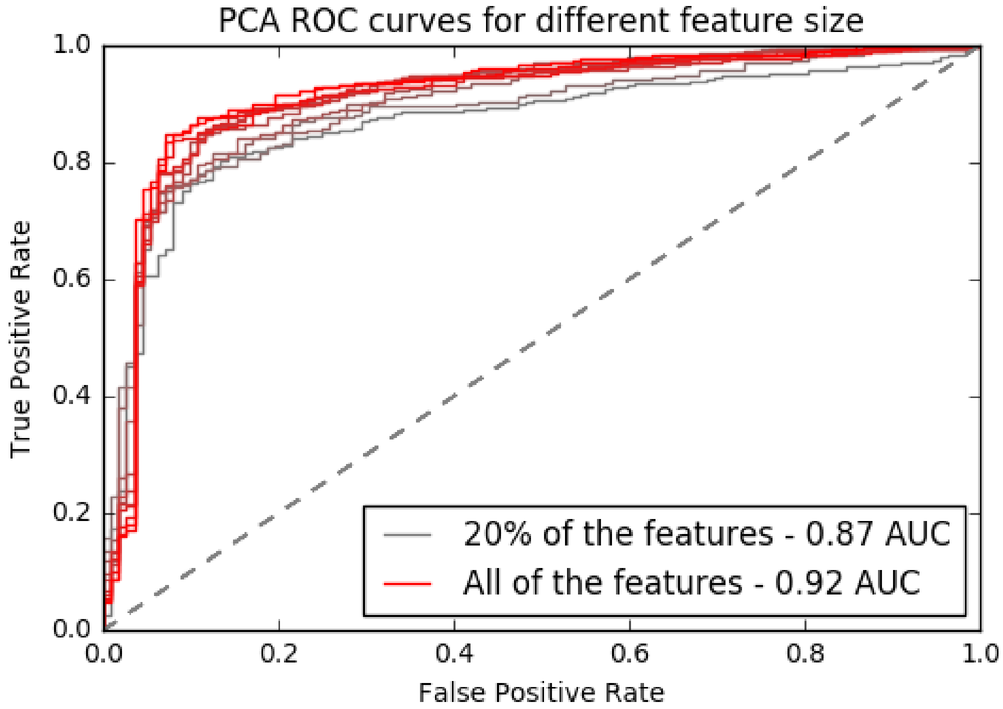


Figure 15: PCA performance across a range of feature numbers based on a class pseudo-score adapted from the cluster center distance of the samples

To assess the computational efficiency of each model, the time taken to train datasets with verying numbers of features was measured. Speed of training is an important resource to track because the higher speed enables faster adjustments and makes the model scale better with more data. Time across all features was significantly higher for PCA and k-means (Figure 16). SVM was both slower and slows down with feature number at a much steeper rate than the performance of other classifiers (Figure 17).

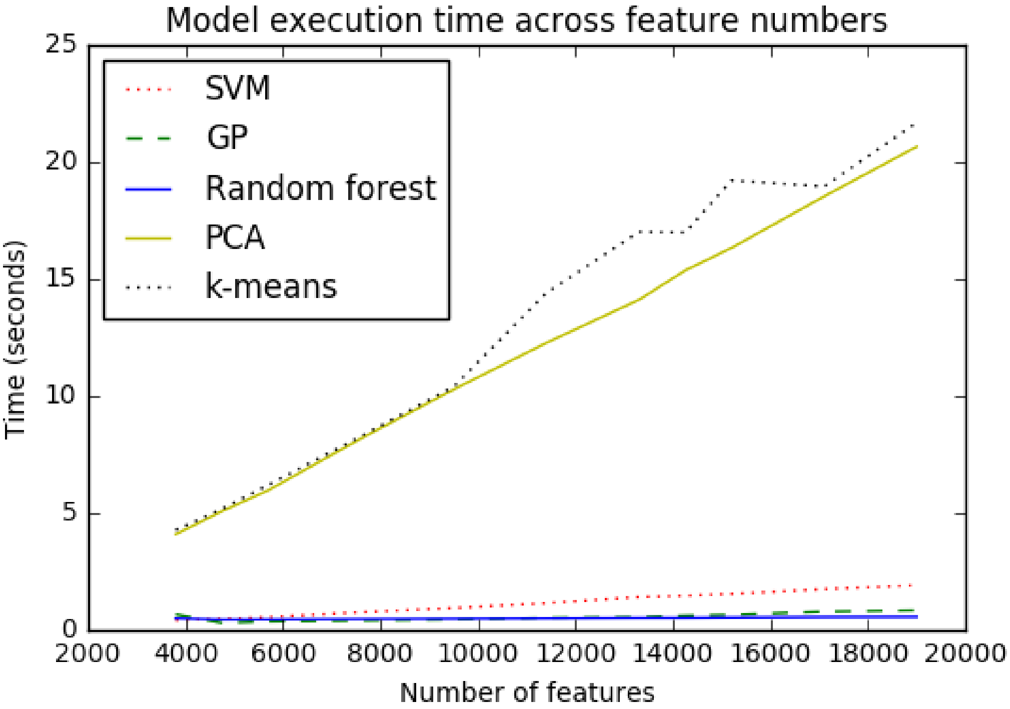


Figure 16: Execution times across a different number of features. Different lines represent the different methods.

As the performance degradation of using only 20% of the total training set was minimal, a dataset consisting of the 2000 most variant features, was used to measure F_1 scores and area under ROC curve (Table 3). The reason why only a minority of the features are as predictive as the total set likely lies in different feature importances. A dataset might contain a number of key predictive features and a lot of unimportant features that do not help increase prediction, potentially due to their high correlation with the other, predictive features or low correlation with the value that is being predicted. Although this dataset is the same in size as the smallest subset of the total feature dataset, it uses preselected features specifically to have the best performance. The total feature dataset measures performance, but it does not select features strategically. The results demonstrate the power of this approach as AUCs are barely lower than models trained on the full dataset.

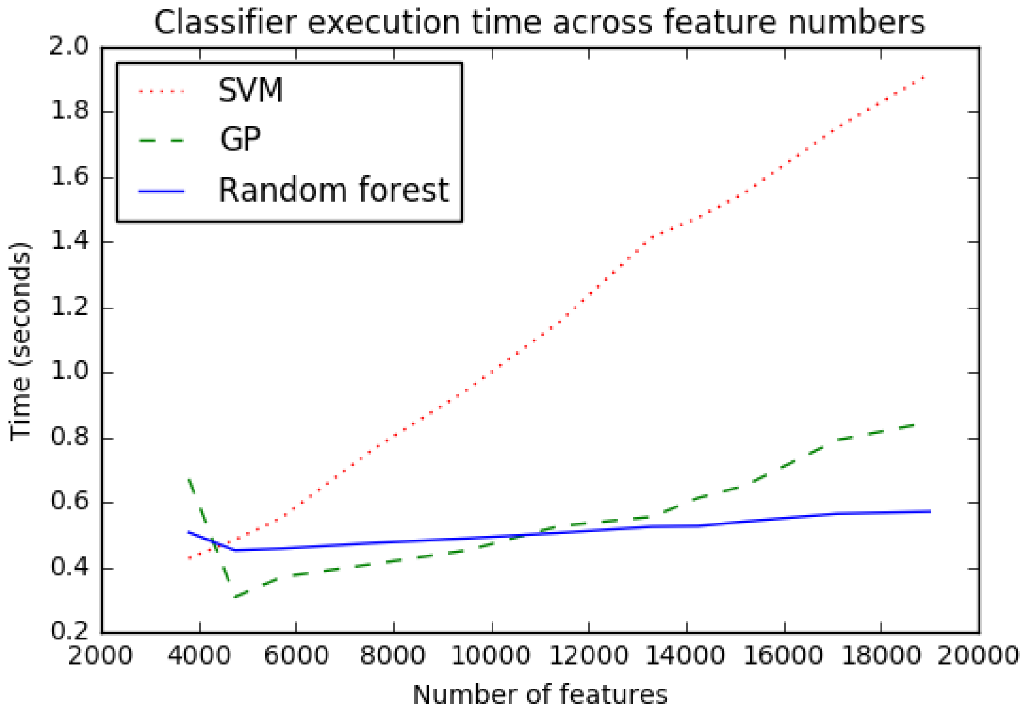


Figure 17: Execution times across a different number of features just for the classifier methods.

Method	F_1 score	Area under ROC curve
k-means	0.40	0.58
PCA	0.91	0.90
SVM	0.91	0.92
Gaussian process	0.89	0.92
Random forest	0.93	0.91

Table 3: Performances of different methods as measured by F_1 score and area under ROC curve.

2.3.1 Selection of a model

Although only the one data type from one cancer was used, as TCGA datatypes are of a fairly common format, results of these analyses might be generalised to other learning problems since the data properties are not uncommon - large amount of samples (100 to 1000) with an even larger amount of features (up to 20 000) with the features being continuous on a the same scale (while this might often not be the case, if the features are continuous, they can be easily normalised to the same scale).

The classification algorithms used all had increases in accuracy as the number of features increased, but not to such a large degree to warrant choosing one over the others for this dataset.

Since, in the feature number size variation case, the features were randomly chosen from the total feature pool, the AUC scores in the case of lower feature numbers could be painting a worse picture than would normally be the case. Some important features which could have a large impact on accuracy might have been left out. However, even if this is the case, it is unlikely to be a big difference since there are no big increases in the AUC at any points and the final case, which uses all the features, is not leaving any features out.

K-means performed very poorly, not too different from a random classifier. This is not surprising since k-means treats all the features the same. Being an unsupervised technique, it has no guiding label (like a class or a score) which would help it identify more important features. When there is a large number of features, many of which might not carry any class differentiating potential, it is essential for a learning algorithm to pick up on that in order to be effective.

Although PCA is another adapted non-classifying analysis, it has a surprisingly good performance, similar to the classifiers. Unlike k-means, PCA treats different features differently. Weights/rotations of principal components determine how influential each of the features are in determining the specific principal component. This coupled with distance to cluster centers being weighted by principal component variance ensures the more influential features are used. Additionally, PCA enables quick visualisation of multi-dimensional data and does not enforce a binary split of the samples.

An interesting thing to note is random forest having the highest F_1 score even though the AUC is slightly lower than SVM or GP classifiers. Because the differences are rather minor, this might not be of importance. Alternatively, it could mean the random forest has better specificity than the false positive rate since AUC score is dependent only on true positive (sensitivity) and false positive rates. Assuming this difference is not due to random variation, this could relevant information when choosing a classifier although not relevant for this case (differentiating between cancer subtypes).

Time-wise, the non-classifiers performed worse (5 to 20 seconds per a model), and importantly this deteriorated with increasing feature number. This is expected since k-means requires several iterations of each sample distance and cluster calculations while the PCA performs several transformations on all of the samples and their features. Of the classifiers, SVM performed worst, but still an order of magnitude faster than the non-classifiers (0.4 to 2 seconds). An important property is the rate of increase with the feature number. Gaussian process and random forest had minimal changes compared to the other algorithms (both staying well below 1 second). This makes them good choices for problems with a large number of features.

While being able to deal with a large number of features is generally advantageous, less features used allows for higher usability with other limiting methods. Some of the methodologies presented were implemented in a form of an interactive web application (Shiny R package) which adds additional constraints on the process - the data need to be handled by the application without using too much memory or time. As can be seen in the (Table 3), accuracy was still high (except for k-means) on the smaller, but with optimised feature selection, dataset.

An important property not analysed is the interpretability of the model. If it is a model used for learning or differentiating between subsets of the dataset in some other way, it can be valuable to easily understand which features are used to do so and how. Of the accurate methods, PCA and random forests have the most accessible features which are also biologically relatable. Weights in a principal component directly correspond to how important one gene is in the principal component. In the case of random forest, it can be seen which genes with which thresholds are used to determine whether a sample is one class or the other. The correlation matrix of a Gaussian process classifier can help understand how similarly classified samples are connected and support vector machine use samples closest to the class boundary which in itself can be quite informative, but is a bit more abstract and requires more analysis than random forest and PCA.

2.4 Single data-type classifiers for endometrial and breast cancer.

Both TCGA and CCLE include data on three types of genomic feature - gene expression data, simple nucleotide variant mutation data and gene level DNA copy number estimates, that might carry information for subtyping samples. To test which of these data types are most effective for classifying cancer subtypes random forest classifiers were trained for each data type separately (copy number, mutation, expression), for both cancer types used.

The training set used to train the classifiers was composed of samples equally split between the two classes (114 patient samples for endometrial, 65 for breast cancer). Each of the samples consisted of 2000 genomic features (expression, copy number aberration or mutation depending on the data type) and a binary class variable representing the cancer subtype information (endometrioid/serous for endometrial cancer, triple-negative/receptor-positive for breast cancer).

After the training phase (with training set having 100 samples for UCEC and 80 for BRCA, both of which have equal class/subtype representation), the classifiers estimated the class of the samples the testing set. The test set samples (424 for UCEC, 830 for BRCA) were composed of only the genomic features.

Their accuracy on the test set can be seen in the upper part of Figure 18 and Figure 19. Their performance shows which of the data types carries the most information for differentiating classes in the random forest classifier. The copy number classifier performs best of all the endometrial cancer classifiers based on a single data type (F_1 = 0.89, AUC=0.92).

In the case of breast cancer the expression-based classifier is the most predictive (F_1 = 0.83, AUC= 0.91) (subsection 2.4).

This is in line with what is known about the genomic background of the subtype classes. Copy number data is the most differentiating data type for endometrioid - serous endometrial cancer (Getz et al., 2013) and gene expression data is the most differentiating for receptor-based breast cancer subtypes (Koboldt et al., 2012).

2.4.1 Gaussian process validation

In order to confirm the accuracy on the patient test set, Gaussian process models were built with the same training and test datasets. As expected, based on their overall accuracy and performance in the method comparison chapter, the results were similar with copy number based breast cancer classifier performing significantly better and expression classifier for endometrial cancer performing much worse. (Table 5).

Data type	UCEC	BRCA
Expression	(0.88, 0.91)	(0.83, 0.92)
Copy number	(0.83, 0.925)	(0.75,0.77)
Mutation	(0.89, 0.91)	(0.72, 0.79)

Table 4: Accuracy scores of the random forest classifiers for endometrial (UCEC) and breast (BRCA) cancers. Data type denotes the type of data used to build a classifier. The first number in the pair is the F 1 score of that specific data type - cancer type classifier, while the second one is the AUC score.

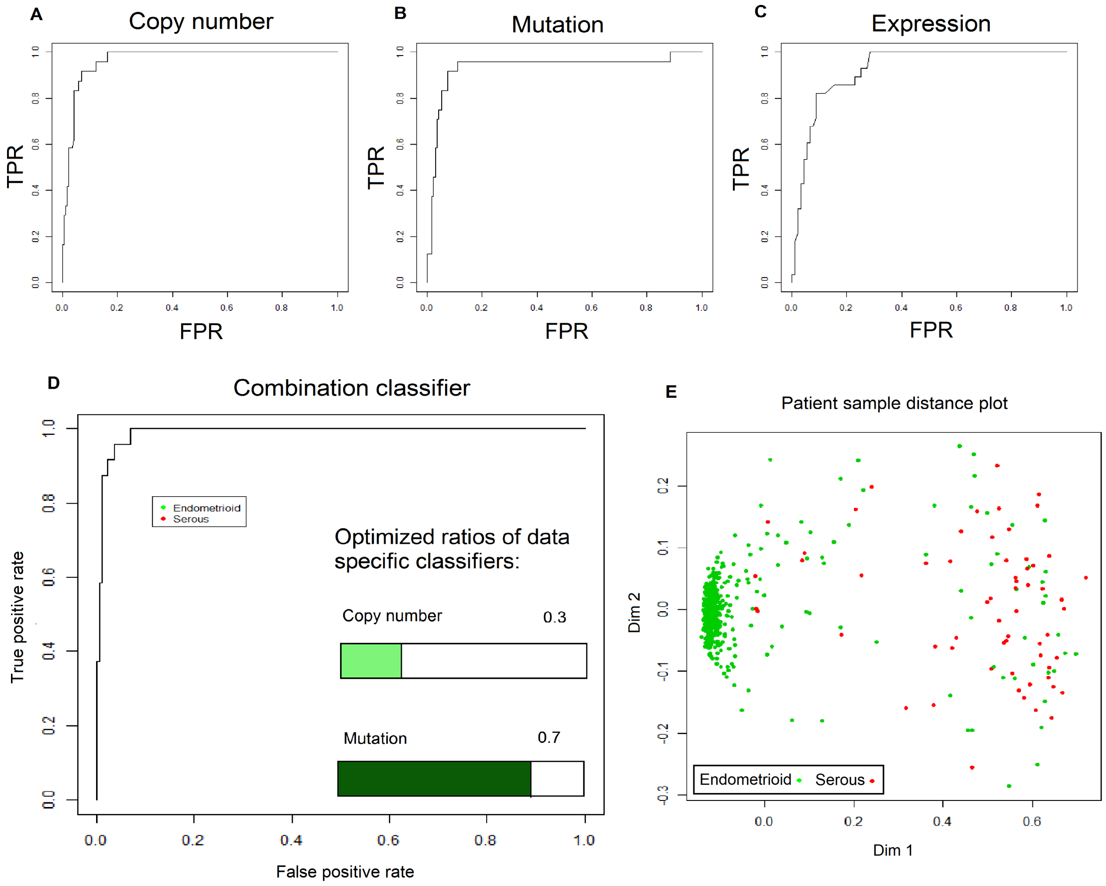


Figure 18: Accuracy (ROC curves) of different data type classifiers for endometrial cancer. Values generated by the classifier for the test set samples and known classes were used to construct ROC curves for each of them. A-c) the single data type classifiers, D) ROC curve and associated weights for the combination classifier, E) visual representation of class separation through and MDS plot.

Data type	UCEC	BRCA
Copy number	0.925	0.901
Expression	0.524	0.935
Mutation	0.520	0.917

Table 5: Accuracy of the Gaussian process classifiers

2.5 Exploration of classifier combinations

Different classification tasks requre different dataypes. Thus a general purpose classifier must consider all data types in order to select the relevant variables. In addition since not all biologically significant alterations can be seen in all the data types (Getz et al., 2013), integration of multiple data types provides a way of expanding scope of the model to potentially relevant predictors across multiple data types. If significant drivers exist across different data types, a predictive model tapping into different data types will have access to more of them and could have higher predictive accuracy.

Data integration of the available different genomic data (mutation, gene expression, copy number aberration) was carried out in order to maximise the potential predictive power of the models.

There were two distinct approaches to integrating the different data types.

Combining different data type features into a single model

One way of data integration is to combine the data types by joining all the features together and rescaling them if appropriate. Rescaling is necessary because significantly larger feature values for one of the data types result in higher variance for those features, assuming all else is equal. This can give the illusion of more important features in the data type containing larger values. However, this is applied only to expression and copy number data because of the binary nature of the mutation data.

An important advantage of this approach is being able to see which features are more important in the classifier score across data types. This is not only interesting because of the inter-type feature importance comparison, but it also a measurement of which data types are more influential with respect to the chosen subtype. Knowing which data type is more predictive for a certain subtype prediction can inform future data generation. E.g. if drug response of an endrometrial cancer is better predicted by gene expression data than other data types then transcriptome sequencing would preferable to genome sequencing when investigating the specific drug.

Additionally, since in this approach the classifier picks out the specific features which seem predictive without regard to the source data type, potentially significant features of a less generally predictive data type are still recognised.

Combining single data type models through weighting their score by model accuracy

A faster way of using all the data types for the scoring process is combining the result of individual classifiers evaluation into a single score. Although this approach does not offer additional insight into the genomic features, it is faster and easier to implement.

A single classifier assigns probabilities of belonging to a class to every sample. These probabilities sum up to 1, so the resulting combination score also needs to sum up to 1. This can be achieved by assigning a weight, between 0 and 1, to each classifier and constraining the weights to sum to 1. For example, 0.2 to copy number and 0.8 to mutation-based classifier, or 0.6 and 0.4. The weight factors are applied to each class probability of the classifiers. The resulting weighted class probabilities for each class are then summed across all the classifiers (Equation 17). This weight assignment approach allows the contribution of a classifier to vary while making sure the resulting class probabilities sum to 1 (Equation 18). The higher the weight factor of a classifier, the more similar the combined score is to that classifier.

$$
\begin{aligned}
&P(A)_{C_1} = p\\
&P(B)_{C_1} = 1 - p\\
\\
&P(A)_{C_2} = q\\
&P(B)_{C_2} = 1 - q\\
\\
&P(A)_{combination} = w \times P(A)_{C_1} + (1 - w) \times P(A)_{C_2}\\
&P(B)_{combination} = w \times P(B)_{C_1} + (1 - w) \times P(B)_{C_2}
\end{aligned}
\tag{17}
$$

$$
\begin{aligned}
&P(A)_{combination} + P(B)_{combination}\\
&= w \times P(A)_{C_1} + (1 - w) \times P(A)_{C_2} + w \times P(B)_{C_1} + (1 - w) \times P(B)_{C_2}\\
&= w \times (P(A)_{C_1} + P(B)_{C_1}) + (1 - w) \times (P(A)_{C_2} + P(B)_{C_2})\\
&= w + (1 - w)\\
&= 1
\end{aligned}
\tag{18}
$$

The weight factor, w, is determined by iteration through a range of factors between 0 and 1 and selecting the one that gives the least number of misclassifications on the test set. While there might be a problem of overfitting (tailoring a model too much to the test set so it ends up performing poorly), there are factors that make that less likely. Firstly, it is a combination of three models using different data types that do not overfit themselves. If a weights combination favours one model, it will overfit as much as that model, but if it favours a combination of models, it effectively favours a combination of different features making less likely to overfit. Additionally, the random forest model itself is well suited to avoid overfitting due to its bootstrapping mechanism (explained earlier). The final accuracy scores being not too different from the non-weighted combination strategy suggests a similar level of overfitting as the other model.

While it is difficult to compare the importance of a feature from one data type with a feature form a different data type, the overall effect of a data type on accurate classification in the context of the chosen class can still be estimated through weights assigned to the two classifiers.

Endometrial cancer combination classifier

An issue in both the combination approaches was the reduction in samples resulting from incomplete coverage across all the data types. Initial sample size of copy number data was 539, but counting the samples which also have the mutation data reduced the number of samples to 242. After intersection with the expression data, the sample size was down to 7, so a complete data type combination, with a meaningful sample size, was not possible for endometrial cancer.

Using only two datatypes at a time, the mutation - copy number weighted combination classifier performed best(Figure 18). The mutation-based classifier was assigned $w = 0.7$ after iteration, and the copy number one $w = 0.3$. Although copy-number classifier is by itself the most accurate classifier, the mutation classifier is more important when combining classifiers through weighting. The accuracy of this kind of combination classifier (through weighting of existing single data type classifiers) was similar to the previous one ($F_1 = 0.92$, AUC = 0.98).

Using the data type combination approach with UCEC samples, a mutation - copy number combination classifier achieved an F_1 score of 0.88 and AUC of 0.92.

Breast cancer combination classifier

Unlike endometrial cancer, there are a sufficient number of samples with all three data types available in the cancer dataset to make training a three datatype model feasible (524 samples from 1080 have all three datatypes available). Since there were

Classifier type	F_1 score	AUC
Expression	0.828	0.927
Copy number	0.75	0.77
Mutation	0.72	0.79
Combination through scaling	0.877	0.941
Combination through merging data	0.877	0.929

Table 6: Performance of different breast cancer classifier types as measured by F_1 score and AUC. The first three rows are classifiers based only on one data type, the latter two combine the datatypes. Combining through scaling uses scores of the basic classifiers and joins them based on the weights obtained through maximising for the least misclassifications on the test set. Combining through merging data creates a new classifier by combining all the existing datatypes into one dataset.

three data types to combine, and the sum of the weights has to be equal to one, a single weight is determined by the other two (Equation 19).

$$\begin{aligned} w_1 + w_2 + w_3 &= 1 \\ w_3 &= 1 - w_1 - w_2 \end{aligned} \tag{19}$$

Thus values for only two weights must be found. To search over two ranges, a grid search over two ranges, one for each weights, was implemented, using a step of 0.01 in each case. For every combination, the third weight was imputed and then all three were used to assign a score to the test set. The final weights were the ones which has the lowest number of misclassified samples on the test set.

These weights were $w_{CN} = 0.22$, $w_{EXP} = 0.36$ and $w_{MUT} = 0.42$.(Table 6). Classifier made by combining the single data types with those weights had F_1 and AUC scores of 0.87 and 0.94 respectively.

The data joining combination approach yields similar, highly accurate, results (0.87 and 0.92 for F_1 and AUC respectively) (Table 6, Figure 19).

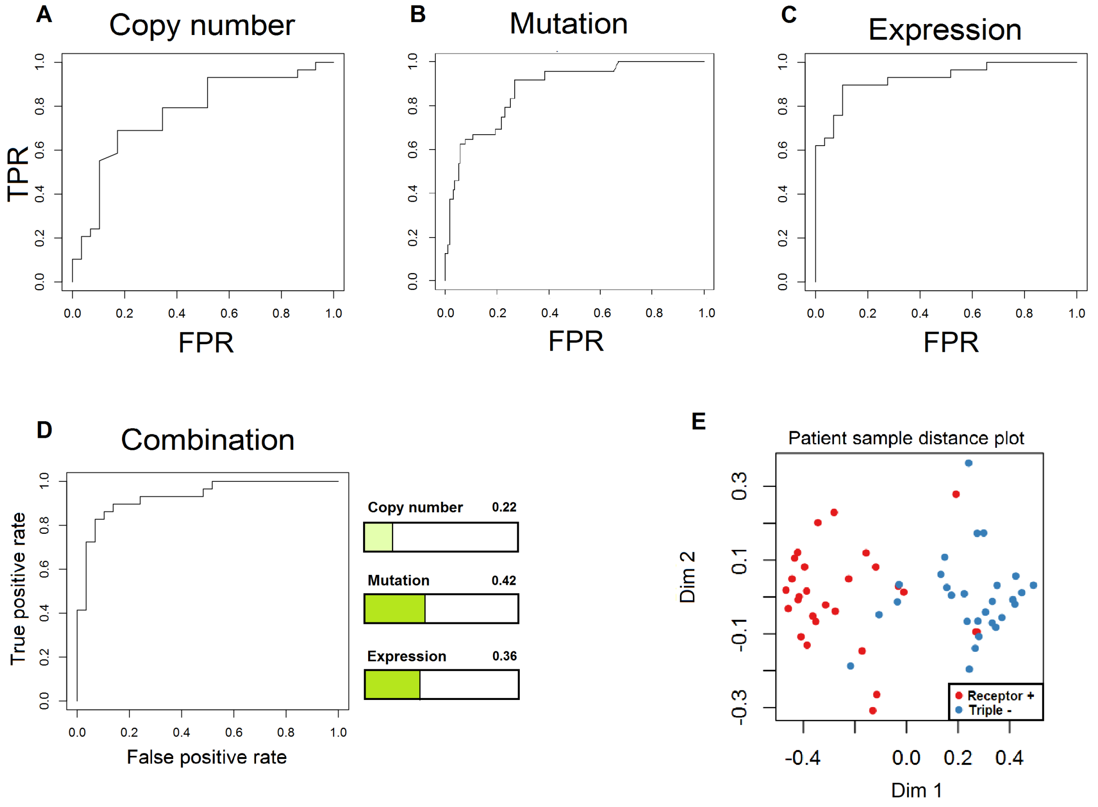


Figure 19: Breast cancer classifier performances as depicted by ROC curves. A-C) performances of single data type classifiers. D) performance of the weighted combination classifier with the weights used next to it. The weights used are the ones found to lead to most accurate (as measured by number of misclassified samples) combination. E) MDS plot of the data joining combination.

2.6 Discussion

Through comparison of different methods for the data set that will be extensively using, I set out to determine what the best approach would be.

Resource-wise, classifiers performed much better than k-means and PCA. Since the size of the dataset in question can be a limiting factor in analyses, this was an important confirmation of my expectations.

The best models in this analysis all perform exceptionally well (F_1 of 0.92 and 0.87 for weighted combination classifiers for endometrial and breast cancer respectively). While there are slight differences, I decided to focus on random forest because of the most convenient way of extracting feature information from the model. Random forest automatically calculates the importance (mean decrease accuracy and Gini impurity) of features it uses, making it much easier to analyse the underlying causes of the classification strategy the model learned. Correlation matrices and kernel mappings of Gaussian process and support vector machines are slightly less straight-forward.

To my knowledge, the adapted use of PCA carried out has not seen much use. Further research into its applications and comparisons with other methods would be desirable before it is selected in place of other methods which perform similarly, but have been extensively applied and documented. Especially in cases such as this one, where standard methods do not perform worse. I would suggest using this approach alongside other, well-established, methods.

Single data type classifiers showed varying degrees of accuracy, with the best ones having excellent performance. The variation in the performance by data type can be attributed to role of those data types in the chosen subtypes of the cancer types analysed. Combination classifiers peformed, unsurprisingly, better than the single datatype ones. Assuming availability of more than one data type, combining them in some way appears to be preferable to using just one. The only potential issue being increase in resource requirements.

While the joined features combination classifier has the advantage of being able to assign importance to features from different data types, thus helping the investigation of causative elements driving the classification decisions, the weight scaled combination classifier is much more convenient because it is faster and requires less memory. Given they both scored extremely well (with the scaled combination one actually slightly outperforming the joined features model), I have settled on using the weight scaled combination model for the most part–examining the biological significance being the case where the other one would be preferable.

3 Subtype classification of CCLE cell lines

3.1 Introduction

In the previous chapter, commonly used machine learning methods, such as random forests and Gaussian processes, were shown to be highly predictive of cancer subtype in patient cancer samples. Models were developed to predict endometrioid/serous subtype in endometrial cancer and receptor-positive/triple-negative subtype in breast cancer.

If a method is predictive of subtype in patient cancer samples, it is a good candidate for prediction of the same subtype in cancer cell lines. Using the previously described classification models allows for systematic evaluation of cell lines. Cancer cell line subtype score can help guide research decisions and reduce costs of using a suboptimal cell line. Determining the score via a model trained on patient cancer samples makes sure the subtype score is derived from real patient data, thus placing the cell lines into patient cancer subtype context and making the score clinically relevant.

Since cell lines can suffer from misalignment with the original cancer and misidentification (Capes-Davis et al., 2010), such a score can help verify their place on the chosen cancer subtype spectrum, assign a subtype to a new or underutilised cell line, and provide evidence when it is suspected that a cell line is incorrectly assigned to a subtype.

This chapter deals with cancer cell line analysis from several different perspectives. Firstly, the literature has been surveyed for information on subtypes of endometrial and breast cancer cell lines. This information is later used to examine how it matches with classifier assigned scores. Non-classifying analyses (PCA and clustering) served to provide a quick overview of the cell line relationships in the feature landscape. Cancer cell line scores based on patient samples were generated using the methodology described in the previous chapter. Because of its performance and interpretability, random forest was chosen as the main learning algorithm for this chapter. Distance from patients, calculated with respect to the classifier's feature importances, lets us know not just how well a cell line matches a subtype, but also how similar it is to the patient samples in general.

3.2 Cell line reviews

In order to better understand the computational cancer-based classification of cell lines, a literature search was conducted to determine the reported classifications. This enables the computational scoring of cell lines to be compared to real-world properties, provided literature descriptions of cell lines are accurate enough.

Most commonly, literature involving cancer cell line research is not focused on cancer subtypes, but rather on cancer reactions to specific drugs and the related biological properties such presence of certain mutations, pathway differences, etc. (Gozgit et al., 2012). This is unsurprising since these kinds of studies utilise new technologies and therapeutics thus adding novelty to the field.

3.2.1 Endometrial cell line type assignments

In the case of endometrial cancer, two classification systems are commonly used: endometrioid/serous and type 1/type 2 (Getz et al., 2013). Serous and endometrioid types are defined by their histopathology - e.g. a typical serous carcinoma tissue might have papillary structures, rounded cells and cilia (Soliman and Lu, 2013). In contrast, the type 1/2 system is based more on predicted clinical outcome. Type 1, as opposed to type 2, cancers are less aggressive, have better prognosis and are linked to excess oestrogen (ACS medical and editorial team, 2019). Despite being different classifications, the overlap between them is significant: all serous are type 2 and all low-grade endometrioid are type 1. However, Type 2 also includes high-grade endometrioid cancer and other some other types of endometrial cancer (clear cell carcinomas, carcinosarcomas) (Kandoth, 2013; Saso et al., 2011).

The relationship between the two subtype classifications in the TCGA endometrial cancer dataset can be seen in Figure 20.

The cancer cell lines covered here are endometrial cancer cell lines found in the CCLE (Barretina et al., 2012). There are quite a few other commonly used endometrial cancer cell lines which are not contained within CCLE: ARK1, ARK2, HEC-155/180, SPEC-2, EN-11, EN-1.

There have been cases of endometrial misidentification in the past. In 2012 (Korch et al., 2012) identified that the cultures of the line labelled as ECC-1 had been identical to another endometrial cancer cell lines (Ishikawa) since at least 2006 (Mo et al., 2006). However, (Weigelt et al., 2013) still used it in their endometrial cancer panel. Other misidentified cell lines include HES being HeLa and the entire hTERT-EEC line being MCF-7 breast cancer (Korch et al., 2012).

There are several cases where the literature contains confusing reports on the type of a cell line. A large panel of cancer cell lines are described as endometrioid by one study (Weigelt et al., 2013), some of which are noted differently in other reports (Kuramoto and Nishida, 2003) (HEC-6 - adenoacanthoma), (Barretina et al., 2012) (RL95-2 - mixed adenosquamous carcinoma) and (Nagamani and Stuart, 1998) (RL95-2 - adenosquamous carcinoma).

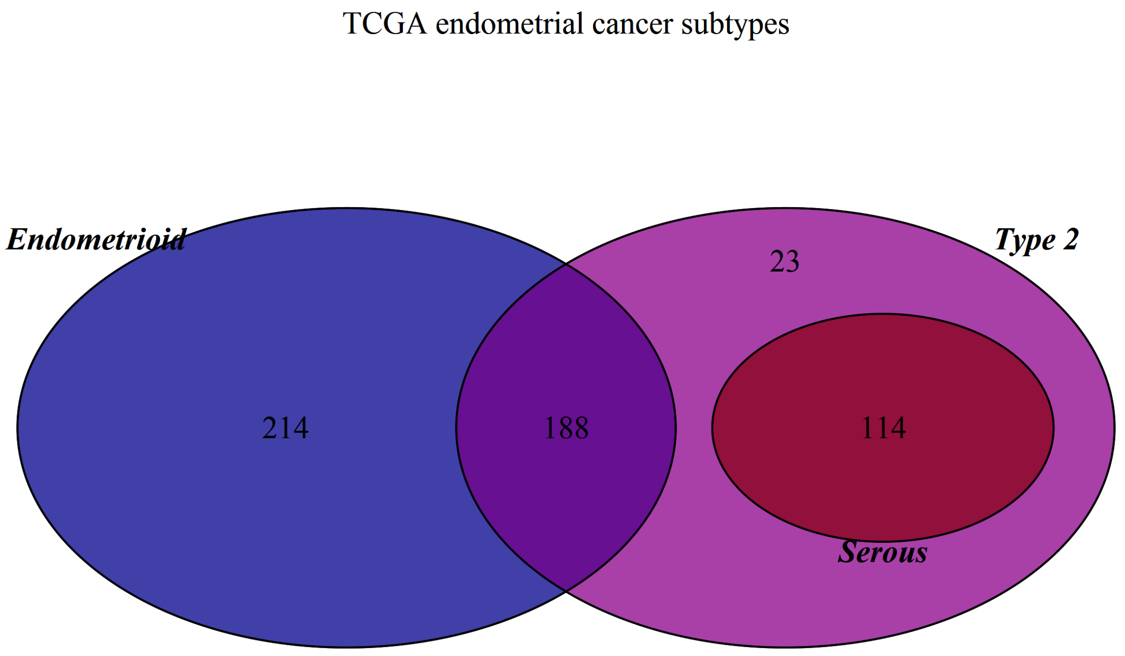


Figure 20: Endometrial cancer subtypes relationship. All samples are from the TCGA's endometrial cancer (UCEC) dataset. All endometrioid that are not type 2 are type 1. Type 2 is made up of high-grade endometrioid, serous and other, non-endometrioid, subtypes (clear cell, carcinosarcoma, mixed types).

According to (Zhou et al., 2014), MFE-280 and KLE are type 1, but all other reports suggest type 2 (Kalogera et al., 2017).

COLO-684 might be of ovarian origin as the COLO-704 cell line was isolated from the same tumour (European Collection of Authenticated Cell Cultures, 2018) and subsequently identified as ovarian (Barretina et al., 2012).

CCLE endometrial cell lines have no literature identifying any of them as serous. Thus, where I wish to compare my subtypes to the previous literature I will use classes based on type 1 and type 2 because of the significant overlap between serous and type 2 subtypes.

A brief summary of the literature and the associated types for each of the cell lines is presented in Table 7 and a summarised version used for contrasting with the analyses results in Table 8. These assignments are not to be taken as definitive and are only used as an approximate guide. Assigned class is determined based on literature descriptions of the specific cell line. If a cell line is noted as Type 1 or endometrioid without being described as high-grade, it is considered to be one class (represented by 0 on the classifier score scale); high-grade endometrioid, poorly-differentiated, serous and clear-cell cell lines are considered a different class (1 on the classifier score scale). This class also subsumes other types of endometrial cancer. In the case of contrasting descriptions being present in the literature, the more common or more recent one was used.

	Histology	Type	Comments	Assigned
AN3-CA	metastatic	Type 2	Assigned type 2 (Korch et al. 2012, Theisen et al. (2014)), described as poorly differentiated (characteristic of type 2) (Kuramoto and Nishida 2003) and as a metastatic undifferentiated cell line (Nagamani and Stuart 1998)	1
COLO-684		NA	Suspected an ovarian (Barretina et al. 2012)	NA
ECC-1		Type 1	Type 1, identical to Ishikawa cell lines and not matching the tumour from which it was reportedly derived (Korch et al. 2012).	0
EFE-184	metastatic	NA	Metastatic (Artimo et al. 2012) endometrioid (Weigelt et al. 2013) .	NA
EN		Type 1	Present only in one publication so far (Isaka et al. 2003), as an invasive endometrioid adenocarcinoma.	0
ESS-1	stromal	Type 2	A stromal cancer (Gunawan et al. 2003, Barretina et al. 2012).	1
HEC-1	endometrioid	Type 2	HEC1-A and HEC1-B cell lines were isolated from the same source (Kuramoto and Nishida 2003). CCLE considers them identical due their high shared SNP identity and both being subclones of HEC-1. Both cell lines are described as type 2 (Korch et al. 2012, Xiong et al. (2015) although also as endometrioid by (Weigelt et al. 2013).	1
HEC-50	endometrioid	Type 2	Mentioned by various authors as type 2 (Korch et al. 2012), endometrioid (Weigelt et al. 2013) and high grade endometrioid (Kuramoto and Nishida 2003). HEC50co, a related cell line, is described as type 2 (Albitar et al. 2007)	1
HEC-6	adenoacanthoma	NA	HEC-6 is suspected to be an esophageal (Cai et al. 1993)	NA
HEC108	high-grade unknown	Type 2	(Kuramoto and Nishida 2003)	1
Other HEC lines	low-grade	Type 1	HEC151, HEC251, HEC59 and HEC265 have all been described as low-grade adenocarcinomas (Kuramoto and Nishida 2003)	0
IshikawaH02ER	endometrioid	NA	Ishikawa cells were originally type 1 (Korch et al. 2012, Albitar et al. 2007), but due to high proliferation since the isolation, there has been a variety of subclones (Nishida 2002)	NA
JHUEM	endometrioid	Type 1	JHUEM endometrial cell lines were isolated from the malignancies in Japanese patients. All of the CCLE JHUEM cell lines are recorded as endometrioid adenocarcinoma by RIKEN BioResource Research Center with JHUEM-1 additionally being described as grade 2 (Riken Bioresource Center 2018)	0
KLE	endometrioid	conflicting	Noted as type 2 (Korch et al. 2012, Theisen et al. (2014), Xiong et al. (2015)), poorly differentiated (grade 3) (Kalogera et al. 2017, Nagamani and Stuart (1998)), endometrioid (Weigelt et al. 2013) and type 1 in (Zhou et al. 2014).	1
MFE-280	endometrioid	Type 2	Described variously as type 2 (Hackenberg et al. 1997), endometrioid (Weigelt et al. 2013) and type 1 (Zhou et al. 2014)	1
MFE-296	mixed type	Type 1	Mixed type (Hackenberg et al. 1997)	0
MFE-319	endometrioid	Type 1	Type 1 (Hackenberg et al. 1997)	0
RL-95-2	adenosquamous	Type 1	Noted as type 1 (Weigelt et al. 2013), adenosquamous (Kuramoto et al. 2003, Barretina et al. 2012) and well differentiated adenosquamous (Nagamani et al. 1998).	0
SNG-M	metastatic	NA	It has been suspected SNG-M is a metastatic (Ishiwata et al. 1977), a metastatic lymph node (Mori et al. 1994) or not a cancer at all (Fukuda et al. 1999).	NA
SNU-685 -1077	carcinosarcoma	Type 2	These cell lines have been isolated from the same Mixed Mullerian (a carcinosarcoma) tumour patient (Yuan et al. 1997).	1
TEN	clear cell	Type 2	Mentioned as type 2 (Fushiki et al. 1997) and as clear cell carcinoma (Barretina et al. 2012).	1

Table 7: Endometrial cell lines' subtypes

Table 8: Endometrial cancer cell line subtypes according to the literature. Summary of Table 7.

Type 1	Type 2	Ambiguous
ECC-1	AN3-CA	COLO-684
EN	ESS-1	EFE-184
HEC-151	HEC-1	HEC-6
HEC-251	HEC-50	ISHIKAWAH02ER
HEC-59	HEC-108	SNG-M
HEC-265	KLE	
JHUEM-1	MFE-280	
JHUEM-2	SNU-1077	
JHUEM-3	SNU-685	
JHUEM-7	TEN	
MFE-296		
MFE-319		
RL-95-2		

3.2.2 PubMed citations

Cell line citations were obtained from the number of search results from PubMed database (Table 9). R packages RISmed (Kovalchik, 2017) and easyPubMed (Fantini, 2017) were used to query the database for each cell line, the results of which were then processed in R. The R script made to execute this (and for the breast cancer cell lines as well) can be found at https://github.com/vuzun/papers/blob/master/src/PubMeding_BRCA_CLs.R.

Some cell lines have alternative spellings (usually with or without a hyphen) that normally would not show in searches, so the final number is the union of all the results for different spellings. All of the searches had “AND cancer” added to the cell line name so to reduce the number of results which had a word matching the cell line’s name, but did not include the actual cancer cell line. For the majority of the cell lines, the content of the literature search was not checked for whether or not it uses the actual cell line or if the setting is relevant to preclinical cancer research. Since each count included fetching article data from PubMed, number of articles was capped at 1000 to reduce the execution time and avoid needless load on PubMed.

Because of their ambiguous names for the purpose of a search query, a few cell lines had slightly altered queries by adding the *endometrial* term. TEN’s query was *“TEN cells” endometrial* and the same alteration was made for EN, ISHIKAWA and COLO-684. Ishikawa cell line numbers might not be reflective of the cancer cell line obtained from the CCLE (IshikawaHeraklio02ER). There is a variety of Ishikawa cell lines which are often not fully described within the Ishikawa cell line family.

	number of citations
AN3-CA	119
COLO-684	1
EFE-184	3
EN	1
ESS-1	31
HEC1-A	183
HEC1-B	92
HEC-108	11
HEC-151	1
HEC-251	3
HEC-265	2
HEC-50B	12
HEC-6	4
HEC-59	26
ISHIKAWA	1000+
JHUEM-1	0
JHUEM-2	2
JHUEM-3	0
JHUEM-7	0
KLE	225
MFE-280	11
MFE-296	16
MFE-319	3
RL95-2	177
SNU-1077	1
SNU-685	1
SNG-M	26
TEN	1

Table 9: Number of PubMed articles for endometrial cancer cell lines.

3.2.3 Breast cell line type assignments

There are several ways of separating breast cancer samples into clinically relevant subgroups. One commonly used way is by the status of the hormone receptors - HER2/ERBB2, oestrogen and progesterone. The combination of low expression of all three receptors is associated with the poorest prognosis and difficulty of treatment (Grigoriadis et al., 2012). While this is a useful typology, there are more detailed characterisations of those subtypes as well (Lehmann et al., 2011). However, the more detailed subtyping leads to less representatives per a subtype and more requires subtype classes to model.

There have been some reports of misidentification in breast cancer cell lines, several of which are present in the CCLE. The KPL-1 cell line turned out to be a different breast cell line - MCF-7 (both in CCLE) (Capes-Davis et al., 2010). MDA-MB-435 was identified as melanoma cell line (Ellison et al., 2002), but was still used in research as a breast cancer cell line for several years afterwards (Rae et al., 2007).

The cancer cell lines covered here are breast cancer cell lines found in the Cancer Cell Line Encyclopaedia (Barretina et al., 2012). There are many other breast cancer cell lines which are not contained within CCLE: HCC712, IBEP2, LY2, MDAMB134, MDAMB175, ZR75B, BSMZ, IBEP1, IBEP3, MDAMB330, ZR7527, 21MT1, 21MT2, 21NT, 21PT, HCC1008, HH315, HH375, KPL-4, OCUB-F, SKBR5, SUM190PT, SUM225CWN, EMG3, HCC3153, HMT3522, KPL-3C, MA11, MDAMB435, MFM223, SUM185PE, SUM229PE, HCC1739, SKBR7, SUM102PT, SUM1315M02, SUM149PT and SUM159PT.

(Dai et al., 2017) pools multiple breast cancer cell lines studies and, among other analyses, lists the consensus of the receptors' status in the cell lines (Table 10). However, there are several cell lines which are present in the CCLE breast cohort, but are not specifically subtyped by (Dai et al., 2017).

HCC1419 and HCC1500 appear to be ambiguous in the literature with few articles available. Receptor statuses of HCC1419 are ER+/PR-/HER2+ in (Riaz et al., 2013) as opposed to ER-/PR-/HER2+ in (Kao et al., 2009). For the purpose of the classification utilised here it does not matter much because the both cases are receptor-positive. HCC1500 is considered triple-negative by (Neve et al., 2006), but as receptor positive in (Riaz et al., 2013). Hs606T is noted as luminal (ER or PR receptor positive) and Hs739T as basal (triple-negative) (Jiang et al., 2016). HMEL, Hs274T, Hs281T and Hs343T have no citations and no easily accessible information regarding their receptor status.

CCLE also contains some lines of uncertain provenance: HMC-1-8 is suspected of

Table 10: Breast cancer cell line subtypes with respect to receptor status. Receptor positive have 1 or more present receptors (progesterone, estrogen or HER2) while triple-negative have no present receptors. Ambiguous cell lines either have conflicting or no information present regarding their receptor status.

Receptor positive	Triple-negative	Ambiguous
AU-565	BT-20	EVSAT
BT-474	BT-549	HCC-1500
BT-483	CAL-120	HCC-1419
CAMA-1	CAL-148	HMC-1-8
EFM-19	CAL-51	HMEL
EFM-19-2A	CAL-851	Hs274T
HCC-1428	DU-4475	Hs281T
HCC-1569	HCC-1143	Hs343T
HCC-1954	HCC-1187	Hs606T
HCC-202	HCC-1395	Hs739T
HCC-2218	HCC-1599	Hs742T
JIMT-1	HCC-1806	
KPL-1	HCC-1937	
MCF-7	HCC-2157	
MDA-MB-134-VI	HCC-38	
MDA-MB-175-VII	HCC-70	
MDA-MB-361	HDQ-P1	
MDA-MB-415	Hs578T	
MDA-MB-453	MDA-MB-157	
SK-BR-3	MDA-MB-231	
T-47D	MDA-MB-436	
UACC-812	MDA-MB-468	
UACC-893		
YMB-1		
ZR-75-1		
ZR-75-30		

being a lung cell line by CCLE as it shares high SNP identity with HLC-1 (a lung cell line) (Barretina et al., 2012). CCLE also notes SK-BR-3 and AU-565 being from isolated from the same person and MCF-7 and KPL-1 sharing high SNP identity (Barretina et al., 2012).

Interestingly, there is a much larger population of breast cancer triple-negative cell lines in the CCLE, compared to patient samples in TCGA (57 triple-negative samples and 467 receptor-positive ones).

Table 11: Number of PubMed articles for breast cancer cell lines

Cell-line	Number of articles	Cell-line	Number of articles
AU565	48	HDQ-P1	3
BT-20	620	HMC-1-8	4
BT-474	1211	HMEL	0
BT-483	22	Hs 274-T	0
BT-549	388	Hs 281-T	0
CAL-120	1	Hs 343-T	0
CAL-148	0	Hs 578-T	562
CAL-51	57	Hs 606-T	974
CAL-85-1	0	Hs 739-T	974
CAMA-1	50	Hs 742-T	1
DU4475	30	JIMT-1	102
EFM-192A	3	KPL-1	30
EFM-19	19	MCF7	1000+
HCC1143	20	MDA-MB-134-VI	6
HCC1187	7	MDA-MB-157	70
HCC1395	11	MDA-MB-175-VII	5
HCC1419	5	MDA-MB-231	1000+
HCC1428	12	MDA-MB-361	183
HCC1500	15	MDA-MB-415	14
HCC1569	13	MDA-MB-436	165
HCC1599	3	MDA-MB-453	488
HCC1806	70	MDA-MB-468	1000+
HCC1937	185	SK-BR-3	1000+
HCC1954	85	T-47D	1000+
HCC202	1	UACC-812	24
HCC2157	0	UACC-893	12
HCC2218	3	ZR-75-1	795
HCC38	38	ZR-75-30	82
HCC70	40		

3.2.4 PubMed citations

I obtained breast cancer cell line citation numbers (Table 11) using the same procedure as described in the endometrial cell lines section. Cell line HMEL had 6 results none of which used the cell line in question, but often referred to hydroxymelatonin as HMEL. Consequently, those results were not taken into account.

3.3 Non-classifier analyses

3.3.1 Principal component analysis

PCA provides a way of summarising high-dimensional data through building new features (data dimensions) which capture more variance than the original ones. While PCA, unlike random forests or Gaussian processes, does not learn any pattern in the data or assign scores, it can still be valuable to learn how variance is distributed in the data and whether that shows any interesting patterns. Interestingly, in the last chapter a pseudo-classifier based on distance to cluster centroid in PCA space performed surprisingly well. This suggests that the information contained in the PCA is valuable for investigating cell subtypes. For both cancer types, variation was split across many principal components (Figure 21), indicating a high number of independent underlying drivers. This complexity is expected in a multi-dimensional, difficult-to-treat problem such as endometrial or breast cancer.

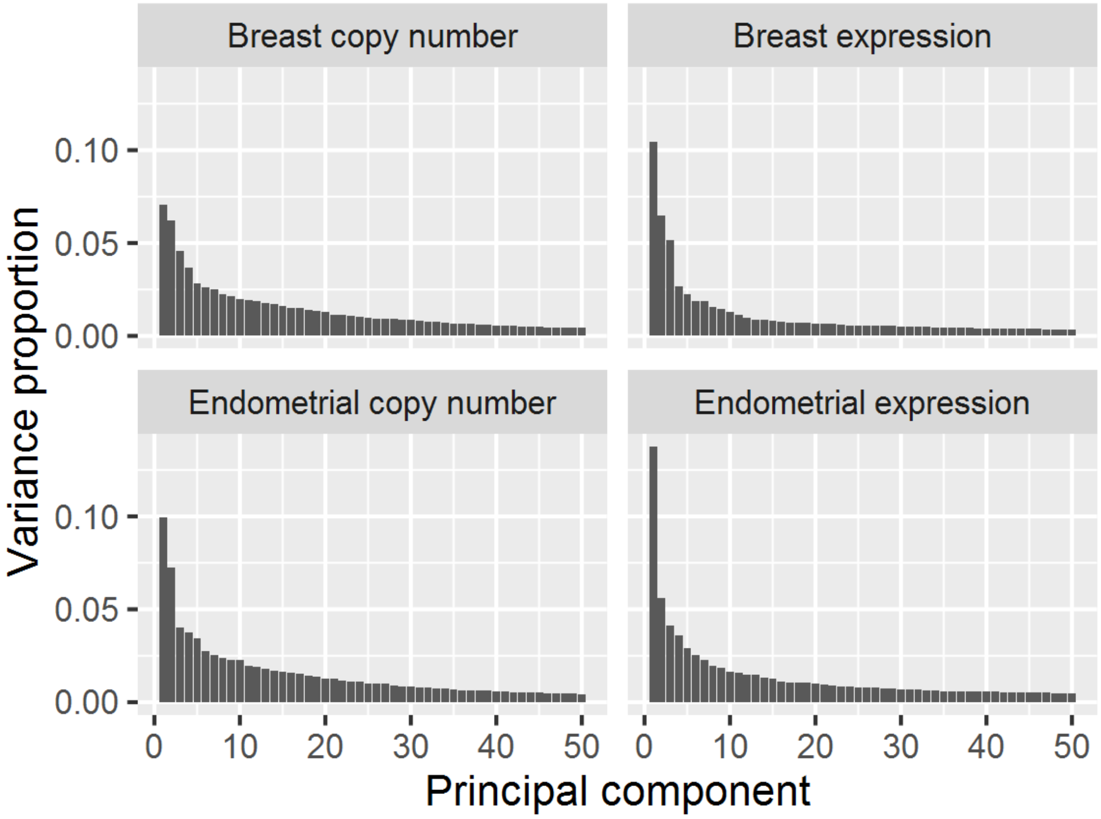


Figure 21: First 50 principal components and the proportion of variance they capture in the corresponding data type - cancer type PCA.

Endometrial cancer

Principal component analyses of transcripts per million and GISTIC2 copy number scores of patient samples were performed using base R's *prcomp* function. The resulting PCA rotations were used to transform the cell line data as well. The result was 539 principal components per each patient sample and cell line for copy number and 175 for expression. The first two principal components are shown for copy number (Figure 22) and expression (Figure 23), split by cancer subtype for patient samples (endometrioid or serous).

PCA analyses captured similar amount of variance for both data types used (10% and 7% for the first two PCs for copy number, 13% and 6% for expression data). In the copy number PCA (Figure 22), endometrial cancer subtypes have very different distributions - endometrioid samples have a dense cluster in a small part of the PC space (upper right corner of the plot) and a large, but sparser, fallout around it; serous samples have the same density throughout the simlarly large space they cover. One explanation for this might be that there is more genomic diversity in the serous subtype, likely due it being more agressive and, thus, more prone to copy number changes (Nakayama et al., 2012).

Even though the overlap of the subtypes is significant, a rough division between the subtypes can be made.

The expression PCA shows similar degree of separability between the patient sample subtypes, although without the distribution differences seen in the copy number PCA - neither subtype has a dense cluster of its samples.

In the copy number PCA, cell lines are placed throughout the plot with several cell lines having a fair distance from the other cell lines and the endometrioid cluster.

Cell lines in the expression PCA are placed fairly compactly, suggesting that expression might be less differentiating data type for cell lines and possibly the subtypes as well.

According to the TCGA-based study of endometrial cancer subtypes (Getz et al., 2013), overall copy number differences contribute more to clinically relevant subtype differentiation than differences in expression.

Breast cancer

Breast cancer PCA generated 582 principal components for the copy number data (PC_1 - 7%, PC_2 - 6%) and 588 for the expression data (PC_1 - 10%, PC_2 - 6%).

Compared to the copy number PCA (Figure 24), the expression one (Figure 25) has much clearer separation of patient samples by subtype, but also more clumped together cell lines. Ideally, I would want the cell lines spread out in order to differentiate between them more easily.

However, in the expression PCA, cell lines closest to the triple-negative cluster

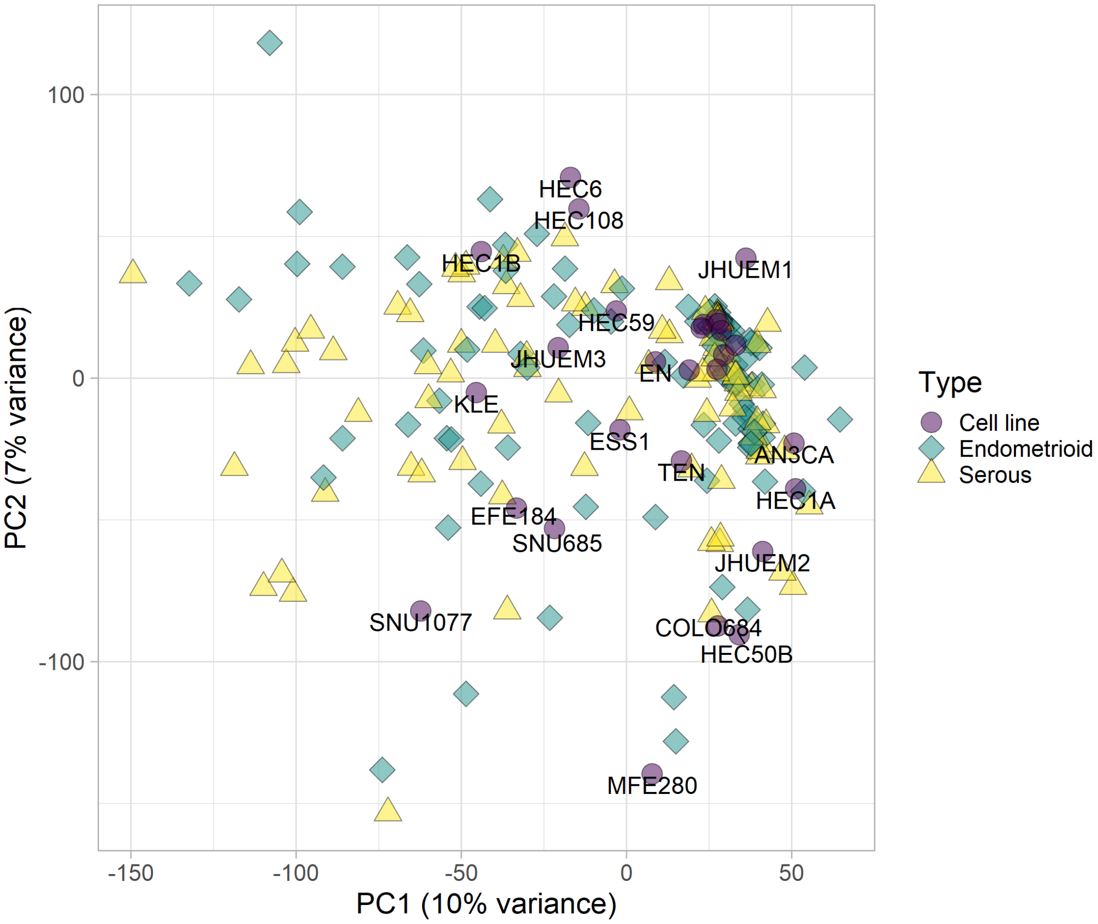


Figure 22: Copy number PCA of endometrial cancer. PC_1 captured 10% of the variance and PC_2 7%. Total number of principal components was 539. Colour/shape correspond to different cancer subtype and cancer cell lines. Labels next to some of the points on the plot are names of the cell lines represented by the nearby circle.

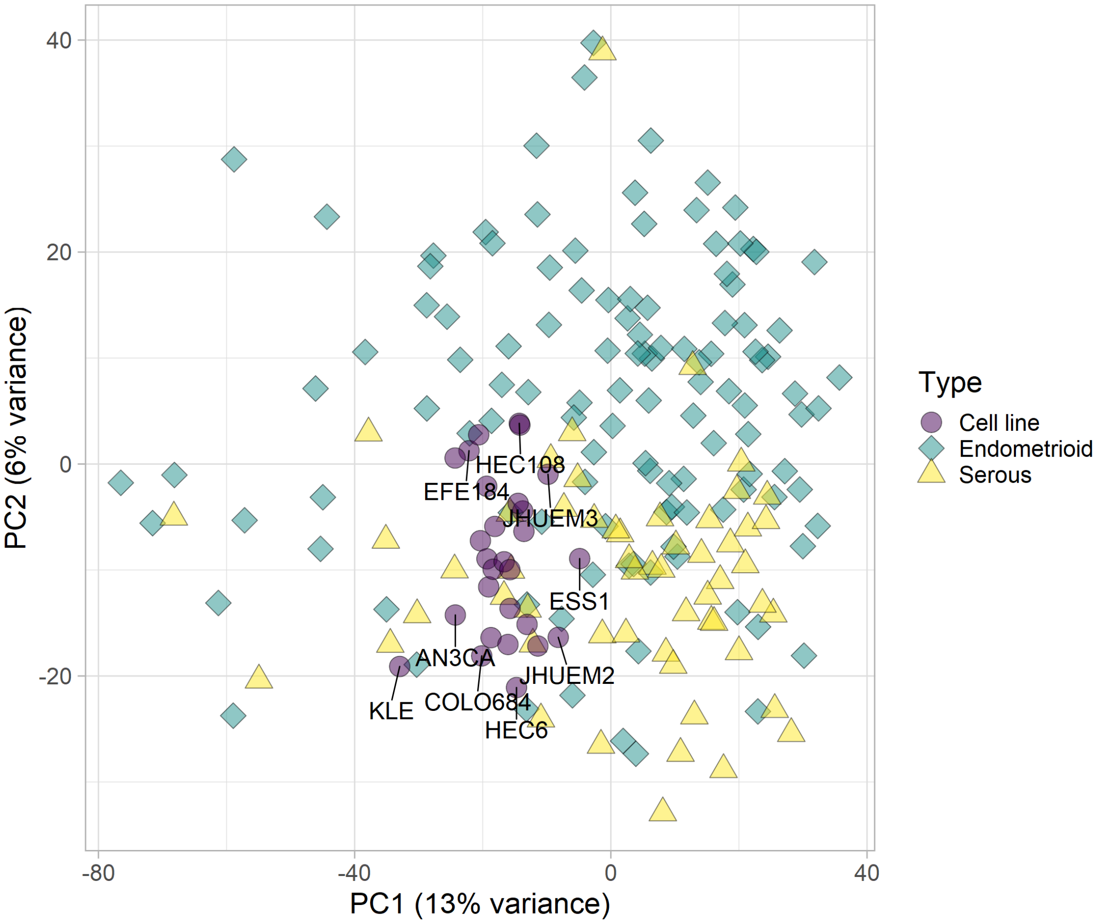


Figure 23: Gene expression PCA of endometrial cancer. First two principal components (out of 175) captured 10% of the variance.

(HCC70, HCC1954, Hs578T, HCC1143, BT549) are triple-negative themselves with the exception of HCC1569 (HER2 positive, ER/PR negative cell line), so the smaller variation might not be as impactful as it might seem at first.

Expression based PCA being clearer at separating subtypes makes sense considering the breast cancer subtype in question is receptor-based.

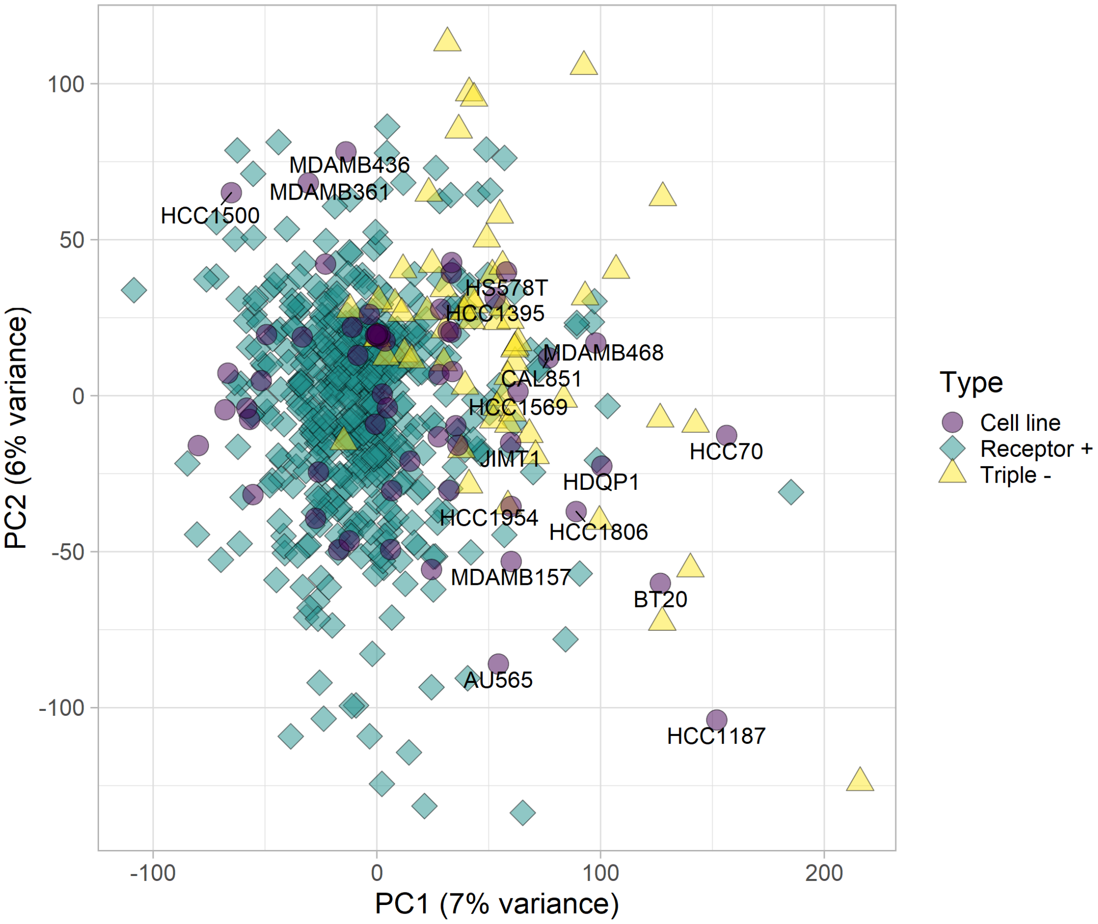


Figure 24: Copy number PCA of breast cancer.

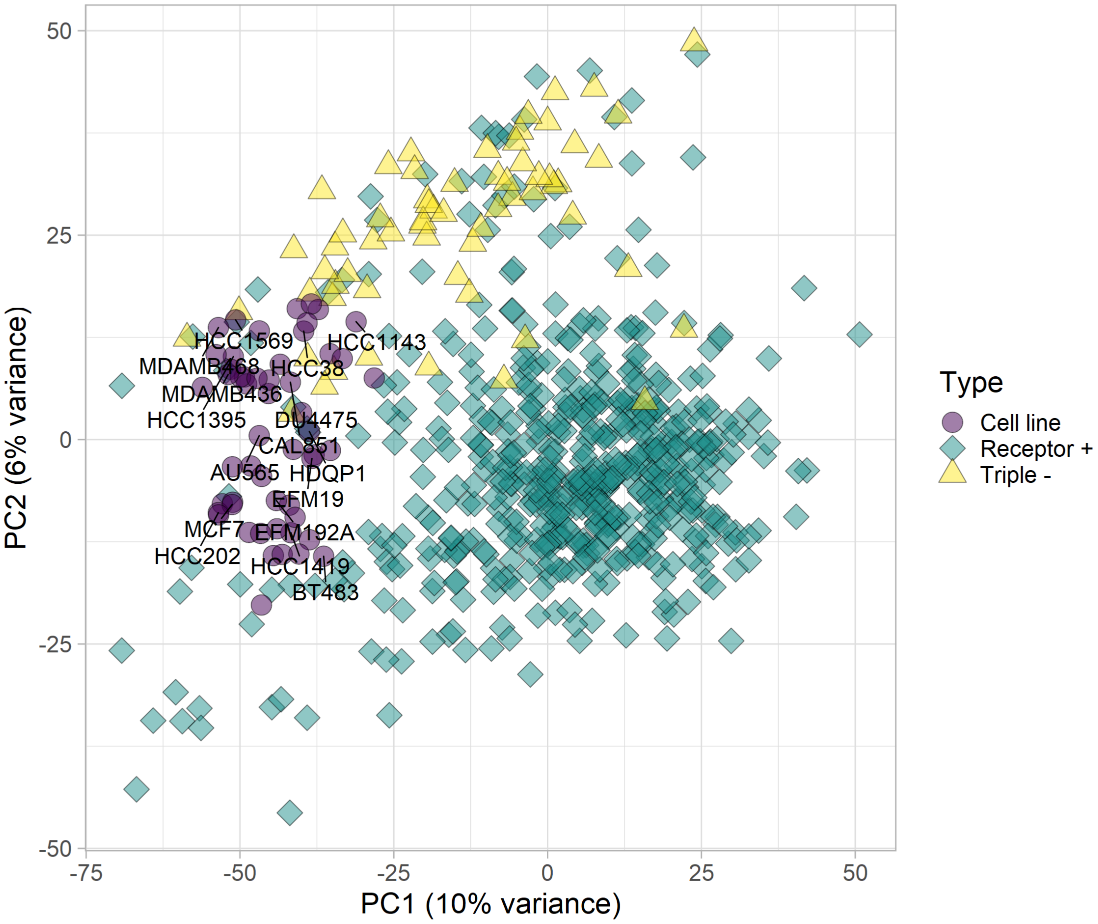


Figure 25: Gene expression PCA of breast cancer.

3.3.2 Clustering

Hierarchical clustering with significance scores was run (Figure 26, Figure 27) using copy number scores from GISTIC2 analysis of CCLE segmented copy number profiles and TPM data. Clusters with a number next to them have $p < 0.05$, indicating strength of the clustering for the specific subcluster. P-values were calculated by *pvclust* function from the R package of the same name through multiscale bootstrap resamplings (Suzuki and Shimodaira, 2015). Clustering used a correlation distance metric and UPGMA agglomeration algorithm.

Although there are some sensible patterns in the clustering results (one significant cluster includes HEC1-A and HEC1-B, two other type 2 cell lines - MFE-280 and ESS-1, unassigned Ishikawa cell line and type 1 HEC-251 which has only 3 citations in the literature), inbetween distance of the SNU pair and overall lack of pattern with respect to type and histology make the results of this type of clustering questionable for endometrial cancer.

In the case of breast cancer, for the copy number clustering, known or suspected identical cell line pairs (AU-565 and SK-BR-3, MCF-7 and KPL-1, ZR-75-1 and YMB-1) (Barretina et al., 2012) cluster close to each other, validating this cluster breakdown. Out of the other 7 significant groups, 5 have cell lines with the same subtypes, according to the literature.

Similar patterns can be seen in the expression-based clustering with HMEL cell line being such an outlier, all the other cell lines cluster together.

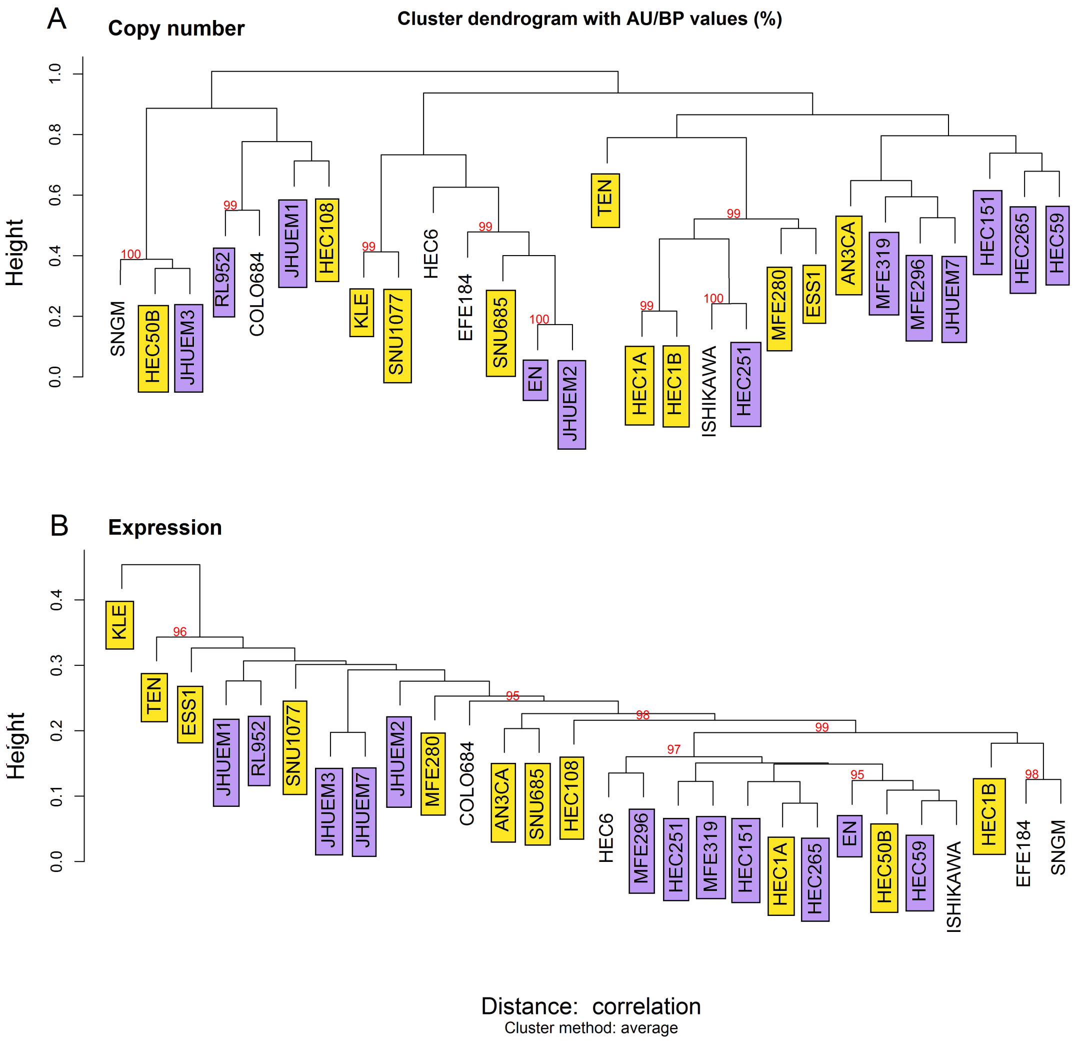


Figure 26: Clustering of endometrial cancer cell lines by their copy number scores (A) and expression values (B). Cell lines are coloured according to their subtype, as assigned in Table 7. Cell lines without a clear known subtype are not coloured. Clusters with high significance (approximate unbiased p-value >= 95%) are marked with a number. Produced with the pvclust R package .

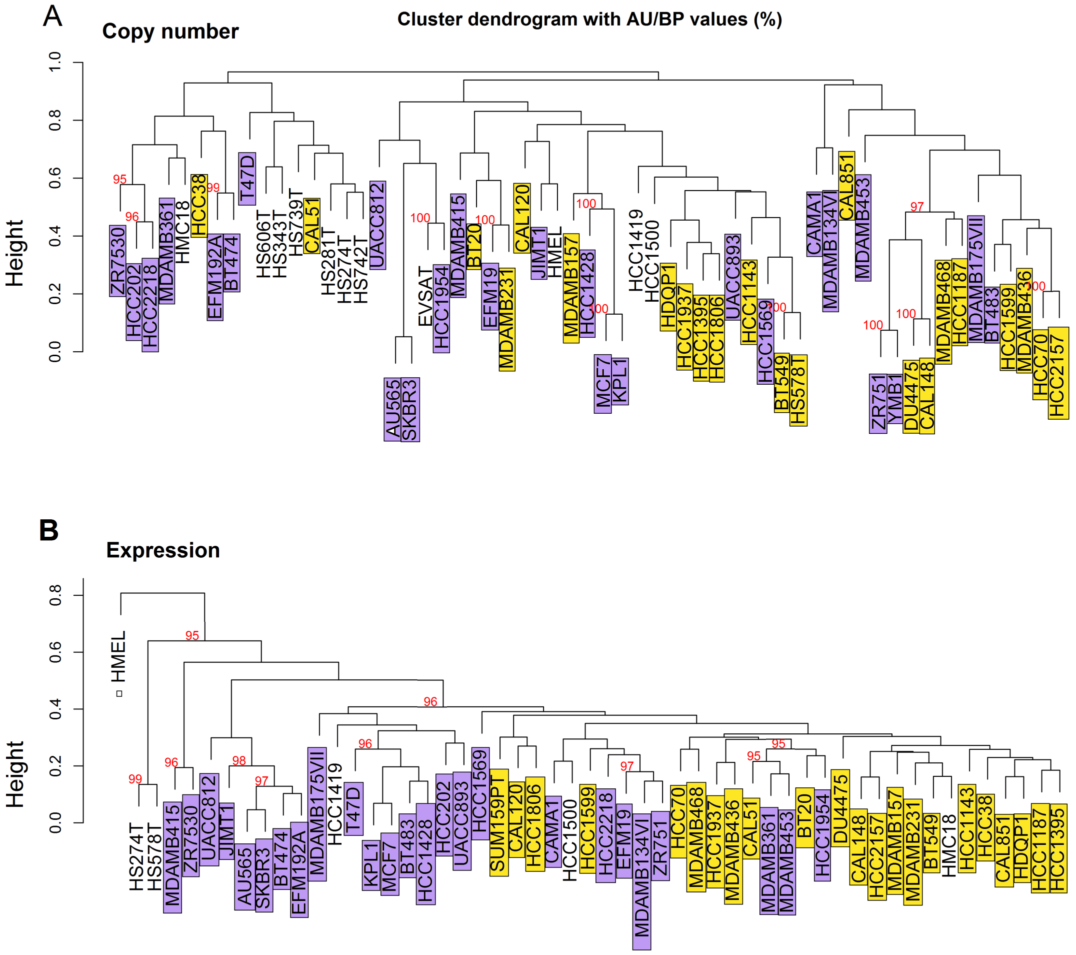


Figure 27: Clustering of breast cancer cell lines by their copy number scores (A) and expression values (B). Cell lines are coloured according to their subtype, as assigned in Table 7. Cell lines without a clear known subtype are not coloured. Clusters with high significance (approximate unbiased p-value >= 95%) are marked with a number.

3.4 Application of created machine learning models to cell lines

The classifiers developed in the previous chapter were then applied to cancer cell line data. The data used for the cell lines matched the patient sample data - the same features were used and the same data type. Literature assignments were sourced from published journal articles and medical sources (see sections on Endometrial cell line type assignments and Breast cell line type assignments)

Results of single data type classifiers and the weighted combination classifier when applied to cell lines were overlayed with the literature assignment of cell line subtype as colour (Figure 28, Figure 29).

For each cell line, we can see how likely it is to be one or the other subtype and with respect to a data type. For example, EFE-184 cell line does not have a clearly define subtype in the literature, but in the endometrial cell line results (Figure 28) it can be seen that it resembles a serous subtype more than most other cell lines, especially when it comes to copy number examination only.

Endometrial cancer

There is a general tendency of type 1 cell lines to score more endometrioid and type 2 to score more towards the serous side. Score ranges vary significantly between the classifiers used. Mutation-based classifier has only 1 cell line scoring (KLE) above 0.5 (meaning it's more likely to be of serous subtype), expression classifier has majority of the scores close to 0.5 and the copy number classifier is the one with the highes amount of variability, not suprising cosidering copy number has been shown as most predictive for endometrial cancer subtypes (Getz et al., 2013). The combination classifier, being the best weighted combination of copy number and mutation classifiers, has 5 cell lines scoring as more serous (KLE, EFE-184, ESS-11, HEC-50B, SNU-1077), none of which are type 1 in the literature. SNU-685 (type 2) scores as endometrioid in the mutation and combination, but serous with copy number and expression. Only consistently endometrioid-scoring type-2 cell lines are AN3-CA and HEC-108.

Of type-1 cell lines, all of them have mutation and combination classifier scoring as endometrioid. Some score as serous in the expression classifier, but are fairly close to 0.5 as are the type 2 cell lines in that classifier.

The cell lines which lacked the literature subtype or were difficult to place in the type 1/2 are fairly spread out. SNG-M and EN have consistently low (more endometrioid) scores. HEC-6, Ishikawa and COLO-684 have serous copy number scoring, but endometrioid otherwise. EFE-184 is the second most serous cell line and consistently

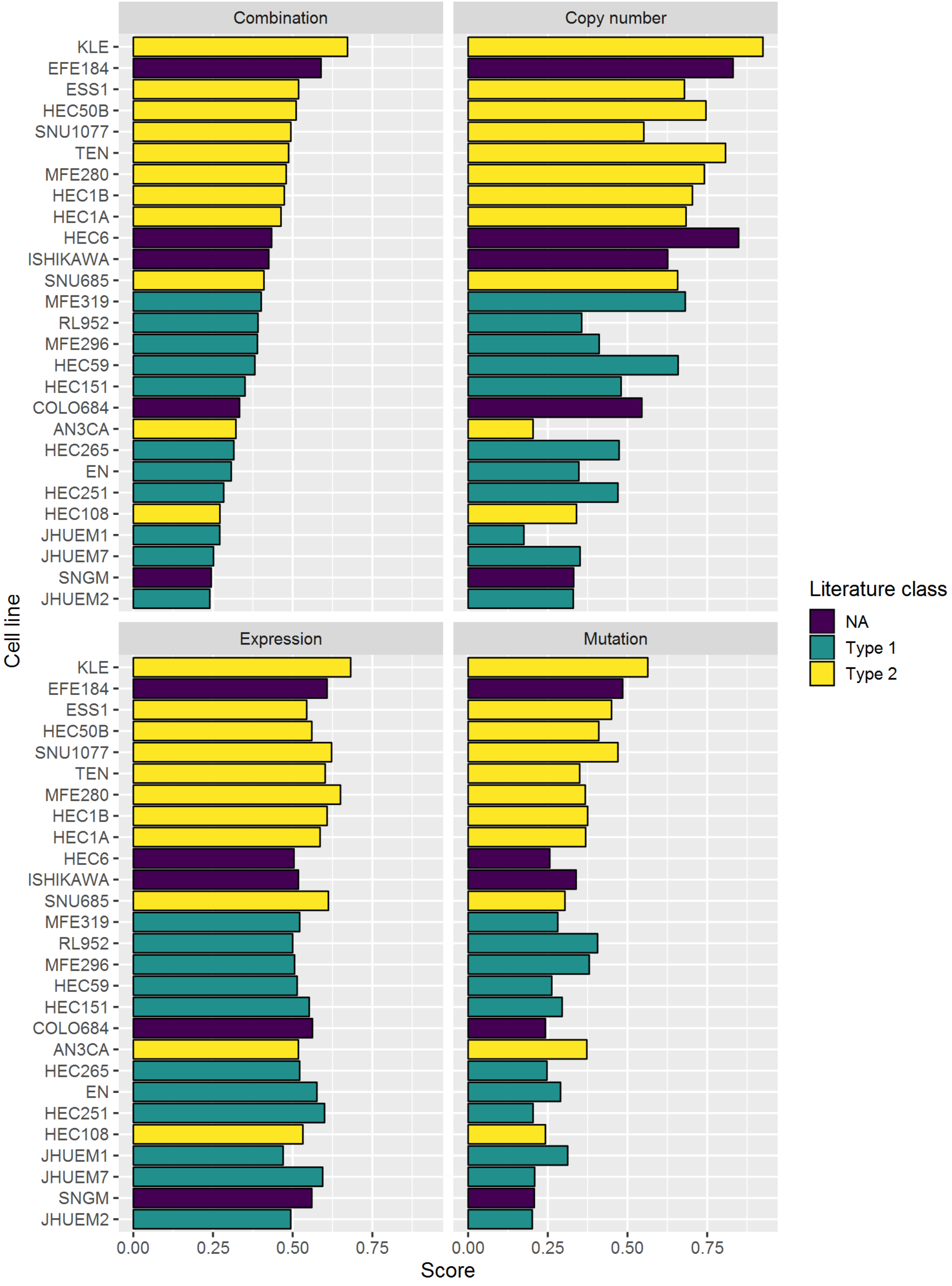


Figure 28: Cell line scores from single data type classifiers and a weighted combination (0.3, 0, 0.7 for copy number, expression and mutation respectively) classifier for endometrial cancer with the dominant literature class marked with colour. Cell lines are arranged according to decreasing combination score as horizontal bars with their names. Score, on the x-axis, is the classifier-assigned probability of the cell line's subtype. The higher the score, the more likely is the cell line to be of serous subtype. The lower scores correspond to endometrioid subtype.

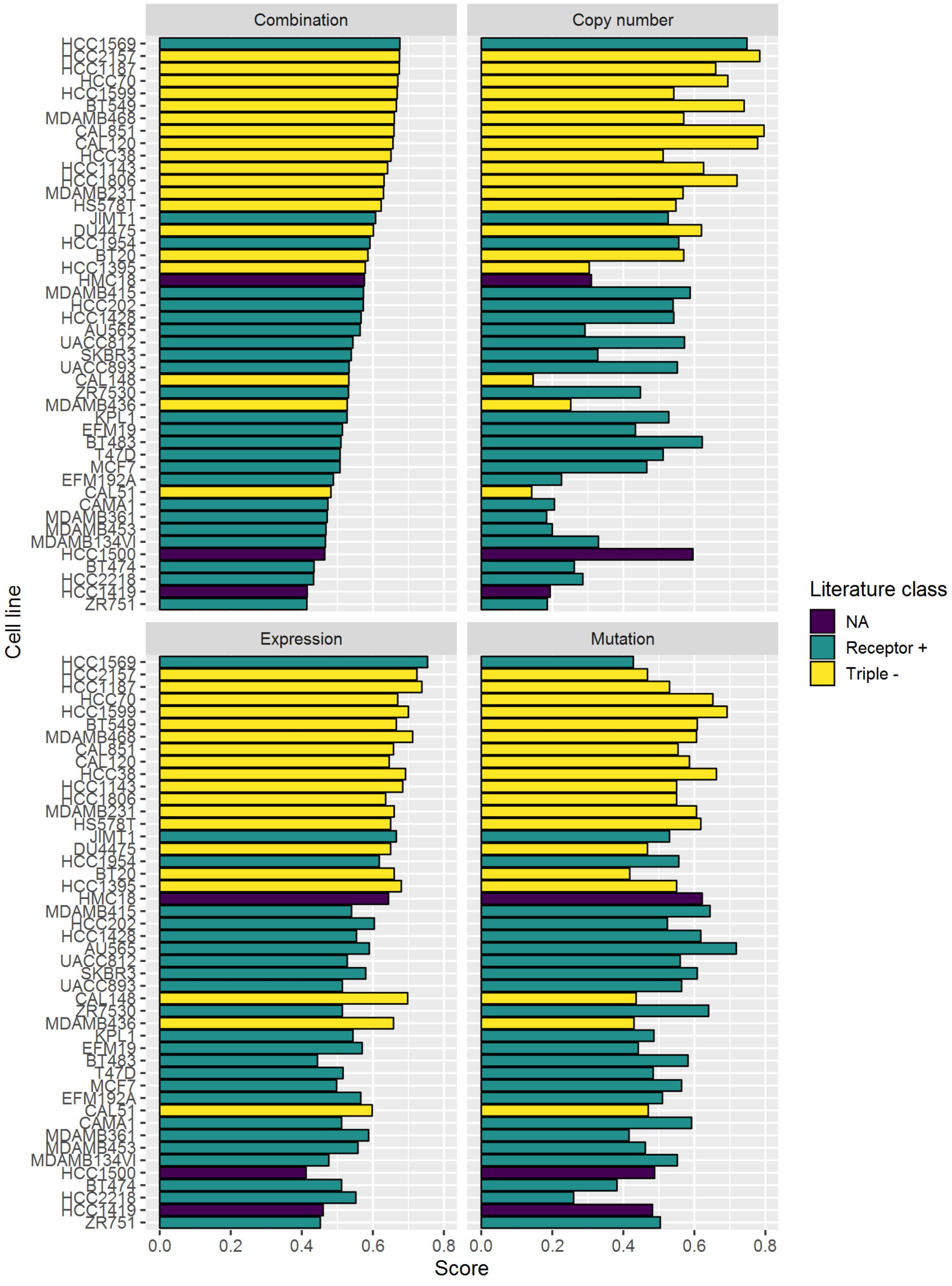


Figure 29: Cell line scores from single data type classifiers and a weighted combination (0.18, 0.58, 0.24 for copy number, expression and mutation respectively) classifiers for breast cancer with the dominant literature class marked with colour. Higher scores indicate, according to the classifiers, triple-negative and lower scores receptor positive cell lines.

scores above 0.5.

Breast cancer

Scores assigned by the classifiers generally conform to how breast cancer cell lines are described in the literature, although the scores tend to be, for all cell lines, skewed towards the triple-negative side.

With respect to the literature, out of single data type classifiers, expression-based on performs the best even though copy number one has higher variance of score.

Ambiguous cell lines, HCC-1500 (triple-negative in (Neve et al., 2006), ER+/PR+ in (Riaz et al., 2013)) and HCC-1419 (ER+/PR-/HER2+ in (Riaz et al., 2013), ER-/PR-/HER2+ in (Kao et al., 2009)), are assigned very low scores by the combination and expression classifiers. While the copy number and mutation classifier give them higher scores, those classifiers had less impact on the accuracy, as determined by the weighting scores (0.18 and 0.24 for copy number and mutation, 0.58 for expression). Unknown subtype HMC-1-8 cell line scored as triple-negative, but is at the lower end, bordering the bulk of receptor-positive cell lines.

HEC-1569, JIMT-1 and HCC-1954 are receptor-positive cell lines which were grouped with the triple-negative ones. Interestingly, all three are HER2+/ER-/PR-. CAL-148, MDA-MB-436 and CAL-51 are triple-negative cell lines grouped with receptor-positive ones. An interesting thing to note is that they have much lower copy number than expression and mutation scores.

Score range

Scores assigned by classifier range from 0 to 1 where 0 represents one class and 1 the other. Boundary between classes is typically set at 0.5. However, I do not think that is the most appropriate threshold for this methodology. Since the score is, basically, saying how close the cell line is to one subtype as opposed to the other, cell lines with score 0.5 might be cell lines which are exactly in between subtypes, but also cell lines which are not close to either, but equally so. For example, in (Figure 22), both EN and MFE-280 seem like they are similarly close to both cancer subtype clusters, but EN is close to both, while MFE-280 is far away from them. Most importantly, the dynamics of subtype scoring when cell lines score close to 0.5 are not known. It could be that, when a cell line does not possess clear characteristics of either subtypes, certain feature of one subtype drives the classification more or less than it should. This seems likely in the case of endometrioid/serous subtypes since it appears the endometrioid subtype is more well-defined, with endometrioid samples closer together, while the serous samples are fairly dispersed (Figure 22, Figure 23). In the case of endometrial cancer, this is further complicated by smaller number of cell lines and larger number of the ones with an undetermined literature classification. Because of this I feel it is more beneficial to show the scores of cell lines compared to each other to give the full picture and avoid assigning a hard border based on score range alone. If there is a need for a specific threshold, I would opt for selecting one which leads to maximal agreement with previous literature (while it would be, based on Figure 29, 0.55 for the breast cancer, for the endometrial cancer it cannot be determined as the literature subtypes differ from the classifier training subtypes).

3.5 Cell line distance measurements

Because the score provides only a one-dimensional measure and cell lines can be poorly representative in a variety of ways not captured by the cancer subtypes used, the scores of the combination classifiers were contrasted with the average cell line distance to a patient sample. Cancer cell lines and patient samples, in these analyses, are defined through their features - values for their copy number aberration, mutation status and expression levels. Since they use the same features (same genes for which the values are obtained), they are placed in the same, multi-dimensional, feature space. This enables an examination of the difference in positions in such space - the proximity or the distance between patient samples and cell lines. While a cell line might score highly for a certain subtype, having much larger average distance to patient samples than other cell lines could still make it a less preferable model. Additionally, it shows a dimension of representativeness which score, in the way used here, does not capture. Since score will always be between between 0 and 1 (in essence one or the other class), by itself it cannot distinguish whether one cell line is inbetween subtypes or completely different from both of them.

Distance of cell lines to patient samples was calculated through standard Euclidean distance ($d(x, y) = \sqrt{[\sum_{i=1}^{n}(x_i - y_i)^2]}$) weighted by importance of the variables (mean decrease accuracy, explained in the previous chapter) in the random forest model. The standard Euclidean distance is calculated as the root of summed squares for each difference in values between two datapoints (a cell line and a sample here). In order to take into account different variable importances, those differences were multiplied by their importance factor, as determined by the random forest model. This was done to base the distance more on the important variables and less on the ones that are not. Otherwise, the difference in unimportant variables would have as much effect on the distance as the same difference in important variables.

3.5.1 Average distance of cell lines to all patients

To measure how representative in general a cell line is of patients I calculated the average distance to all patient samples (Figure 30, Figure 31). This would be expected to show a uniform distribution of distances with a slight skew making cell lines of the more represented class closer. Cell lines not fitting into this pattern might be particularly bad or good models.

Endometrial cancer cell lines appear to have greater variability in their distance. HEC-251 and HEC-108, a type 2 cell line with a surprisingly low score, are noticeably more distant, on average, from the patient samples than the other endometrial cell lines. This might be due to some unique properties of these cell lines or to type

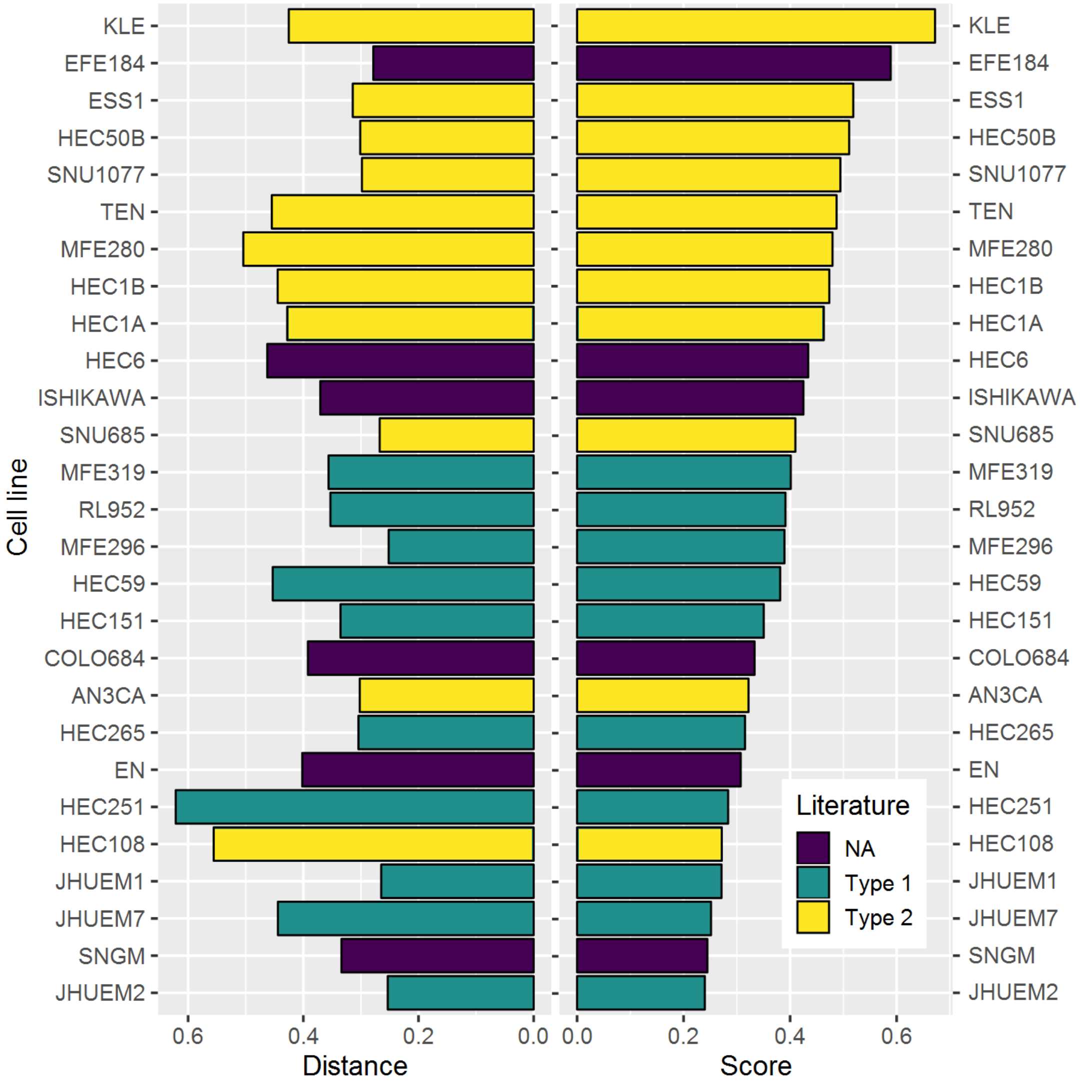


Figure 30: Endometrial cell line score and their average patient distance, ordered by score. Literature assignment is shown in colour.

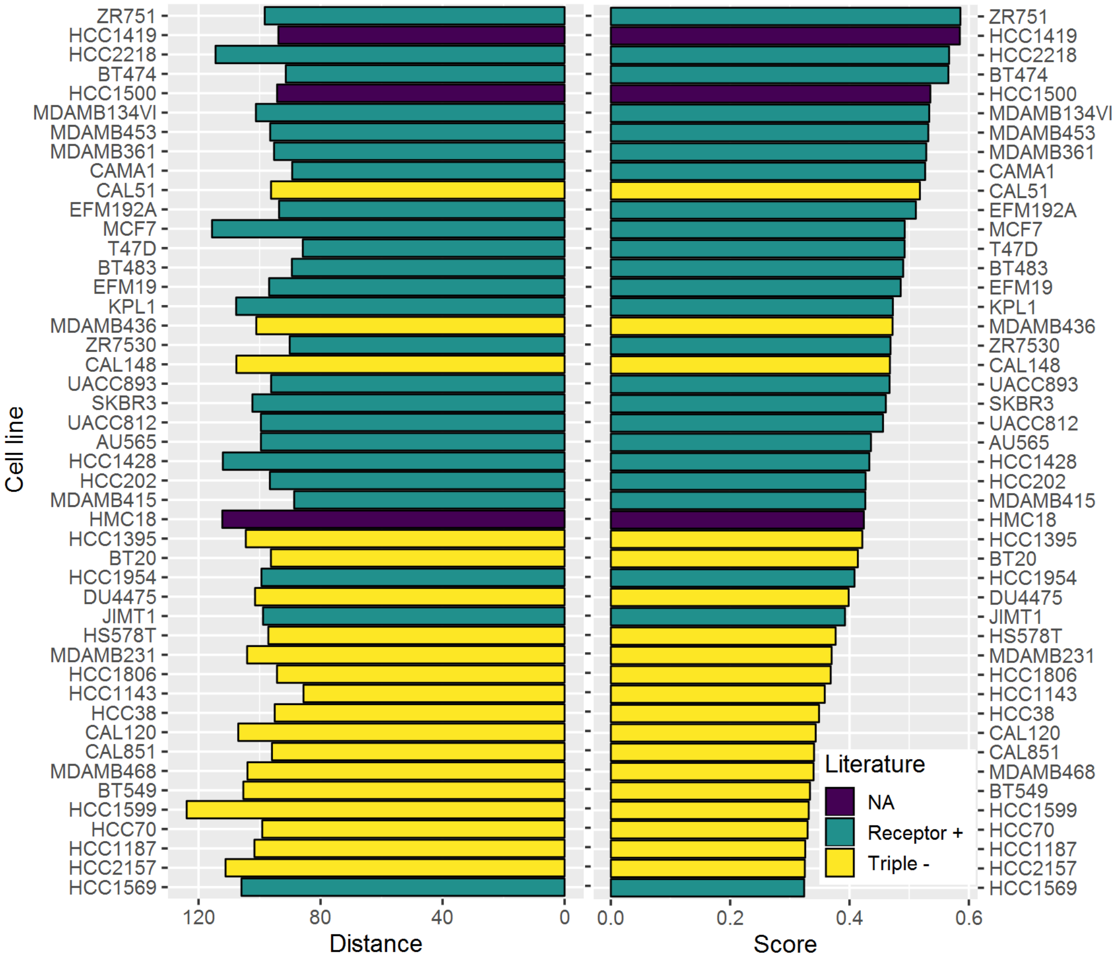


Figure 31: Breast cancer cell line score and their average patient distance, ordered by score. Literature assignment is shown in colour.

2 being more aggressive and thus having cell lines that could drift away further from the patients due to accumulated mutations. The later does not seem to be supported judging by other type 2 cell line distances. Cell lines scoring near the end of the range and having low distance are likely a good choice for the respective subtype. JHUEM2, even though it is rarely used, has the lowest score and it is one of the closest cell lines to patient samples, making it a great choice for pre-clinical endometrioid subtype research. Interestingly, EFE-184 is one of those cell lines despite not clearly defined by the literature aside from being mentioned as metastatic and endometrioid (Artimo et al., 2012). Other ambiguous cell lines have noticeably higher distances to patient samples. Ambiguity could be a result of being very distant to both subtypes, thus being marked as an inbetween case.

Despite majority of the patient cohort being endometrioid and no representation of histologies not included in the subtypes (clear-cell, sarcomas), patient distances do not follow a particularly defined distribution along the classifier score range. With majority of the TCGA samples being endometrioid and some of the cell lines being of neither subtypes (Table 7), this is somewhat surprising. While it could be that the cell line specific alterations present in all the cell lines have pushed them all equally away from the original cancer population, that cannot explain a clear variation of distance between the cell lines. Especially considering the distances were calculated using the most important features, as determined by the high-accuracy classifiers.

Breast cancer cell line distances to the TCGA cohort show much higher uniformity. This could be attributable to the size of the cell line pool being mucg larger than in endometrial cancer. Cell line closest to the patient samples is T-47D which has been extensively used in reasearch (Table 11) while the one furthest apart is HCC-1599, a triple-negative cell line with only 3 citations.

None of the misclassified cell lines is particularly close or distant, when compared to the rest of the breast cancer cell lines. Not being distant is particularly interesting since it suggests their misclassification is not due to not fitting the patient cancer landscape.

3.5.2 Cancer cell line distance to patient samples

For each patient the closest cell line was annotated. Then for each cell line the number of patients with that cell line annotated as the closest one was counted ("counts" in further text). An examination of which data types of patient endometrial cancer samples are closest to how many cancer cell lines can be seen in Figure 32, and in Figure 33 for breast cancer. Combination distance was the result of combining the other distance with the same weighting factors the weighted combination classifiers were created. While some previous patterns do repeat (majority of patient cancer samples are closest to a cell line of the matching type), there is a significant incongruence across the data types. These results emphasise the importance of integrating multiple data types in genomic analyses when possible as well as highlighting different genomic events in same cell line subtypes (e.g. MFE-319 and MFE-296 are both type 1 endometrial cell lines classified as endometrioid, yet they have very different patient sample count profiles).

Compared to the average distance of a cell to all patients, this approach is more specific, informative and does not get skewed by general patterns.

In the case of endometrial cancer - endometrioid counts are more variable, while also having a higher total count (since there are more endometrioid patient samples). The counts do not correlate well between the data types.

The cell lines with the highest counts are MFE-319, MFE-280 and JHUEM-2 for endometrioid ad SNU-1077, MFE-296 and SNG-M for serous subtype. This makes for a large discrepency with respect to classifier score - scores of MFE-319 and MFE-296 are in the middle of the cell line score range and close to the type 1/2 "border" (Figure 30), MFE-280 is amongst the least endometrioid cell lines by score yet every 4th endometrioid sample has it for the closest cell line by mutation and combination distance and SNG-M is the second most endometrioid cell line by score. JHUEM-2 is the only expected high count result here, being the cell line with the best endometrioid score and described as endometrioid in the literature.

Additionally, some cell lines from the end of scoring range, EFE-184, KLE, ESS-1 and JHUEM-7 have low counts for both subtypes.

Combining the data types has varying effects on the patient sample counts, indicating the effect of such integration is not purely additive. For example, MFE-280 and HEC50B have a high number of close endometrioid patient counts for expression data only, but none when using the combination model. At the same time, MFE-296 has a high count only for the combination model counts. Similar lack of additive effect for data types can be seen in receptor-positive breast cancer patient counts for MDAMB175VII and HCC1143 (lower expression-based close counts than the previous one, but higher combination ones).

The breast cancer cell lines (Figure 33) have a much more extreme counts topography. T-47D and HCC-1143 completely overshadow all the other cell lines when it comes to the counts of patient samples that have them as closest cell lines. While T-47D has high counts for both subtypes (albeit somewhat fewer for the triple-negative), HCC-1143 has only for the triple-negative subtype. Unsurpisingly, HCC-1143 is a

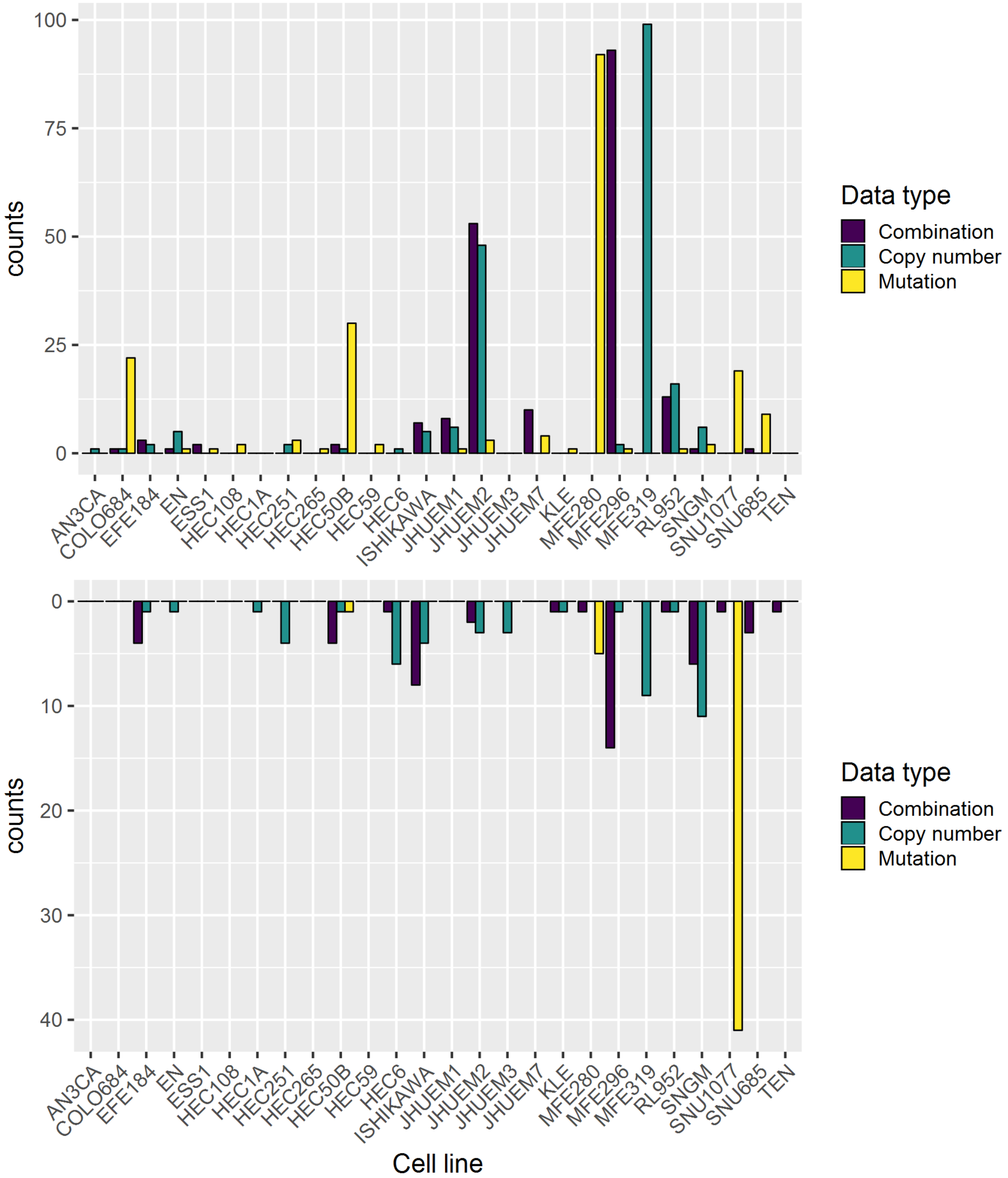


Figure 32: Number of patient samples with a specific cell line as the closest, by subtype and by data type. Length of a bar is the number of patient samples (of the subtype in that area) closest to a cell line. The upper part shows endometrioid patient samples, the lower one serous ones. Different cell lines and their data types are along the x-axis. Colour corresponds to data type (copy number, combination of data types, mutation). Endometrial cell lines not shown were not the closest one for any patient sample in any of the cases.

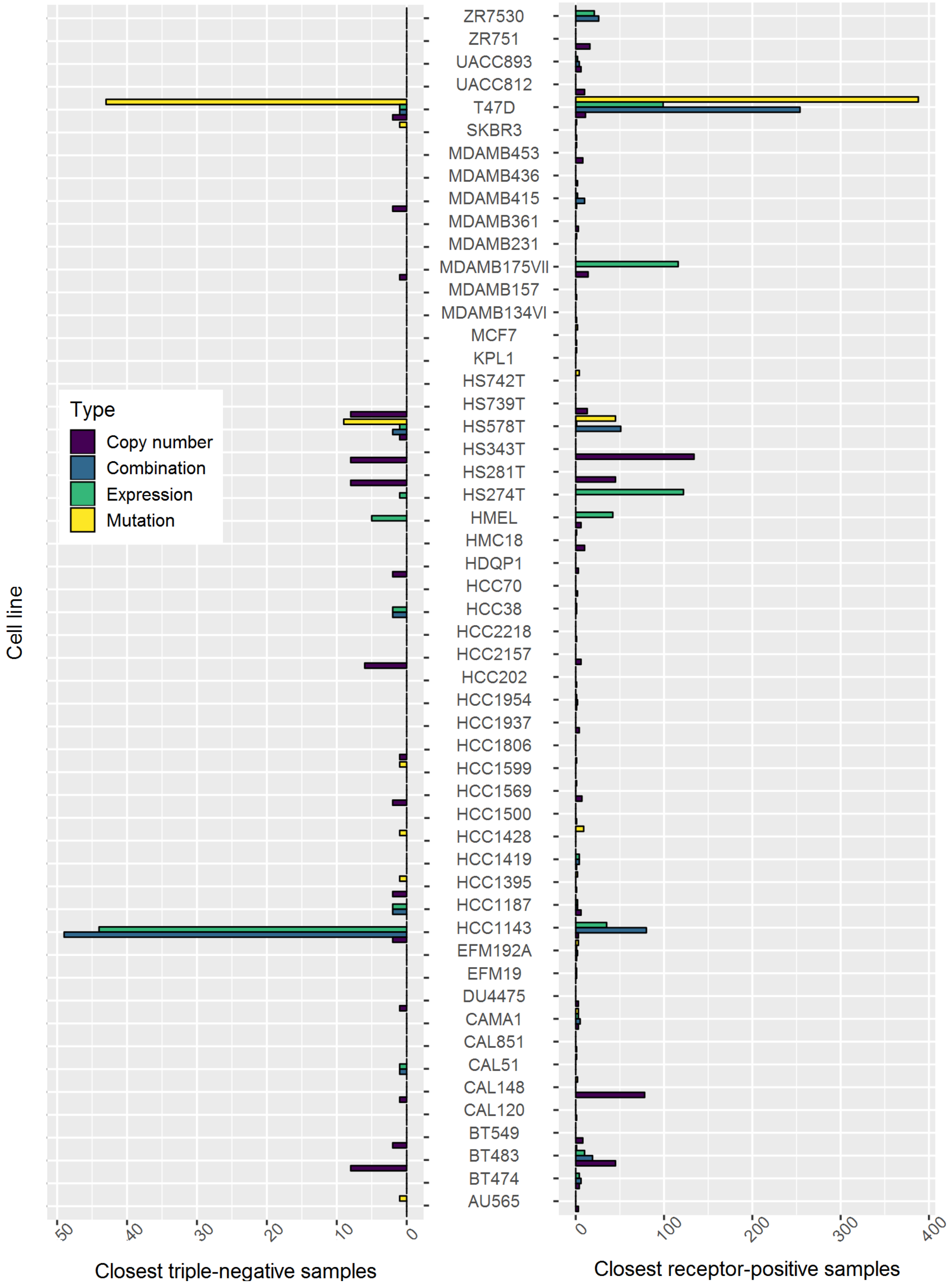


Figure 33: Number of breast cancer patient samples with the specific breast cell line as the closest cell line, by subtype (receptor-positive/triple-negative) and data type.

triple-negative cell line scoring as one as well. Although T-47D is a highly cited (Table 11) receptor-positive cell line scored as such by the classifier (Figure 31), it has a mutation profile which is closest to the majority of triple-negative samples. This makes it unlikely it is just the receptor values that drive the closeness counts. T-47D being shown earlier as very similar to all the patient samples (Figure 31) goes well with the effect seen here. Unlike, T-47D, HCC-1143 has a rather low number of citations (20).

None of the ambiguous or misclassified breast cancer cell lines have an interesting pattern.

3.5.3 K nearest neighbours

A similar distance-based approach to determining a class is k-nn which assigns a class to a sample based on k of its nearest neighbours. In the case of endometrial cancer (Figure 34), such method, while highlighting as serous the same cell lines as a classification approach, appears to be noticeably more conservative in assigning a serous subtype. Similarly for breast cancer (Figure 35), no cell line receives consistent classification as triple-negative. It is rather peculiar that HCC-1569, cell line misclassified as triple-negative earlier, is the most classified as triple-negative.

Although they share a core idea, k-nn, counting which samples have a cell line as closest and average distance to patient samples all bring something unique to cell line distance evaluation. K-nn is a classification method, in a way inverse from looking at which patient samples have a cell line as its closest as it looks at which patient samples are closest to a cell line. While a cell line might not be closest for any of the samples, it is still more similar to some than others and k-nn determines that.

Average cell line distance is an examination of how much a cell line has diverged from the patient cancer population. While it might not play a role in being a certain subtype, highly-divergent cell lines which are of the desired subtype are a riskier investment than the less divergent options, if they exist. They might match the markers of that subtype, but being more different from the overall patient cancer population makes them more likely to have alterations that would make them respond differently to treatment even if they match the subtype markers (e.g. triple-negative cell line that is from a different tissue). The number of samples which have a specific cell line as their closest cell line does not say anything about general cell line - tumour divergence, but it does help in determining which cell lines are better representatives for larger number of samples. However, those samples might not be representative enough of the subtype since they are only a subset of all the subtype samples, thus making it hard to confidently tie the subtype information with the

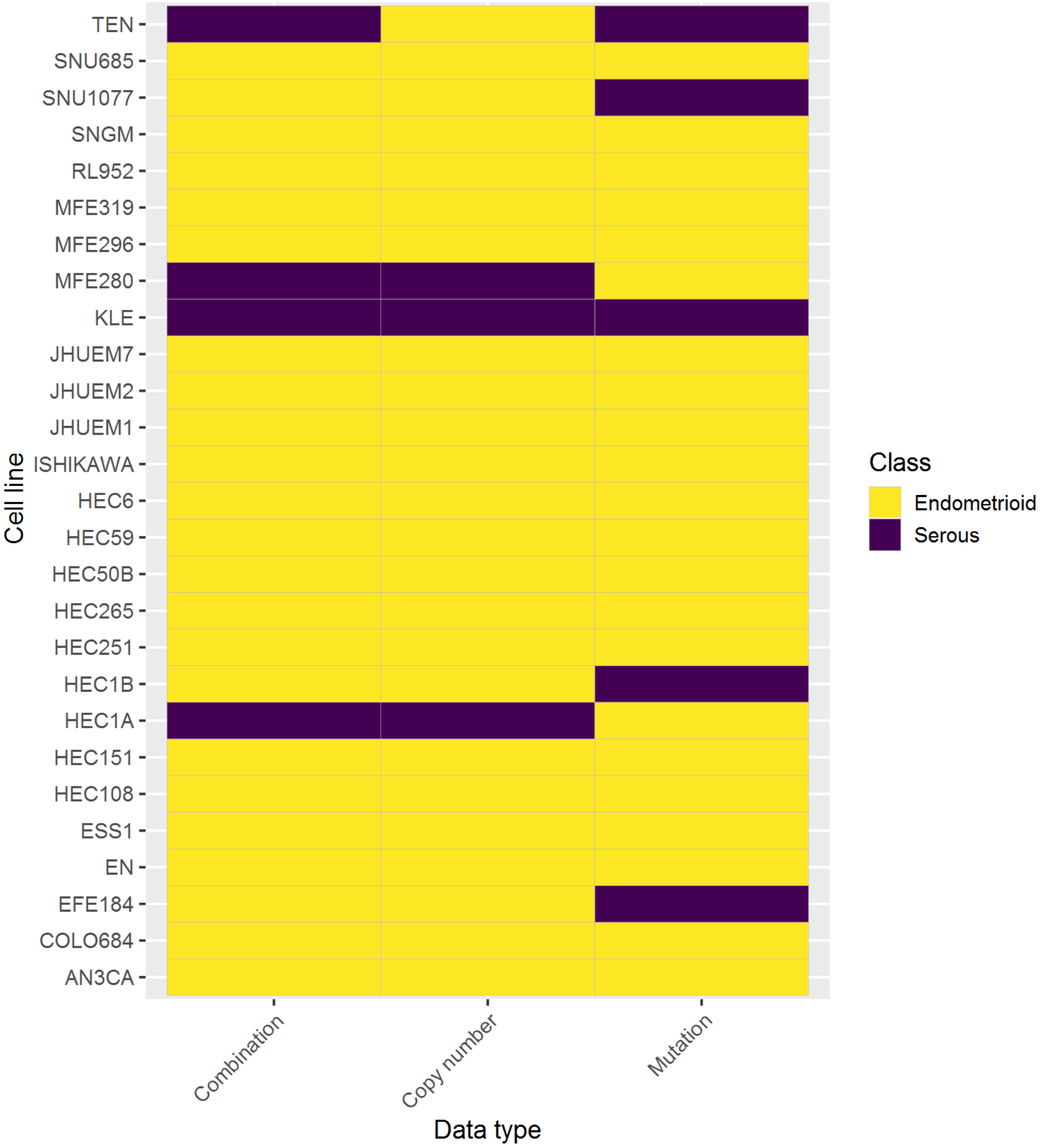


Figure 34: Result of 3-nn class assignment for endometrial cell lines by data types. The distance metric used was Euclidean distance after weighting for variable importance as described earlier. Class assigned (colour) is the dominant class amongst 3 nearest samples, as determined by the distance metric.

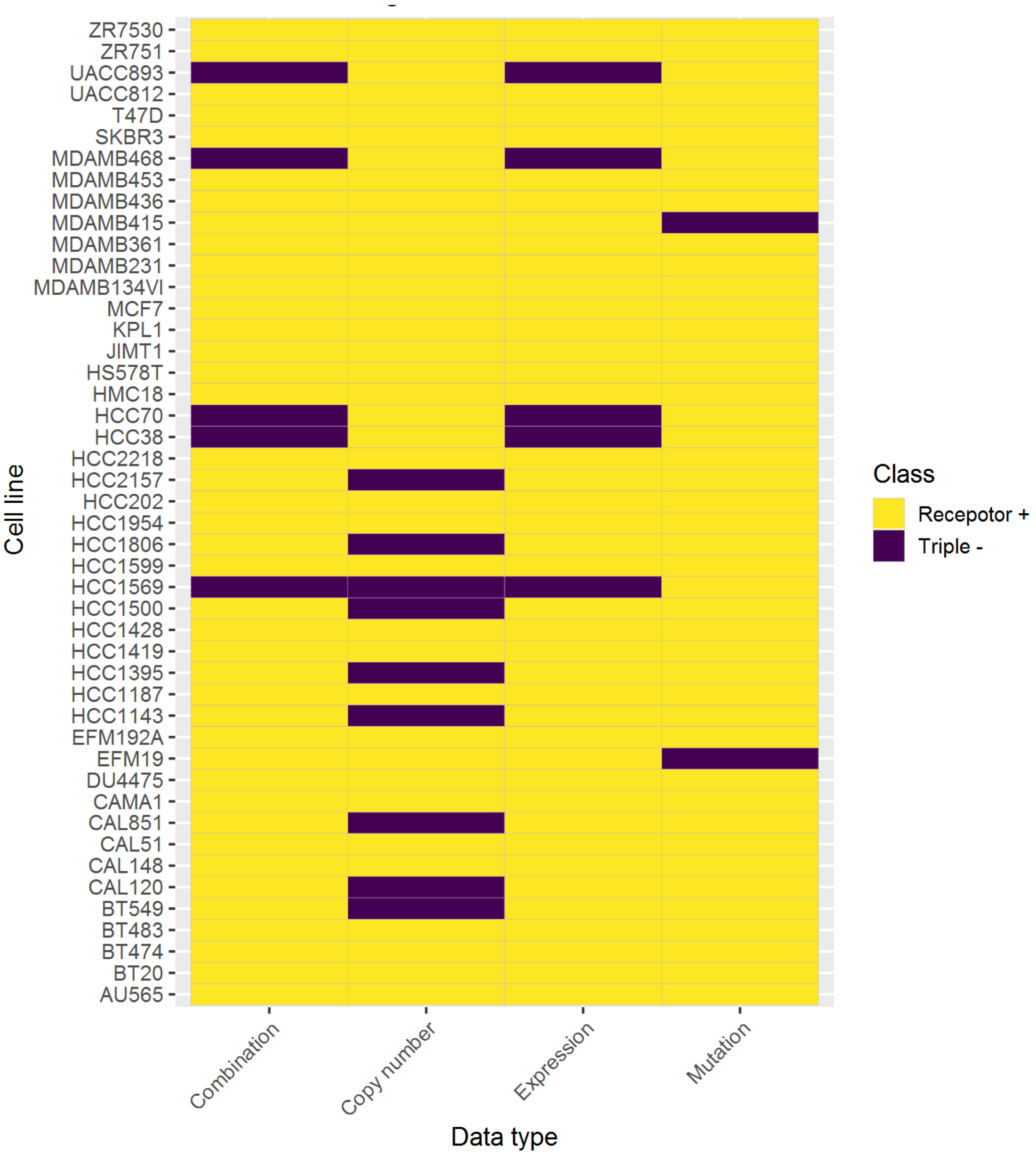


Figure 35: Result of 3-nn class assignment for breast cancer cell lines by data types. The distance metric used was Euclidean distance after weighting for variable importance as described earlier. Class assigned is the dominant class amongst 3 nearest samples, as determined by the distance metric.

specific subset in question without further detailed investigation (e.g. making sure that the subset of the closest samples to the specific cell line is representative of the general subtype population and checking if there is some feature significantly different in that subset). An approach that would complement this issue would be a method which can generalise the subtype information from the entire cohort into a scoring system and then apply it to cell lines, for example the classifier approaches described and conducted earlier.

3.6 Discussion

In this chapter, I have collected literature information on the subtypes for endometrial and breast cancer cell lines, applied standard non-classifying methods to patient samples and cancer cell lines, used the classifying methodology described in the previous chapter to assign a subtype score based on patient data to cancer cell lines and explored different approaches to determining and analysing cell line - patient sample distances.

In the literature search of PubMed, several cell lines were found to have ambiguous or non-existant information. Additional complication was the lack of serous endometrial cell lines in the CCLE, likely because that is the less common and more aggressive subtype. I chose to focus on type 1/type 2 classification scheme for endometrial cancer as there exists a significant ovelap with the endometrioid/serous classification scheme, the cell lines used fall within both subtypes and doing so allows to see how one classification scheme maps on another one.

PCA showed (Figures 22-25) a degree of separation between the cancer subtypes, albeit with principal components which capture low amount of variance. Variance being split across a large number of principal components indicates large dimensionality of independent forces in the dataset, since PCA transforms the data to maximise variance capturing. It could also be attributed to a high-level of noise in the data. In the space of first two principal components, the endometrial subtypes have different centres of their samples with the serous subtype being more dispersed in the copy number PCA. The breast cancer subtypes appear to have similar variances, but their centres are still distinctly apart in both cases. In all of the PCA analyses, cell lines varied between the cancer subtypes, although, with respect to the literature-derived cell line subtypes, meaningfully so only for the copy number endometrial PCA and the expression breast cancer one. In those cases, there is a greater tendency for cell lines marked as one subtype in the literature to be placed closer to the respective subtype cluster in the PCA. However, low variance captured by single principal components and a large overlap of patient subtype groups in the PCA make it difficult to draw strong conclusions from the PCA analyses shown here.

Random forest classifiers applied to cell lines were highly-accurate on the patient test set (see previous chapter). I focused on the combination classifier since it was made from other classifiers in a way that maximised accuracy on the test set, thus being the most accurate of all. Scores assigned to cell lines were, generally, in line with the literature with a few interesting exceptions. Breast cancer cell lines JIMT-1, HCC-1569 and HCC-1954 have been described, unambiguously as far as I can tell, as HER2+ (ER-/PR-) in the literature while being scored by classifiers as triple-negative. This misclassification persits in k-nn classification for HCC-1569. Excluding model suboptimality, which is not unreasonable considering high accuracy scores (AUC = 0.941, F_1 = 0.877), a potential cause of this might be a cancer cell genomic profile overall more similar to a triple-negative patient sample despite the difference in receptor expression (HER2+ as opposed to HER2-). This could a result of cell lines having drifted from the original patient sample through laboratory culture sustaining or maybe because the original patient sample was an atypical HER2+ cancer, more similar to triple-negative cancer population than HER2+ typically are. If the later is the case, that type of cancer would probably manifest clinically as more resistant HER2+ type making these cell lines good candidates for a more specific cancer subtype.

The scoring range was also similar to what can be seen in the PCA analyses. Copy number PCA shows a degree of similarity to cell lines scoring of the random forest classifier. SNU1077, SNU685, MFE280, KLE and HEC1B are scored as serous by the classifier and, in the copy number PCA plot, they are quite isolated from the big cluster of cell lines near the endometriod subtype centre (Figure 22). An interesting similarity can be seen in the breast cancer expression PCA - cell lines closest (using Euclidean ditance in the space of first two components) to the triple-negative cluster (HCC70, HCC1954, Hs578T, HCC1143, BT549) are triple-negative themselves with the exception of, again, HCC-1569.

Although random forest classifiers had high accuracies on the patient test set for all the data types, number of samples with a specific cell line as their closest (Figure 32, Figure 33) varies greatly by data type. This count of patient samples shows interesting patterns unavailable through other used methods. Some cell lines can be a rather good model for a patient tumour of a different subtype (T-47D for triple-negative, SNU-1077 for serous and MFE-280 for endometrioid), judging by what cell lines are closest to which samples.

My cell line recommendations, as a result of this chapter, would be using JHUEM-2 for endometrioid, or generally endometrial, cell line. It scores as highly endometrioid, the literature notes it as endometrioid (although it is rarely used), it is generally representative of patient cohort when compared to other cell lines on overall average distance to patient samples and a large number of endometrioid patient samples have it as their most similar cell line.

For the serous subtype, it is difficult to confidently pick a candidate cell line since serous cell lines do exist, but none were found in the CCLE. KLE has a rather high serous score, however it has been mentioned as endometrioid in the literature. Further investigation of high-scoring (i.e. scoring as serous) cell lines would help elucidate the proper use of those cell lines, as well as CCLE including serous cell lines (e.g. ARK-1, SPEC-2) in their cohort.

Based on the breast cancer cell line analyses here, I would pick HCC1143 for triple-negative and T-47D for receptor-positive based on them being the closest cell line for a large number of samples of the respective subtype, their classifier scores and the literature.

Different approaches picking up HCC-1569 as a triple-negative cell line, despite the literature classificationas HER2+ and other breast cancer misclassifications being of HER2+ cell lines warrant further research of their biological background (see next chapter). Additionally, classifiers having an importance determination as their key characteristic which can be investigated for more insight into what drives the score differences also lends itself to biological causation investigation.

4 Exploration of the molecular drivers behind subtype assignments

4.1 Introduction

In the previous chapter focus was on the patient and cell line scoring in the context of computational subtype modelling. The highly-accurate classifiers developed in the first chapter were used to assigned scores to cell lines. Generally these were in line with the literature-sourced classifications. Here, the underlying genetic properties of these models will be examined, providing insight into which markers determine subtype of specific cell lines and what lies behind some of the unexpected classifications.

Thus the methods implemented previously are useful not only because of their accuracy or ability to determine a subtype, but also provide additional insight into the underlying molecular drivers. This will be determined through a model's choice of important variables or relationship between differently scored cell line groups, especially the cell lines which were not assigned the expected class.

First the status of features important for the random forest will be examined in these cell lines. Following this a more unbiased examination of the difference between misclassified cell lines and those of the cell lines in both the expected and assigned class. Together these shed light on what differentiates these cell lines from the others, showing not just why they might not make a good model for studying the cancer in question, but also suggest situations in which they might be a useful model.

Another important function of these analysis is validation of the drivers of classification results. While the biological background for endometrial and breast cancer, when it comes to the subtyping scheme I focused on, is fairly established, if the underlying genomic drivers are picked up by my scoring methodology, using the same approach with a new or underresearched data set could not only help cell line scoring, but also the general understanding of underlying causative elements.

4.1.1 Determining feature importance in random forests.

One of the key advantages of random forests, in addition to their accuracy and ease of implementation is that they are transparent in that they allow easy access to the features that are important in classifications. Random forest can measure how influential the variables of the data are on the desired classification through two variable importance measures - mean decrease accuracy and mean decrease Gini.

Mean decrease accuracy can be used to measure the importance of a variable by looking into the overall accuracy (properly classified samples) when that feature is missing. It relies on the fact that each tree in a forest is built using only a subset of the samples. The so-called "Out-Of-Bag" samples are those that were not used for a given tree. For a feature m and tree t, the decrease in accuracy is calculated as the number of correct classifications there are in the out-of-bag (OOB) samples for t, minus the number of correct classifications when m is randomly permuted in the OOB samples. This is then averaged across all the trees to give the mean decrease accuracy for feature m.

Mean decrease Gini, for a specific feature, is the decrease in gini impurity index in the case of removal of that feature, calculated in the same way as mean decrease accuracy, but with gini impurity index instead of the number of correct classifications.

Gini impurity index is used to measure, in the case of random forest classification, node impurities. That is, how likely is a randomly chosen sample that ended in that node to be mislabelled if it was labelled according to label distributions of that node's samples (Equation 20, N = the node for which the impurity is being calculated, C = all class indices, p_i = probability of a randomly chosen sample from the node being of class i).

$$I_{gini}(N) = \sum_{i=1}^{C} p_i \sum_{k \neq i} p_k \tag{20}$$

Since the features consist of gene expression, copy-number variation and/or mutation status; and the classification in question is cancer subtype; random forest variable importances can show role of specific genomic events in chosen cancer subtypes provided class prediction is accurate enough.

Gini impurity index is particularly relevant in analyses dealing with some sort of equality or inequality value, e.g. Gini is often used to measure income inequality. While the Gini impurity index can be quite informative in understanding the relationship between a certain feature and the data set, I focus on mean decrease accuracy (MDA) more since it directly relates to model accuracy and role of the variable in it, unlike the Gini impurity index.

4.2 Methods

4.2.1 Data

For differential expression analyses, Cell line RNA-seq count data was obtained from the CCLE https://data.broadinstitute.org/ccle/CCLE_DepMap_18Q1_RNAseq_r eads_20180214.gct.

4.2.2 Drug analysis

IC_{50} (substance concentration needed for 50% inhibition effect) for breast cancer cell lines was obtained from https://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources/Drug_screening_data.html

There were 52 breast cancer cell lines available in the dataset that had at least partial data. The drugs and cell lines which had more than 50% missing values were removed from the dataset while the rest of the missing values were set to the mean of the drug across cell lines.

To determine average drug response, for each drug and group of cell lines (either HER2+, receptor positive or triple negative, excluding HCC1954, HCC1569 and JIMT-1) the mean IC_{50} was calculated. The distance of each cell line from each of these means was then calculated. Finally this was summarised per cell line as mean absolute distance over all drugs/euclidean distance across drugs.

In order to determine if a cell line's drug response was more like group A (e.g. HER2+ or receptor positive) or group B (e.g. triple negative), the distance of the cell line from group A was subtracted from the distance to group B.

4.2.3 Differential expression analysis via DESeq

Cell lines were classified as triple-negative, receptor +, HER2+ or "misclassified" (cases of model prediction not being compatiable with the known subtype information) on the basis of receptor expression from the RNAseq data. Differential expression analysis was carried out using DESeq2 (Love et al., 2014). Pathway enrichment was computed using the GOSeq package (Young *et al* 2010). For scripts see https://www.github.com/vuzun/thesis_methods.

4.3 Gene subtype significance as determined by computational models

The models use genomic information as features and assign different importance to features depending on how much they help the prediction of the selected classes, that is, how predictive they are. Knowing which features the computational models find predictive can give clues about biological functions underlying the classification not captured in the score alone.

In the models examined, I have chosen the joined feature combination model (created by joining the different data types into one and then training one model on that combined dataset) instead of the weighted combination model (combining just the final scores of single data type models through optimised weights) since the latter one is, technically, not one random forest model, but a weights vector used to create a linear combination of the existing single data type models.

Additionally, the joined features model gives information on how important different data type features are by themselves in the context of the other data type. For example, if there is an extremely important copy number feature that plays more subtype-determining role than all of the mutation features, but the rest of copy number features are irrelevant, it would show in the joined features model as important, but that might not be obvious from the weighted combination model since it could assign a very low weight to copy number dataset because of the overwhelmingly irrelevant dataset.

4.3.1 Endometrial cancer

Mutation model

Features with a mean decrease accuracy less than zero are irrelevant for a model's decision in the classification assignment. Thus the number of features with mean decrease in accuracy of greater than 0 is a measure of how many features the model is using. The endometrial mutation model had an AUC of 0.91, which is only slightly lower than the full combination model. However, out of 1406 features used in the this model, only 143 had higher than 0 mean decrease accuracy.

Looking at the highest importance scores (Figure 36) and the distribution (Figure 37), it is clear a small number of mutations completely drive the decision-making of this model. Many of these features are well known oncogenic mutation drivers, including 6 of the top 10 features: tumour supressors - *TP53*, *PTEN*, *PIK3R1*; and proto-oncogenes - *CTNNB1*, *KRAS*, *MTOR*. *TP53* mutation are found almost exclusively in serous tumours, while the others are almost entirely excluded in serous tumours.

Of the top nine most serous cell lines, 8 carry a mutation in *TP53* (shown by green marked cells in the upper part of Figure 38A), and only two a mutation in *PTEN*, while of the remaining 20 only 8 carry a mutation in *TP53*, but 17 carry a *PTEN* truncation. Interestingly *TP53* and *CTNNB1* are almost completely mutually exclusive, and most cell lines carrying a PTEN mutation, carry one or the other of these. Only a single line carries a mutation in none of these genes.

Copy number model

Out of 4000 features in the copy number random forest classifier (which had an

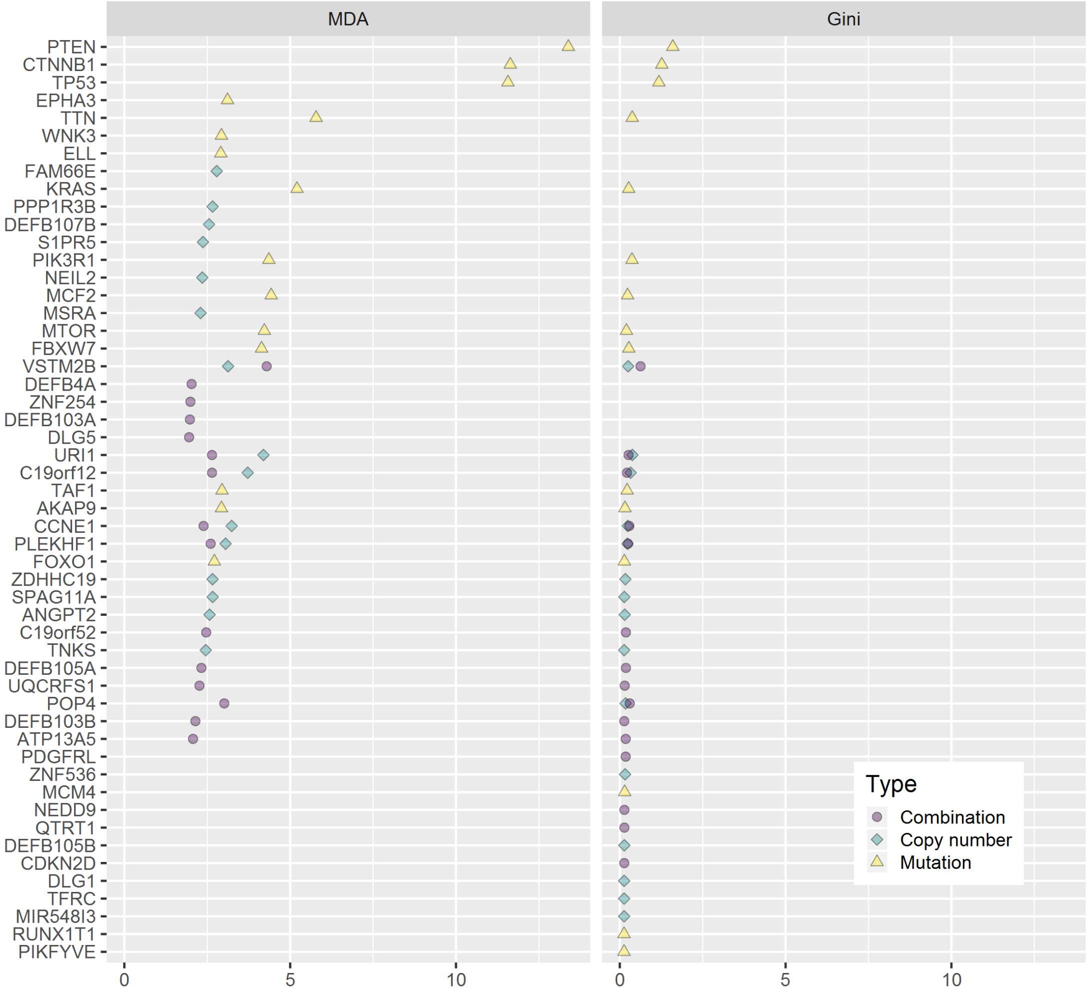


Figure 36: Most important endometrial cancer features. The shown features were the ones showing in top 15 features by Gini or MDA in any of the used classifiers (pictured as different colours and shapes).

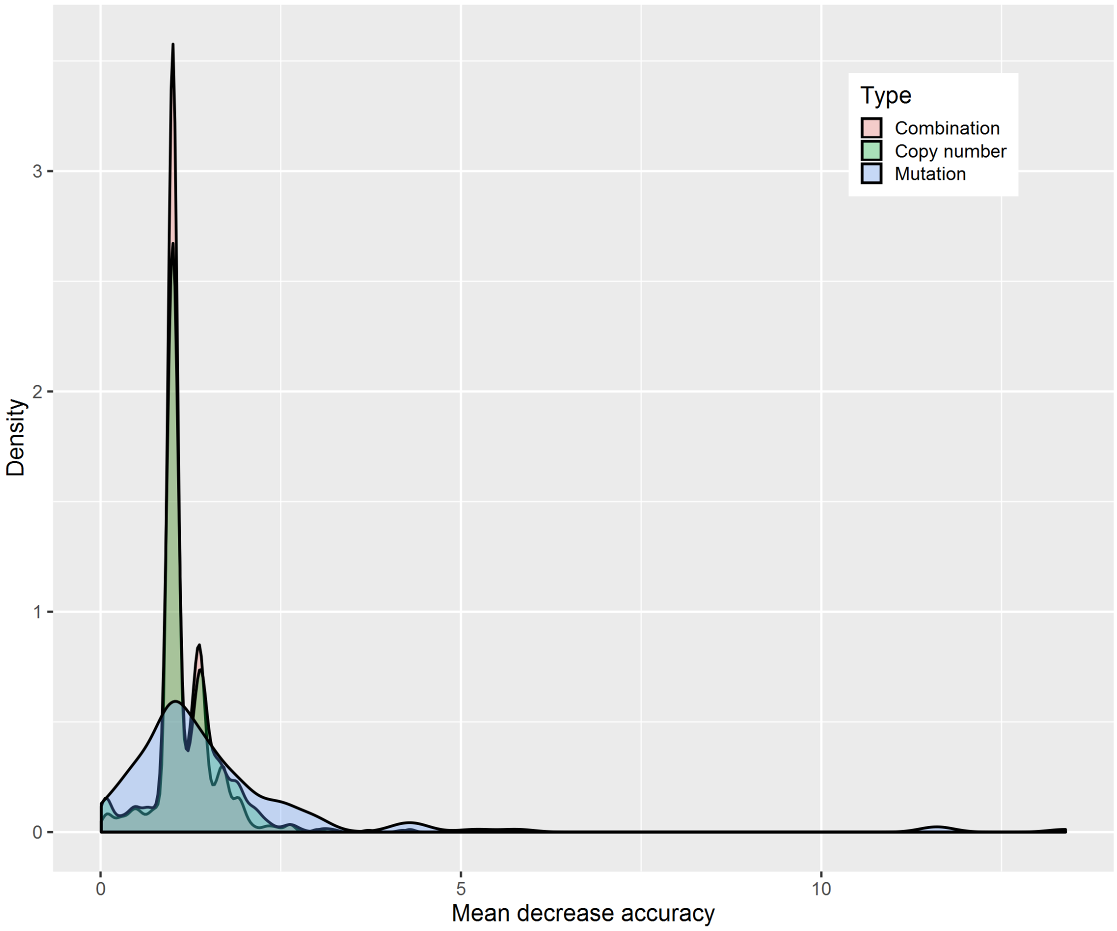


Figure 37: Distribution of MDA values of random forest features for endometrial cancer classifiers.

AUC of 0.925), 657 had greater-than-zero mean decrease accuracy. Most of the importance-positive features fall between 1 and 2 mean decrease accuracy (Figure 37, Figure 36).

Of the top ten most important features in the copy number model, 7 map to 19q12, which has been detected to be amplified in breast cancer. Most of the patients carrying this amplification are serous and around 40% of serous patients carry it (Figure 38B). One of the genes in this amplification is the E1 Cyclin *CCNE1* and is a common target of copy number alteration in several tumours (Zack et al., 2013). More specifically, in high-grade and type 2 endometrial cancer, *CCNE1* and *URI1* were found to be amplified (Noske et al., 2017). Two of the remaining features, *PPP1R3B* and *FAM66E* are also located in close proximity on 8q23.1, and several of the other genes in this region appear further down the list of important features for the copy number classifier. It is more than 3 times more common in serous compared to endometrial cancer (7.1% vs 2.2%).

Amplification of 19q12 is seen in 4 of the 7 most serous cell lines (KLE, SNU1077, TEN and MFE280). Cell lines carrying the amplification of 8q23.1 seem to be more in the middle of the rankings - more serous than the clearly endometrial cell lines, but not classified as high as the top 10. Some lines carry amplification of some of the genes in 8q23 and these tend to be amongst the most serous (e.g. HEC265, EN and JHUEM2).

Interestingly neither ESS1 nor EFE-184 carry either of the high scoring CNA alterations, yet still score highly on copy number classifier, suggesting that their high score must be due to a combination of features, further down the list.

Feature join model

Of the 3406 features in the joint feature model, 2728 features had a mean decrease in accuracy less than or equal to 0, leaving 518 features with a positive for mean decrease accuracy. Only 2 (*PTEN* and *MAP3K5*) were mutation features with the other 1404 mutation features not contributing. These 2 features ranked moderately high within the important 518, only appearing at the 55th place. This demonstrates the importance of a wide variety of copy number features, as opposed to few mutation ones, in determining the endometrioid/serous subtype since there is so many copy number features that contribute (i.e. have MDA higher than 0) to the joined model compared to the number of mutation feaures. Potential cause could be some copy number features covering for the same subtype-determining mechanism as a large number of mutation features.

Unsurprisingly, most of the top features by importance in the copy number model

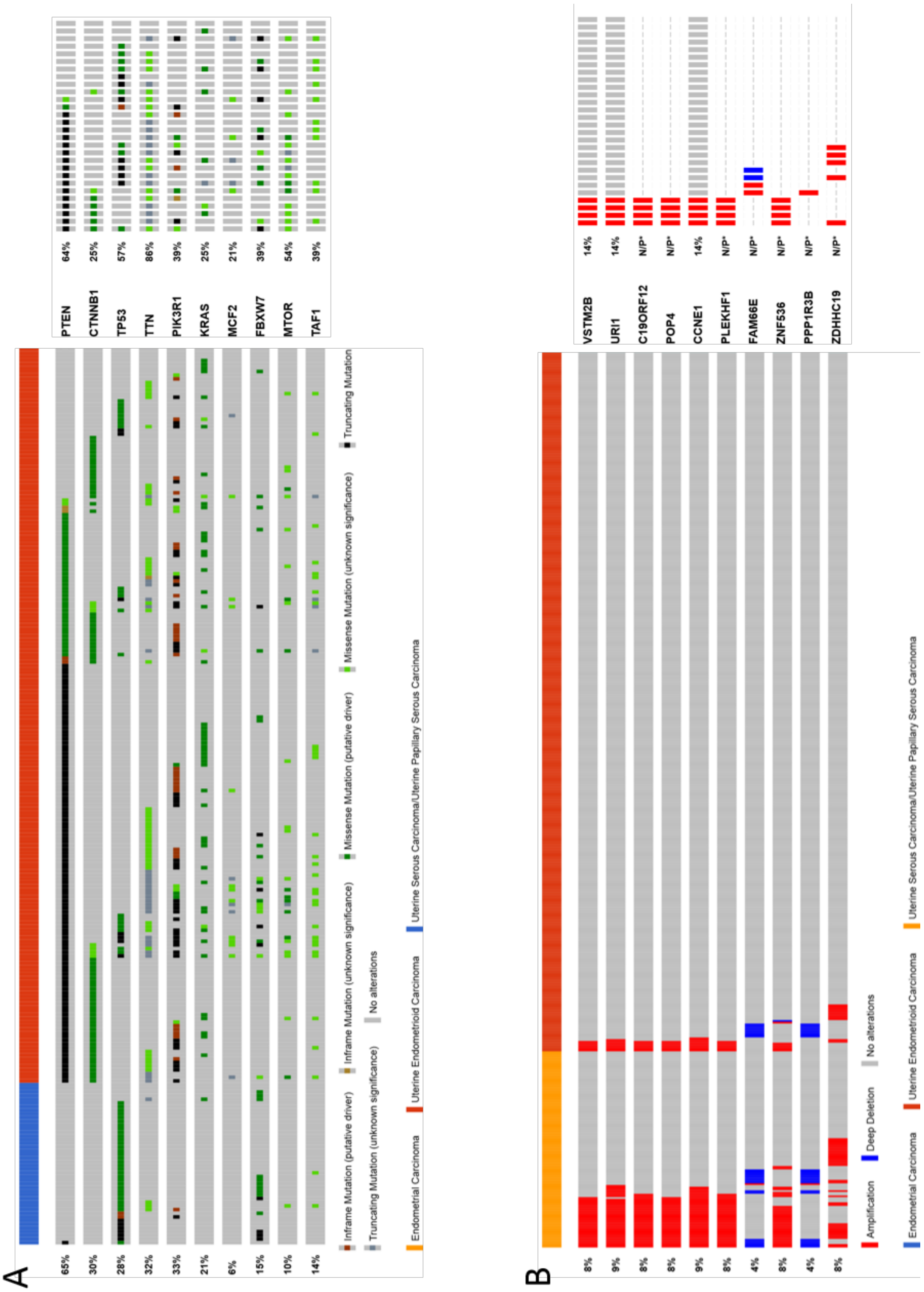


Figure 38: Alteration profiles for highly predictive features in both TCGA patient samples and CCLE endometrial cell lines for A) Mutations and B) Copy number alterations. Plots were generated using the cBioPortal website and the top 10 mutation and copy number features selected by the random forest using MDA. TCGA samples are at the bottom, CCLE on the top. The leftmost bar next to the TCGA samples is the cancer subtype.

are shared with the combined model. The overwhelming majority of the non-zero importance features have between 1 and 2 mean decrease accuracy (Figure 37), similarly to the copy number. Looking at the features with the highest importance (Figure 36), genes in the 19q12 amplification are clear outliers for the combination model through feature joining.

Interestingly, the decrease in accuracy is much larger for the top features in the mutation model than the copy number model, meaning fewer features are driving the classification in the mutation model than the copy number one. However, that might not make them overall more important as can be seen from the joined model. It could be more expected for features like that to be less important in the joined model since it is easier for the other data type to cover for a large part of the importance - it is concentrated in few features so the added new features (copy number here) have more "opportunity" to cover for them.

While *PTEN* has high importance in the mutation-based classifier (Figure 36), that is only in the context of other mutation features. When copy number features join the competition in the combination model, it falls behind (since PTEN has no combination model marker in Figure 36, it is not in the most important features for the combination model).

TP53 was the 55th feature by mean decrease accuracy and 61st by mean decrease Gini, while still being the most important mutation feature in the model. Only 5 mutation features (*TP53*, *PTEN*, *PIK3R1*, *KALRN* and *TPR*) out of 1406 had greater than zero importance scores. This means that, for the combined features model, all other mutation features play no role in determining classification score or are better left out. The cause could be adaptation to cell culture leading to high number of new mutations not normally present in the examined subtypes. In the case of copy number, 449 out of 2000 had greater than zero importance scores.

While these genes being ranked as highly important is expected, it shows how through a high-accuracy model, important variables with biological functional causality can be picked up.

4.3.2 Breast cancer

Mutation model

Only 20 out of 428 mutation features have mean decrease accuracy higher than zero (Figure 39). Interestingly, 186 mutation features have mean decrease gini higher than 0, albeit still rather low (< 0.1 with the range of 0 to 0.6 for all data types). This incongruence shows how the ability of a feature to reduce impurity does not

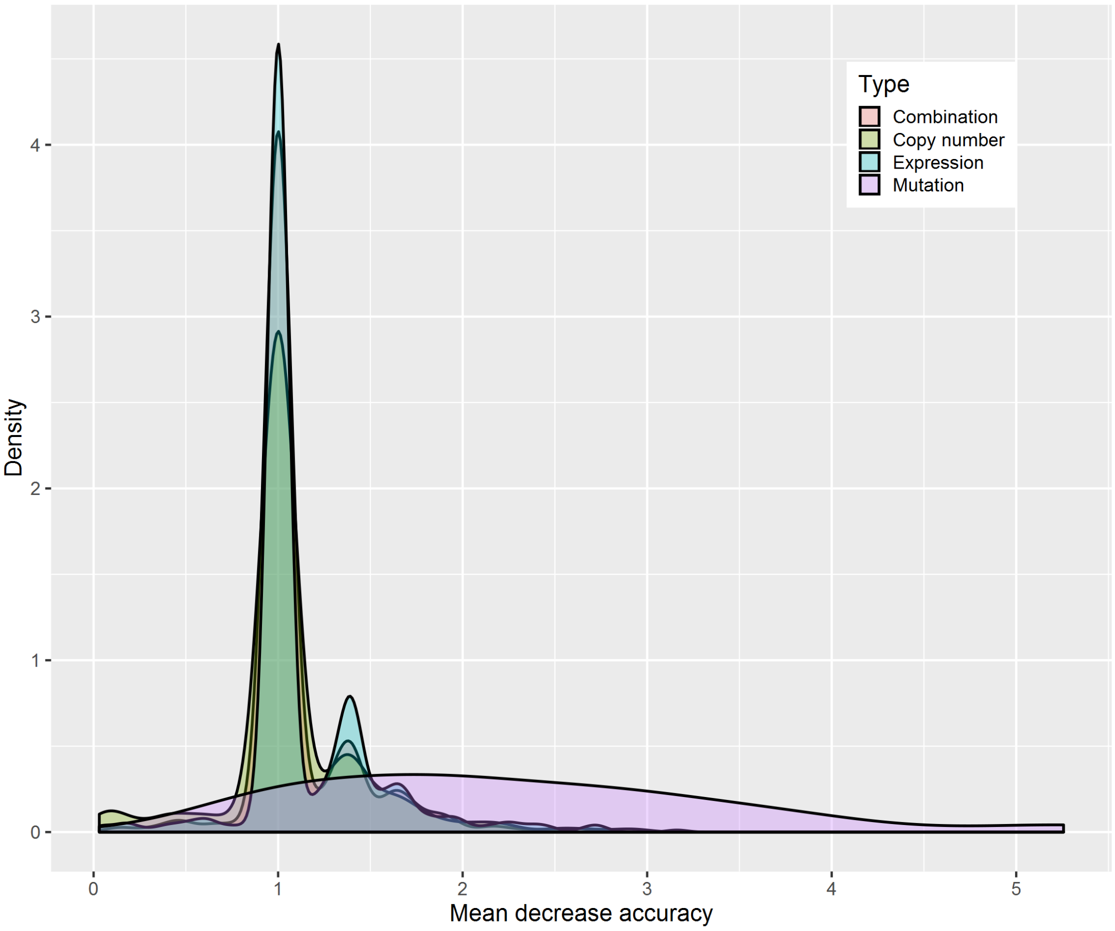


Figure 39: Distribution of MDA values of random forest features for breast cancer classifiers.

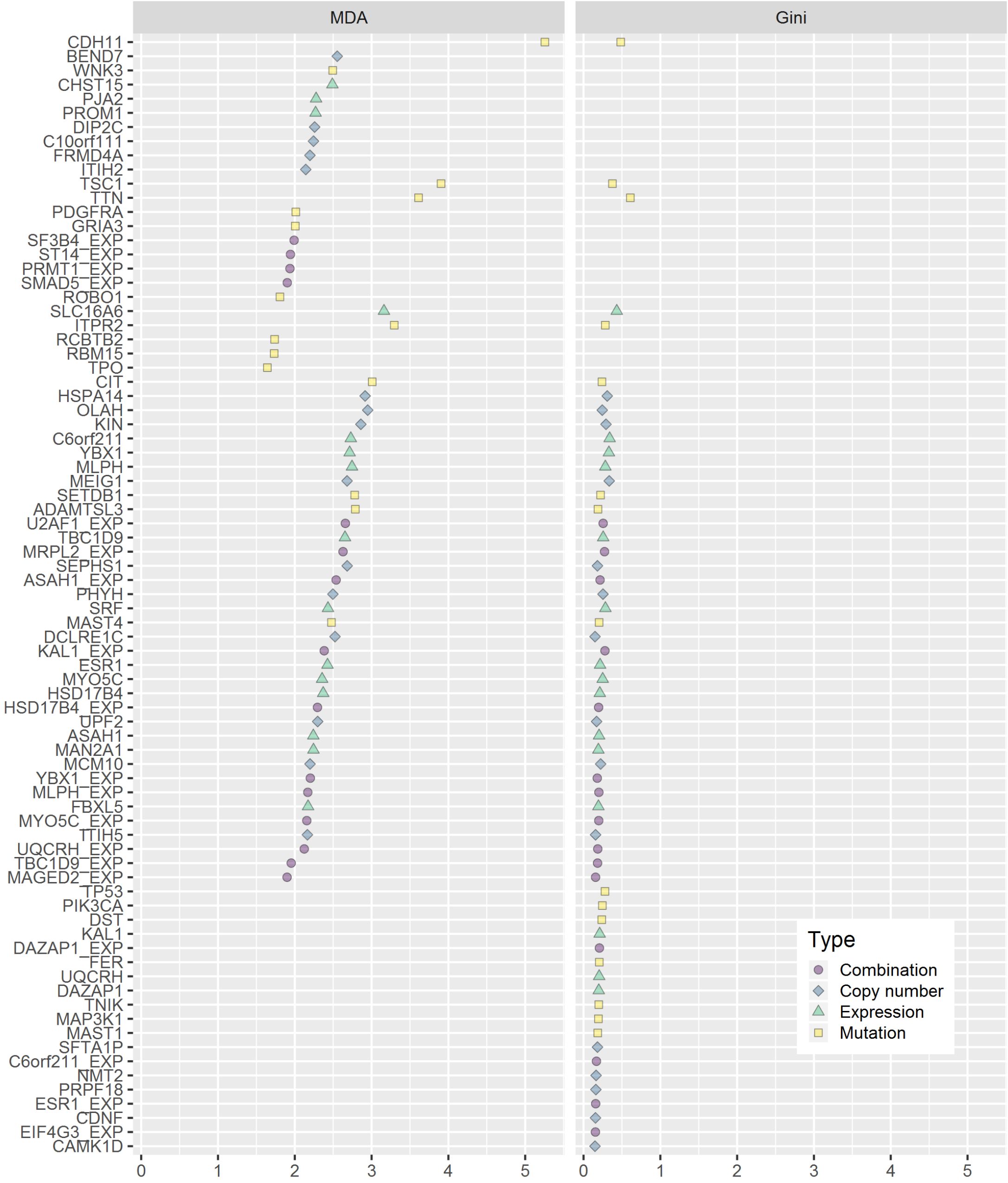


Figure 40: Most important breast cancer features.

necessarily mean it also helps improve the model's accuracy. The low number of features perhaps makes sense, since the mutation based random forest had an AUC of only 0.54, barely greater than chance.

Highest raking feature by importance (Figure 40), *CDH11*, is a tumour suppressor gene varying between more and less invasive breast cancer malignancies (Luo et al., 2013), with role in the development of other cancer types too (Bowles et al., 2007). Other highly important features also have links to various cancer types. *TSC1* is a tumour supressor whose downregulation helps activate mTOR pathway and promote angiogenesis, leading to higher oncogenic suscpetability (Lee et al., 2007). *TTN* is commonly mutated in a variety of cancer types, but this could be more because of its size and less so because of the oncogenic driver potential (Tan et al., 2015). *ITPR2* is long non-coding RNA disregulated in renal cell carcinoma (Blondeau et al., 2011). Lower expression of *ADAMTSL3* and its mutation are a common feature of colorectal cancer (Koo et al., 2007).

Copy number model

The copy number breast cancer classifier proved to be more effective than the mutation based one, with an AUC of 0.83. 674 out of 4000 copy number features have mean decrease accuracy higher than zero, with concentrated importance around 1 MDA (Figure 39).

All of the ten most predictive features are located on 10q13-10q14 and are found almost exclusively in triple negative breast cancer patients. The complete amplification is carried by 2 cell lines, one of which is receptor positive and one triple negative.

With the large number of genes in this region, it it not immediately clear which gene might be contributing to the phenotype. The most important feature, *OLAH*, is a hydrolase playing a role in lactation (Heinz et al., 2016), but with no cancer links to my knowledge. Upregulation of *HSPA14* is correlated with poor prognosis in hepatocellular carcinoma (Yang et al., 2015). A meiosis gene, *MEIG1*, has no strong links to cancer yet, but aberrations of its copy number was found in ovarian disorders (Ledig et al., 2010). *SEPHS1* is overexpressed in rectal carcinoma when compared to normal tissue (Choi et al., 2011).

Expression model

The expression model was the most accurate for breast cancer. 577 expression features have a greater than zero MDA, out of a total of 4000 features. Similarly to the copy number case, a substantial number of them have MDA of 1 (Figure 39).

Amongst the most important features, triple negative samples are associated with a lower expression of 9 of the top 10 features, with only *YBX1* showing a clearly higher expression in triple negative patients (Figure 41) In cell lines, the expression patterns follow the expected profiles. Interestingly, the cell lines HCC1569, JIMT-1 and HCC1954, which are HER2+ lines, that were classified as triple negative by the classifier, all have expression patterns that are more similar to triple negative patients, with the exception of *ERBB2* (HER2) expression. Such atypical expression pattern can be attributed to those cell lines being isolated from a less representative (for the entire HER2+ population) HER2+ cancer sample.

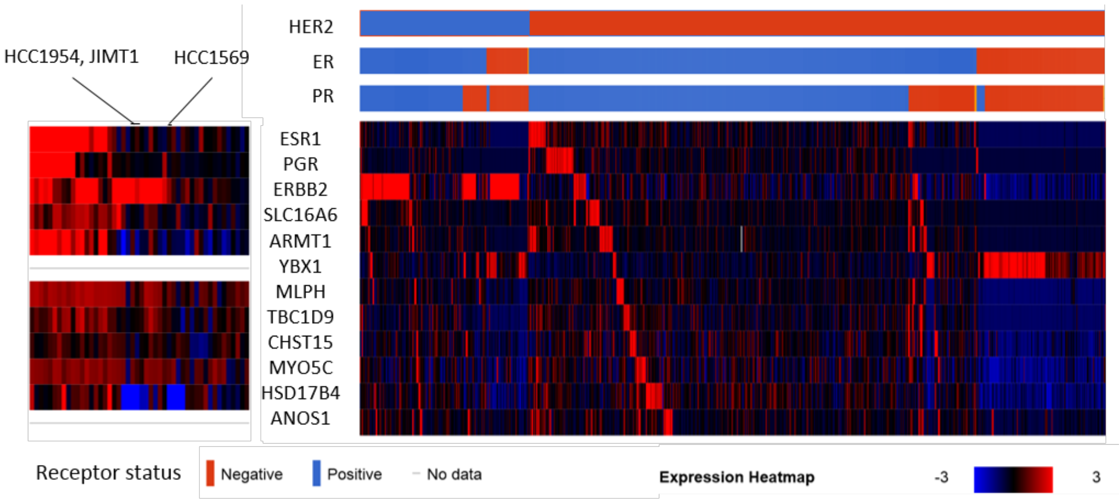


Figure 41: Expression heatmaps of the most predictive genes in the expression classifier for breast cancer showing: right panel - cell lines and left panel cancer patients. Receptor status by IHC are shown for patients on top. Location of HER2+ triple negative classified cell lines is marked. Figure produced by cBioPortal website.

Among the top features, *SLC16A6* is noticeably more important than the others with MDA of 3.16 next to the runner-up *MLPH* at 2.74 MDA and the top 10 mean of 2.6 (Figure 40). Variation in *SLC16A6* is associated with breast and cancer risk (Haiman et al., 2013) and it shows a lower expression in triple negative patients.

Other highly important genes have cancer connections as well. MLPH is associated with resistance to panitumumab treatment for colorectal cancer (Barry et al., 2016) and microRNA activity which differentiates between lung cancer subtypes (Molina-Pinelo et al., 2014). Of all the top ten expression genes, it shows some of the most strikingly triple negative-like expression patterns for the HER2 expressing cells classified as triple negative. *YBX1* is a gene for an RNA-binding protein highly expressed in multiple cancers (Goodarzi et al., 2015). Alterations in *C6orf211* and *ESR1* (gene coding for estrogen receptor) expression levels are correlated to breast cancer risk (Dunbier et al., 2011). Methylation of *HSD17B4*, and consequently its expression, is a marker of HER2 positive breast cancer (Fiegl et al., 2006).

Interestingly, the other receptors, *ERBB2* and *PGR* were not among the more predictive genes.

Joined model

In the joined features model, 641 out of 8425 features had greater than zero importance - 517 expression, 124 copy number and zero mutation features. All the most important features are expression values with the most important copy number feature (*CDK5RAP3*) 76th in the MDA importance ranking. Greater reduction in the importance of copy number (from 4000 to 124) and mutation (from 428 to 0) features than in expression features (4000 to 517) shows gene expression is much more predictive in breast cancer when it comes to the receptor-based subtype. This is expected since the receptor-based subtype vary in pressence of hormone receptors which are tied to their gene expression levels. Subtypes with no expresison of the genes coding for the receptors will not have those receptors. Top 10 copy number features appear only after 114th place.

Interestingly, the top 10 copy number features in the combined model do not match the ones from the copy number only model, suggesting some sort of a correlation to the features from other single data type models. For example, a gene expression feature might be highly correlated, but also slightly more predictive, than a certain copy number feature. In that case, the gene expression feature would be detected as important, but because the copy number feature does not improve the prediction next to it, the copy number feature would end up not important in the model despite being biologically meaningful. The only overlapping copy number features, deemed highly important among copy number features in both models, are *MEIG1* and *HSPA14*.

The overlap of most important features with the expression only model is much better with the top 10 features in both cases containing *MRPL2*, *YBX1*, *HSD17B4* and *MYO5C*.

4.4 Differential expression and pathway analysis of misclassified cell lines

Potential biological causes of misclassification were investigated through differential expression and gene set analyses of the cell lines with respect to the subtype group they belong and the one they were assigned to.

Based on the literature, most cell-lines were assigned receptor statuses and classes (see previous chapter). If a cell-line was supposed to be of one class (triple-negative or receptor-positive) based on the literature, but the classifier assigned it a different score, it is considered misclassified.

In the case of breast cancer the cell-lines JIMT-1 and HCC1569 were both assigned the triple-negative class but express the HER2 receptor according to both the literature and also RNAseq data (Figure 42).

Based on their receptor status and classifier score, cell-lines were grouped into triple-negative, receptor-positive, misclassified and HER2+ cell-lines which were properly classified.

Differential expression analysis using DESeq2 (Love et al., 2014) identified differentially expressed genes in various combinations of groupings (triple-negative - receptor-positive, misclassified - other HER2, misclassified - triple-negative, misclassified - all other) (Figure 43), resulting in many more differentially expressed genes between the HER2+ triple negative classified cell lines than these lines and the triple-negative cell lines (FDR < 0.05, logFC > 1). It is immediately apparent from these numbers the scale of the expression differences differ between the HER expressing triple negative cell lines and other HER2 expressing cell lines. The top changed genes in either direction are shown in (Figure 43E). Interestingly in a hierarchical clustering of the differentially expressed genes, HCC1569 and JIMT1 are outlying branches and do not appear any more similar to triple negative lines than to receptor positive ones. However the number of differentially expressed genes between these cell lines and the triple negative cell lines is far fewer than with the receptor positive ones.

Enrichment analysis was performed with SPIA R package (Tarca et al., 2013) to obtain KEGG pathways (Kanehisa et al., 2016) of the results of the differential expression analysis to find the common themes in the genes that distinguish the misclassified cell lines. The Malaria (Figure 44) and cytokine-cytokine receptor interaction networks (Figure 45) both appear to be altered in triple-negative and misclassified HER2+ cell lines as opposed to receptor positive cell lines or only HER2 positive lines. Cytokines, as a component of the immune system, are often involved

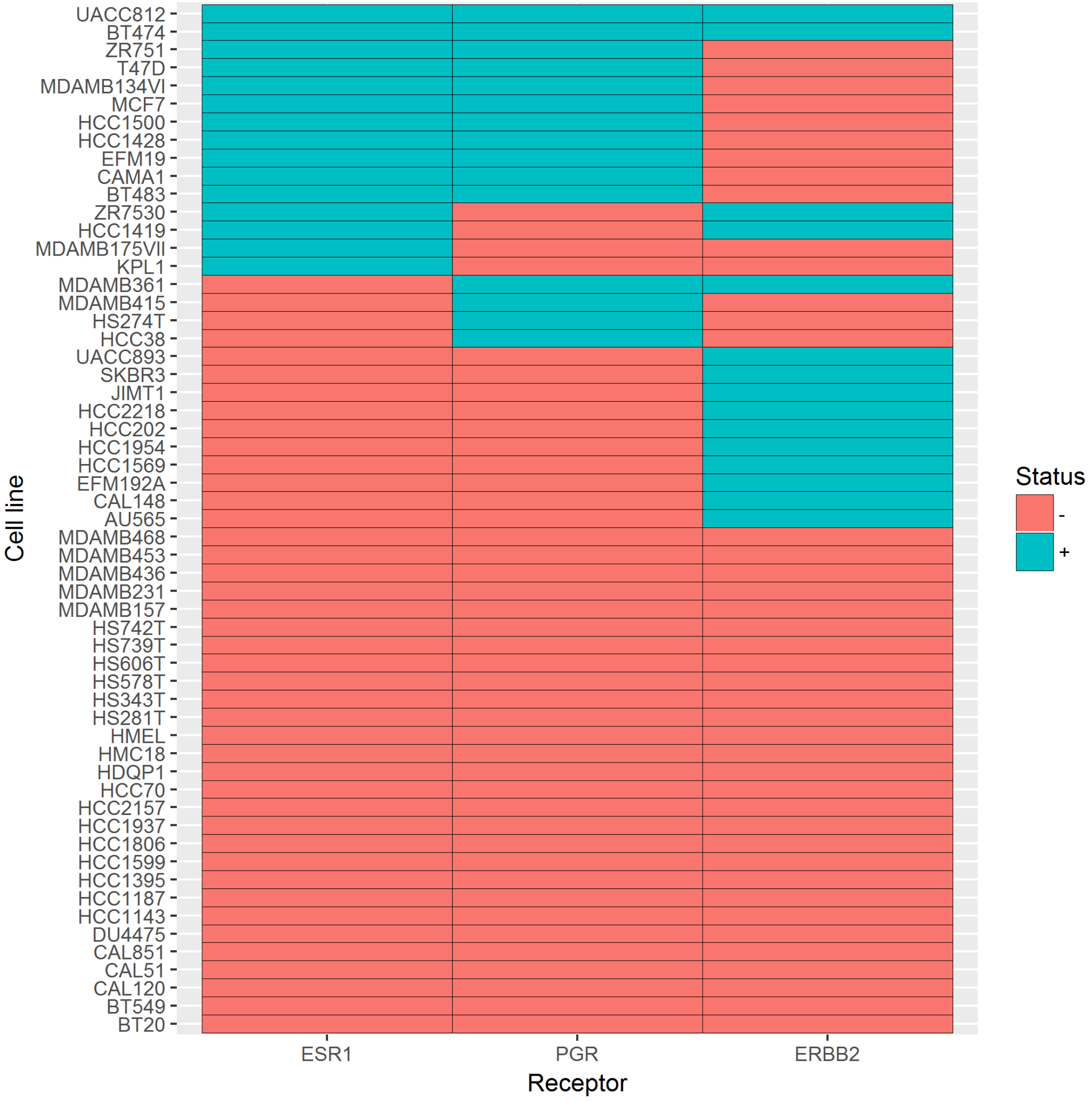


Figure 42: Receptor expression in breast cancer cell lines. ERBB2 (HER2), ESR and PGR are genes coding for HER2, estrogen and progesterone receptors which can determine the breast cancer subtype.

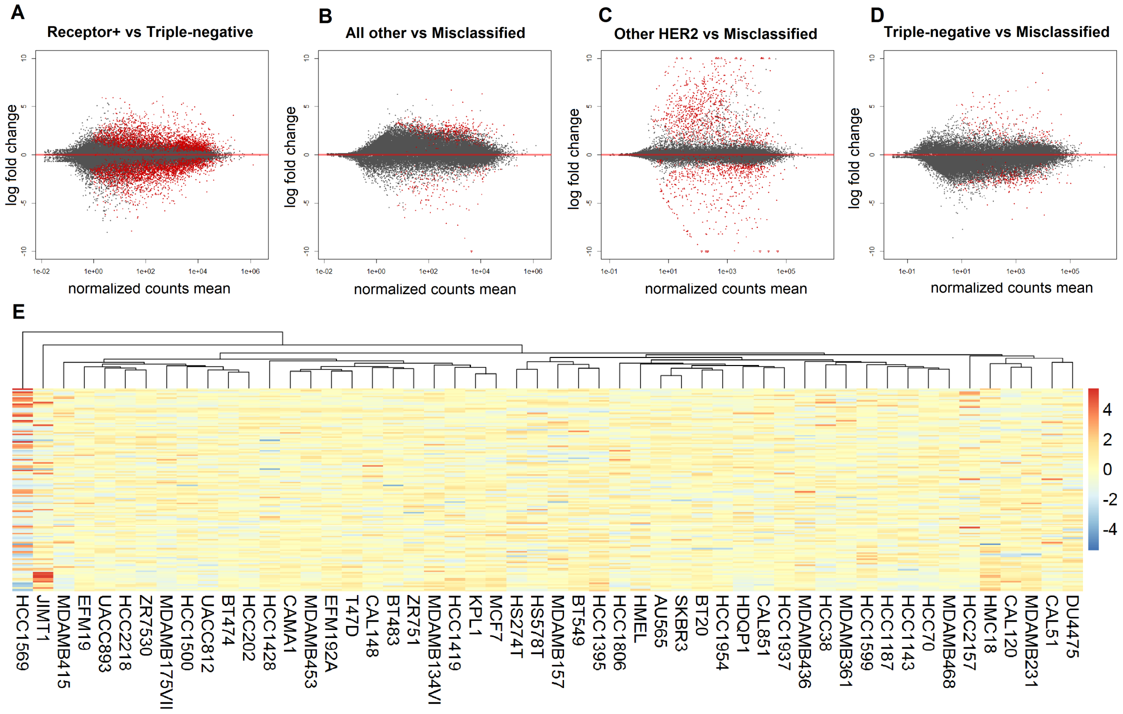


Figure 43: Differential expression analysis for breast cancer cell line groups. MA plots for all analysed groupings (A - D) and heatmap (E) of the genes which were found to be significantly differentially expressed. Number of samples per group are 25 for receptor positive, 22 for triple-negative, 2 for misclassified and 6 for other HER2. MA plots are showing shrunken log fold changes in order to reduce noise normally present due to low count genes.

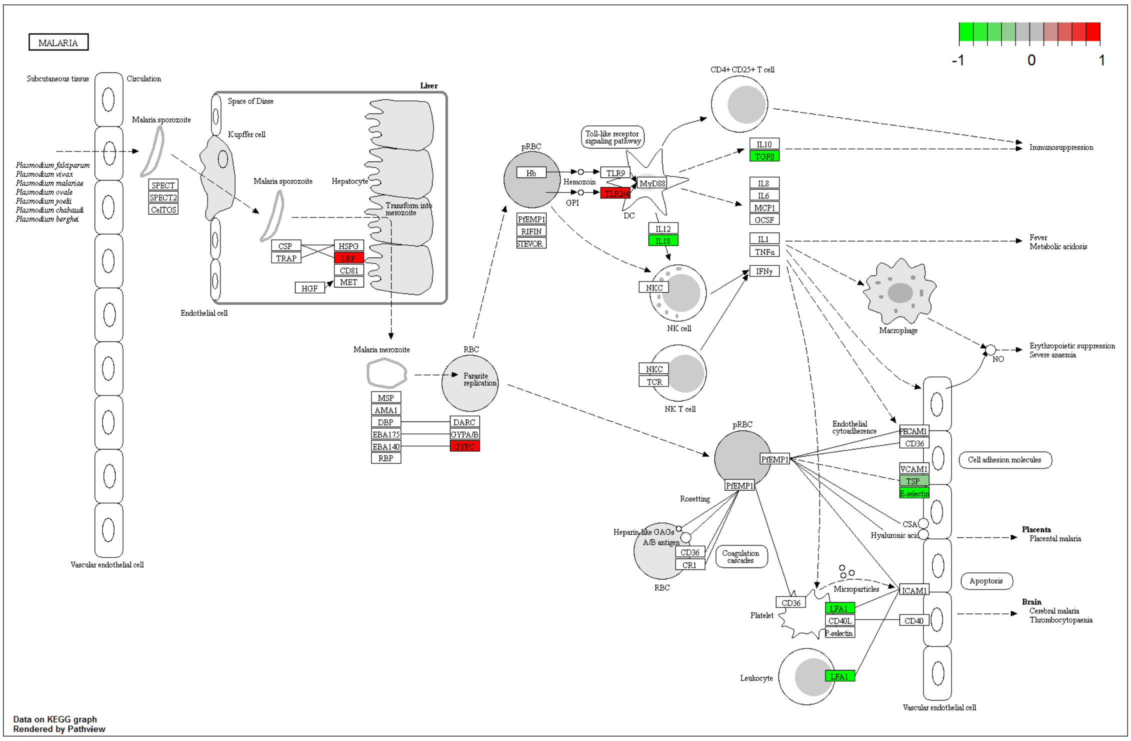


Figure 44: Alterations in Malaria pathway of the misclassified breast cancer cell lines. Colour represents downregulation or upregulation of the differentially expressed genes. Image was generated wiuth SPIA package from the KEGG database.

in the organism's response to cancer. Although Malaria seems unrelated, many of the genes annotated as part of the malaria pathway are relevant to cancer, notable TGFβ, Toll-like receptors, and a variety of interleukins. It has been previously noted that some key pathway elements are relevant for cancer leading to repurposing of Malarial drugs for cancer treatment (Das, 2015). TGFβ signalling has been associated with an increase in cancer-stem cell like behaviour in an IL-8 dependent manner and subsequent resistance to treatment.

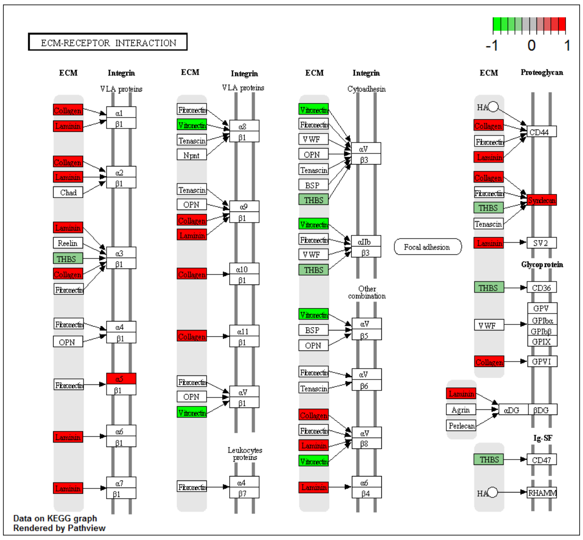


Figure 45: Alterations in the cytokine pathway of misclassified breast cancer cell lines in comparison to other receptor-positive ones. Colour represents downregulation or upregulation of the differentially expressed genes. Image was generated wiuth SPIA package from the KEGG database.

Endometrial cancer

Because the literature assigned classes for the endometrial cancer were type 1 and type 2, while the classes used in the classifier were endometrial and serous, identifying where cell lines were expected to be classified is more difficult. The endometrial cancer cell lines were AN3-CA and HEC-108 are both type-2 cell lines, but were classified as endometrial by the random forest. Further on they are referred to as "misclassified", although they are not in the true meaning of the word like JIMT-1 and HEC-1569 in the breast cancer case since the "mismatch" is type 2 cell line being endometrioid when most type 2 cell lines are scoring closer to the serous side.

There were notable differences in expression between the literature assigned, the score-based grouping and the misclassification-based groupings (Figure 46), although these differences were less pronounced than in the breast cancer case. The only pathway significantly (adjusted $p < 0.05$) altered in misclassified cell lines as opposed to other type-2 cell lines was the Pantothenate and CoA biosynthesis pathway.

This in addition to unusual pair in the group (HEC-108 cell line is part of the HEC family of cells (Kuramoto and Nishida, 2003), unlike AN3-CA which is a poorly differentiated metastatic cell line) makes the meaningfulness of this group questionable.

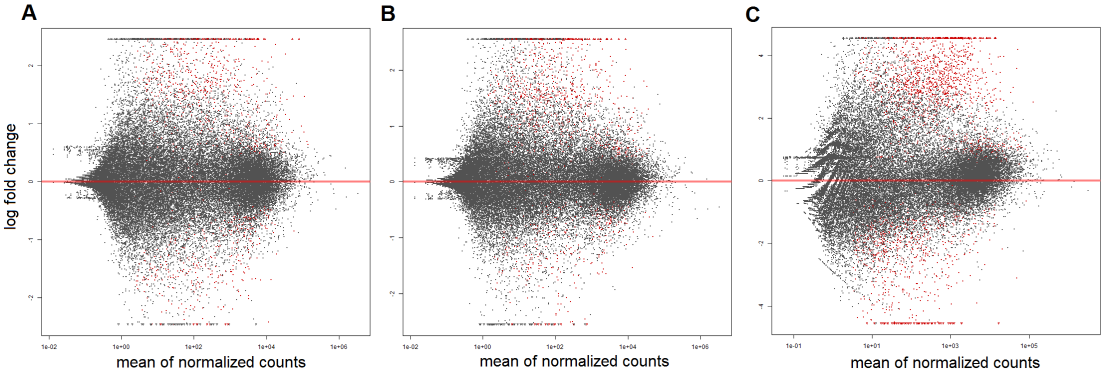


Figure 46: Endometrial cancer MA plots for differential expression of the following groupings: A -Sscore-based (cell lines scoring as one type vs cell lines scoring as the other type), B - lLiterature-based (cell lines scoring as type 1 vs ones scoring as type 2), C - Misclassified cell lines vs other type 2 cell lines. The unusual banding pattern might be the result of there only being 2 misclassified cell lines.

4.5 Cell line drug profiles

Genomic profiles of cancer samples can be used to predict how effective a potential treatment will be. In breast cancer the effectiveness of drugs depends greatly on the receptors expressed in the malignancy, with tumours positive for a particular receptor generally being sensitive to drugs that target that receptor. However, the HER2+ cell lines that are classified with the triple negative cell lines show characteristics that might suggest a resistance to treatment (the previously mentioned TGFβ signalling disregulation associated with resistance, being grouped with the more resistant subtype and, most importantly, drug resistance to common HER2+ drug as discussed here).

In order to address this, data concerning the response of cell lines to anti-tumour drugs was examined. The Genomics of Drug Sensitivity in Cancer (Yang et al., 2013) is a Wellcome Sanger Institute repository containing datasets related to drug research in cancer. One dataset of interest is effectiveness of drugs on cancer cell lines. It contains measurements of drug effectivness for a range of cell lines in the form of $lnIC_{50}$. The IC_{50} value is the amount of drug needed to inhibit half of the cell population.

Cell lines were broken down into triple negative, HER2+, other receptor positive and cell lines which expressed HER2 but were classed as triple negative and examined the IC_{50}s (Figure 47). In several cases the “misclassified” cell lines look more like either triple negative or or receptor positive cell lines than HER2 cells. For example HER2+ cells generally have a lower IC_{50} for CP724714 than other groups (meaning they are more sensitive). However the cell lines classed as triple negative despite expressing HER2 were no more sensitive to this drug than triple negative or cells that expressed other receptors.

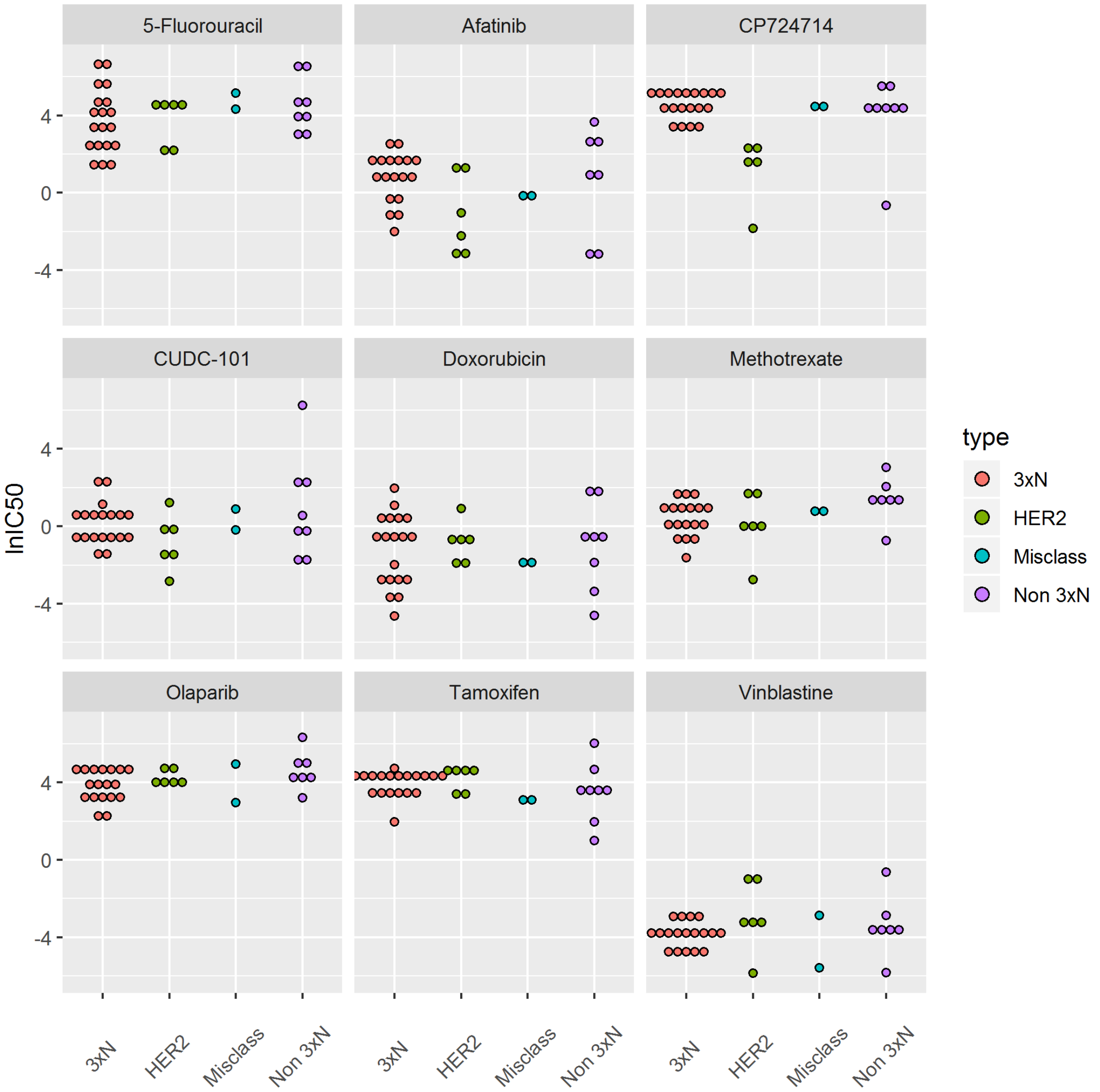


Figure 47: $lnIC_{50}$ of drugs for cell lines per classification group. Shown are breast cancer drugs which had non-missing values for the misclassified cell lines. The different groups are cell lines with respect to receptor status and misclassification by the classifier.

Given the low number of cell lines in each group, especially those I am interested in, it is difficult to say for each drug separately whether the difference between HER2+ but triple negative cell lines is significant. However the fact that the difference holds across many drugs suggests that the effect was real.

To address this, a measure of the distance a cell line's drug profile is from the mean of each group was devised (see methods). By looking at whether a cell line is closer to the mean profile for one group or another we can determine if the drug responses of the triple negative HER2+ cells is more like that of other HER2+ cells or non-receptor expressing cells. All non-receptor expressing cells lines had drug response profiles that were more like the triple negative group mean than the HER2+ group mean (Figure 48). Conversely all HER2+ cells that were classified as HER2+ showed a response closer to the HER2+ group mean than that for triple negative cells. However, both HCC1569 and especially JIMT-1 showed a response that was closer the mean triple negative drug profile than the HER2+ one. Unfortunately data were not available for HCC1954. Interestingly, the cell line with the biggest difference between distance to the triple negative mean and the HER2+ mean was MCF7, being much closer to the triple negative mean than the HER2+ one. This cell line was even closer to the triple negative mean than the general receptor positive mean.

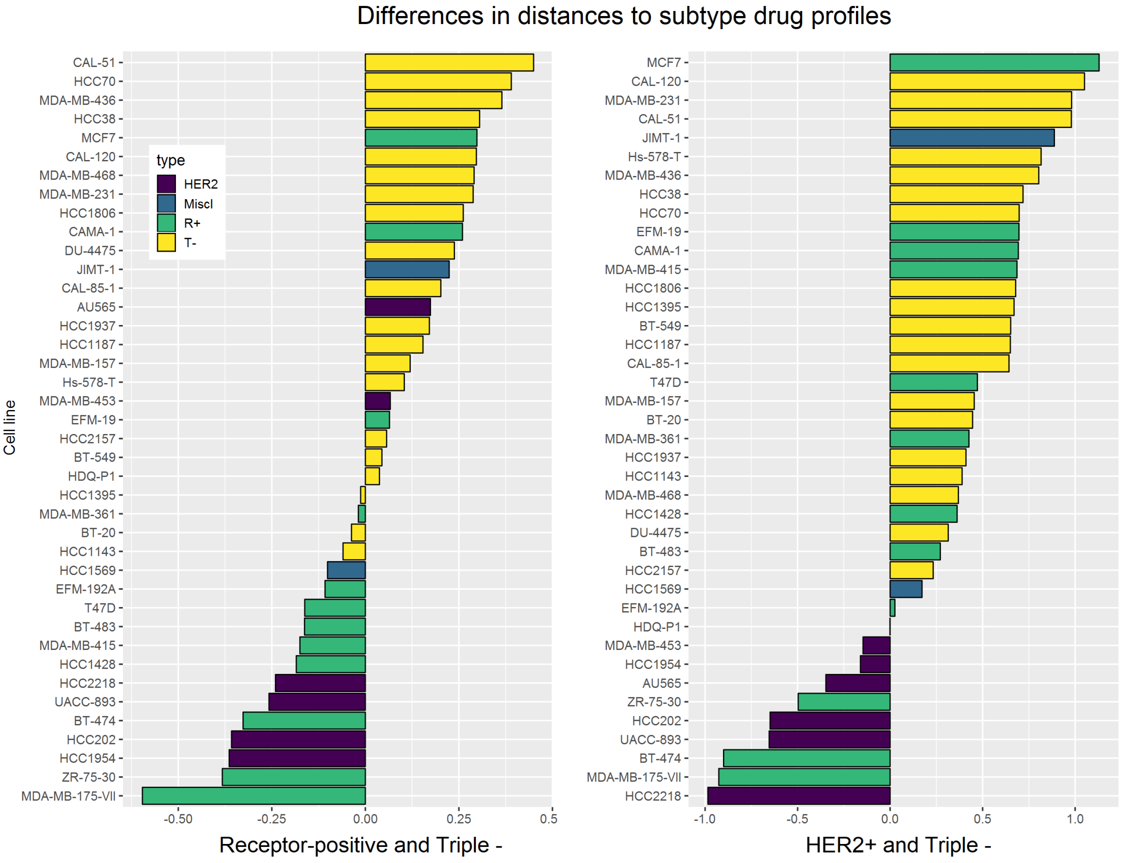


Figure 48: Difference in distances to drug profiles of subtype centres. A - difference between distances to receptor-positive and triple-negative subtype centres, B - between HER2 and triple-negative.

4.6 Discussion

Features picked up as important by the random forest classifier tend to be linked with an oncogenic mechanism. The majority of the features used in the random forest models here have 0 importance score, meaning they are completely irrelevant for the decision-making process of the model. Given the ability to detect features linked to causative oncogenic effects and to easily discard a large number of features while still maintaining accuracy (all of the models used in this chapter are the same as the ones having high accuracy in the earlier chapters), I conclude this strategy of investigating the significance of genomic features is an apt strategy which could contribute significantly to cancer subtype biomarker exploration and analysis if it was to see more usage.

Differential expression analysis detected, for their subtype, atypical pathway activation in the misclassified cell lines (JIMT-1, HCC1569). This affirms the biological relevance of the computational model and potentially defines a new niche within an existing subtype. Anti-malarial drugs have been repurposed as cancer therapeutics (Pascolo, 2016), suggesting that this might be a way of detecting promising new treatments. This could be especially relevant for cell lines and patient samples resistant to standard treatments.

HER2-positive cell lines which were assigned triple-negative by the classifier showed drug response pattern atypical of the other HER2-positive cell lines. One of those cell lines, JIMT-1, has been found to be a Herceptin-resistant HER2-positive cell line. Herceptin is a drug used to treat HER2+ breast cancer. This suggests that the patient-based classification models I used, in some cases, misclassified a cell line that is extremely unrepresentative of its original classifications–HER2+/ER-/PR- cell line that is unresponsive to HER2-targeted treatments.

While this strengthens the validity of my approach, it also provides an alternate view of its applicability. Instead of finding the cell line that scores highly for a subtype in order to obtain good represetative models for it, finding a cell line that is unsuitable for the type its supposed to somewhat resemble helps eliminate that cell line from the pool of cell lines used for that subtype and potentially repurposes it. JIMT-1 is rather undesirable as a receptor-positive or HER2-positive model, but it seems a great fit for a treatment resistant model.

5 Scoring tool

5.1 Introduction

In the previous chapters I set out to add to the information on cancer cell lines as models for the original cancer, hopefully making it easier to select the best cell line in preclinical cancer research.

Research results and techniques developed which could have great impact in different scientific and professional communities must, aside from their inherent scientific use, also be available and known to the wider audience if they are to make significant contribution.

While publicly available software packages (such as ones on Bioconductor (Huber et al., 2015)) allow the wider community to gain access to cutting-edge applicable research methods, they can often be accompanied by a rather steep learning curve.

Online tools, such as NCBI's BLAST (Johnson et al., 2008) or DAVID (Huang et al., 2009), which do not require a certain level of programming ability and can be easily accessed without needing to go through, a potentially complicated, installation procedure can be used by an extremely large and diverse set of researchers.

Enabling easy access to my cancer cell line scoring methods would increase the practical utility of the developed methods and provide an additional feedback venue for method tuning. While the current available options might make the scoring method suboptimal for certain research settings (e.g. the need to provide a binary classification scheme or the scheme being constrained with the TCGA's dataset), the ease of use and availability would still make it useful to researchers wanting to know more about the cell lines of interest before they start an experiment.

Here, I present a Shiny (Chang et al., 2018) application which enables anyone to use the core of the methods I have described in the previous chapters.

Shiny applications are interactive online applications which have proven to be a popular way of presenting data and analysis through a web-based interface due to its relatively simple development and integration with R language (R Core Team, 2018).

They are commonly used for a variety of data visualisations–some interesting examples include genome browser (https://shiny.rstudio.com/gallery/genome-browser.html), tumour growth modelling (http://shiny.webpopix.org/TGImodel/tgi1/) and Ebola epidemiology modelling (https://gallery.shinyapps.io/Ebola-Dynamic/).

CELLector (Najgebauer et al., 2018), the previously mentioned cell line selecting tool, has been deployed as a Shiny application as well (http://ot-cellector.shiny.opentargets.io/CELLector_App/).

5.2 Implementation

I have implemented my application in R using the Shiny framework. Other than the Shiny framework, the dependencies are minimal and are restricted to the `randomForest`, `ROCR` and `tidyverse` packages.

The core requirement of the Shiny app development is the R package of the same name (Chang et al., 2018). While some form of server platform is convenient and typical, the app can be easily be distributed as R source code and then run from the local machine of the user after downloading the source.

One of the strengths of a Shiny app is the ease of development and, consequently, reduction of the time the developer needs to spend on the creation of the app. Thus, it is even more beneficial if software development or computer science are not the creators main skills, as it allows a wide range of professionals to take part in app development despite that not being their primary professional path. This is especially convenient for the research community since the communication of results plays a crucial role in the science domain. Additionally, since it is an R package, and R is commonly used in a variety of research and data science endeavours, it uses a language that the people that would be interested in developing a Shiny app are already familiar with.

All Shiny apps consist of a `ui` and a `server` function. `ui` (user interface) determines the app's visual and organisational properties. How the user interacts with the app is, for the most part, defined in the `ui` function. `server` function contains all the underlying modelling, data processing and other calculations which are then presented through the `ui`.

The core of a Shiny app is shown in the code skeleton below.

```r
library(shiny)

ui <- ...

server <- ...

shinyApp(ui = ui, server = server)
```

`ui` communicates to the `server` through use of "widgets"–user interface elements which obtain input from the user.

`server` commences analyses when it receives input from `ui` and afterwards sends the results of the analyses back to the `ui` to present to the user through a variety of rendering functions, depending on what kind of result is in question.

Several key aspects of the analysis described in the previous chapters are supported in the app:

- building a random forest model through training and testing on the endometrial or breast TCGA patient cancer samples, for the subtypes I defined here, for the chosen data type
- scoring CCLE cell lines with that model
- visualising scores in order to enable easier cell line comparison
- comparing cell line distances to patient samples

Additionally, the user can define their own classification scheme. Being flexible with the classification scheme allows users to go beyond reproducing the analyses presented here, to obtain new data. The default subtypes (endometrioid/serous for endometrial cancer (UCEC), receptor-positive/triple-negative for breast cancer (BRCA)) might not be relevant for the cell line scoring the user would like to know–there might be different subtypes of research interest, but the patient samples and cancer cell lines would still be relevant. In the case of new subtypes, not yet examined in the cell lines, cell line scores from my app can make selecting cell lines much more informed in the context of the new subtype.

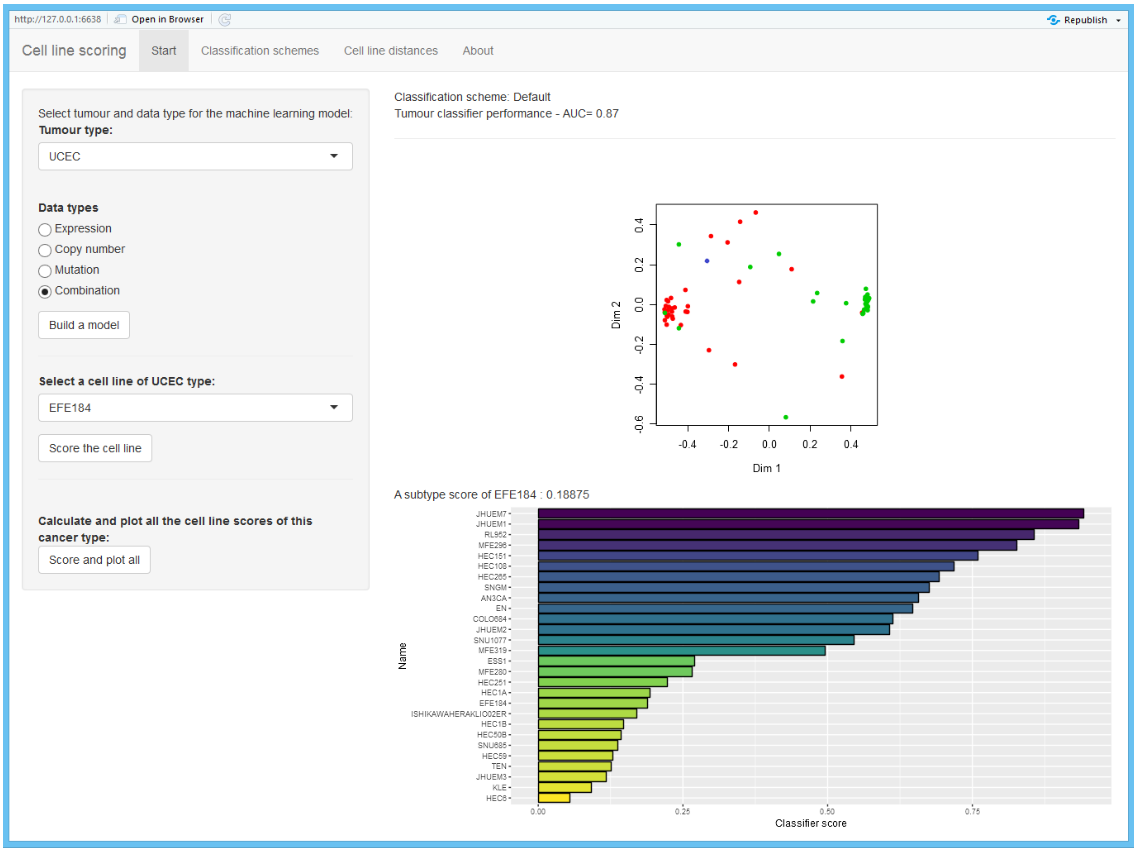
The limitations of the app, in the context of the methodology described in previous chapters, are the lack of feature importance investigation, use of only one machine learning model and necessary limitation of features to just the most variant ones.

However, it is unlikely addition of other methods would be of much merit, as their performance was very similar to the random forest. The number of most variant features chosen was 2000, and with that number of features, random forest classifiers were still highly accurate.

5.3 Usage

The app is segmented into several tabs.

Upon launching, the user is presented with the starting tab of the app (Figure 49). Here, through the menus on the left side, several selections and actions are available. Tumour type and data type selections determine the model which will be build when the `Build a model` button is activated. Following a short period of training and



Select tumour and data type for the machine learning model:

Tumour type:

UCEC

Data types

○ Expression
○ Copy number
○ Mutation
● Combination

Select a cell line of UCEC type:

EFE184

Calculate and plot all the cell line scores of this cancer type:

Classification scheme: Default
Tumour classifier performance - AUC= 0.87

A subtype score of EFE184 : 0.18875

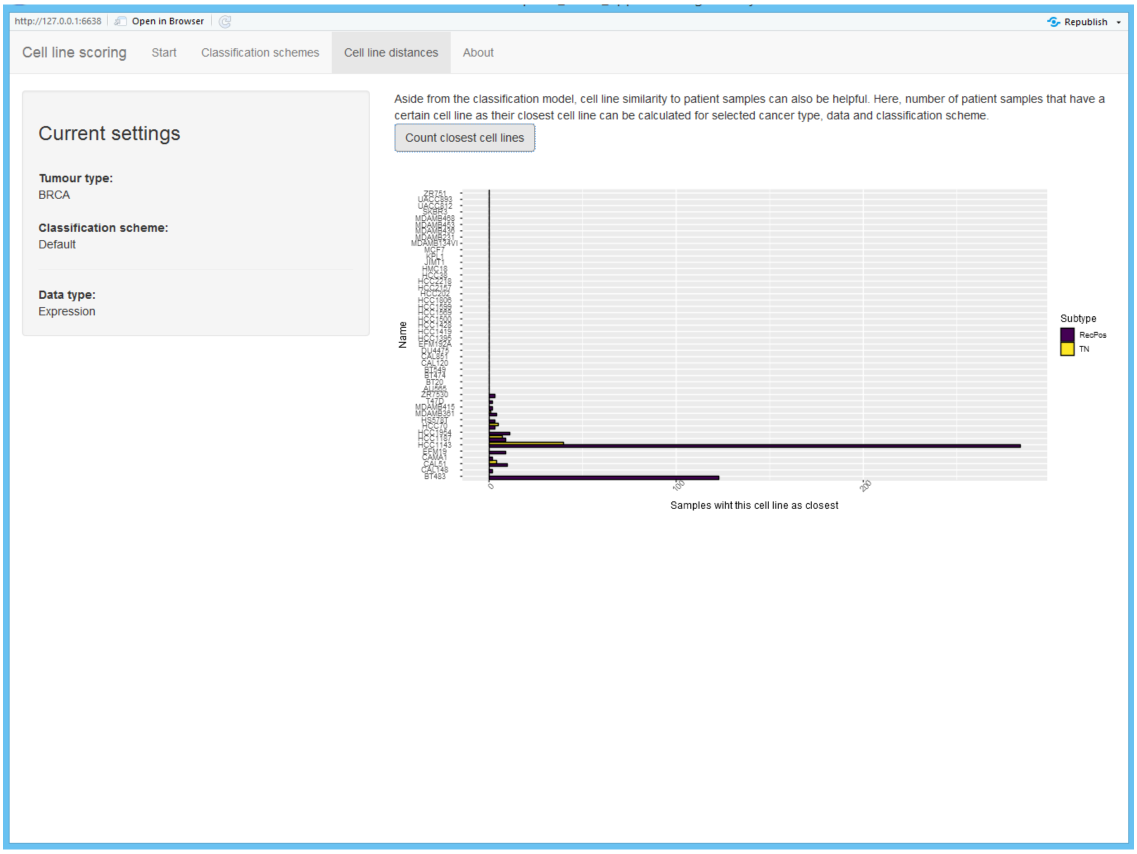
Figure 49: Model building and scoring tab of the application.

testing the model, model's AUC score is shown on the top with MDS plot of the patient samples as produced from the random forest model.

Cancer	Expression	Copy Number	Mutation
BRCA	12.4 sec	10.5 sec	2.0 sec
UCEC	2.0 sec	3.2 sec	2.5 sec

Table 12: Time required to build a model for different cancer and data type combinations. Times are averages of three attempts running on a Xeon E3-1246 v3 based machine

Building a model is fast (Table 12), short enough that providing the necessary compute power on a public server should not be an issue. After the model is built, `Score the cell line` prints a score under for the cell line selected as determined by the active model and places that cell line on the MDS plot above. Scores of all the cell lines, of that cancer type, can be visualised together in the bottom right using `Score and plot all`.



Current settings

Tumour type:
BRCA

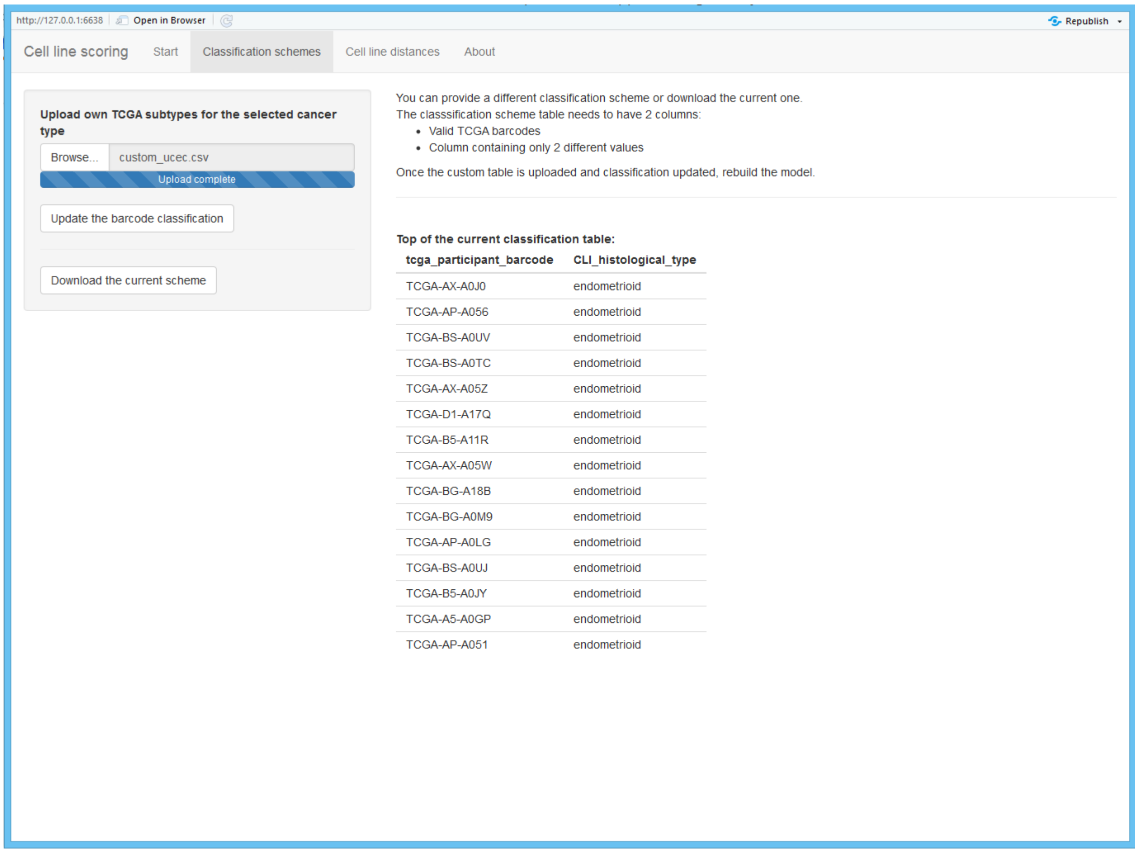
Classification scheme:
Default

Data type:
Expression

Aside from the classification model, cell line similarity to patient samples can also be helpful. Here, number of patient samples that have a certain cell line as their closest cell line can be calculated for selected cancer type, data and classification scheme.

Figure 50: Cell line distances tab has the option to visualise how many patient samples have which of the cell lines as closest, separated by defined subtypes.

The second tab (Figure 51) allows the user to download or upload classification schemes–tables with cancer subtypes for patient samples defined through TCGA barcodes. On the left the user can select a local file to upload, which will become the active classification scheme after pressing the `Update classification scheme` button. This requires the table to be in a compatible format–the first column must be TCGA barcodes (12 characters long as shown in Figure 51) which intersect (at least in part) with the barcodes of the default classification scheme, the second column can contain any type of values as long as there are only 2 unique since the model used is a binary classifier.



Upload own TCGA subtypes for the selected cancer type

custom_ucec.csv

You can provide a different classification scheme or download the current one.
The classsification scheme table needs to have 2 columns:

- Valid TCGA barcodes
- Column containing only 2 different values

Once the custom table is uploaded and classification updated, rebuild the model.

Top of the current classification table:

tcga_participant_barcode	CLI_histological_type
TCGA-AX-A0J0	endometrioid
TCGA-AP-A056	endometrioid
TCGA-BS-A0UV	endometrioid
TCGA-BS-A0TC	endometrioid
TCGA-AX-A05Z	endometrioid
TCGA-D1-A17Q	endometrioid
TCGA-B5-A11R	endometrioid
TCGA-AX-A05W	endometrioid
TCGA-BG-A18B	endometrioid
TCGA-BG-A0M9	endometrioid
TCGA-AP-A0LG	endometrioid
TCGA-BS-A0UJ	endometrioid
TCGA-B5-A0JY	endometrioid
TCGA-A5-A0GP	endometrioid
TCGA-AP-A051	endometrioid

Figure 51: The classification tab allows the user to provide their own classification scheme to be used for model building, scoring and distance calculation, provided it is in the compatible format.

The active classification scheme can be downloaded in the .csv format with `Download the current scheme`. Part of the current classification scheme table are shown on the right side of the tab. Name of the currently active classification scheme is shown in the first and the third tab. The classification scheme used through this thesis is called “Default” while the user-define ones are “Custom”.

The “Cell line distances” tab (Figure 50) shows the currently active app parameters: cancer type, classification scheme and data type used. These can be changed in the first two tabs. The `Count closest cell lines` uses those parameters to determine which cell line is the closest for each of the patient samples, by subtype of the patient samples. It then visualises those findings per cell line, with differently coloured bars corresponding to patient sample counts of different subtype.

5.4 Availability

For the purpose of hosting my application on a publicly available server, I have used the official Shiny applications hosting website (https://www.shinyapps.io/).

Currently, the application is hosted on https://cell-line-scoring.shinyapps.io/clscoring/. Additionally, all the source code and data used (which cannot be accessed through the shinyapps.io version) are available on https://github.com/vuzun/scoring_tool. This mode of access also allows the user to download the code on their local machine and alter it. However, it does require some knowledge of R and version control.

5.5 Discussion

Shiny applications are easy to use and enable presentation of research results to a wide range of audience. The goal of this chapter was to abstract away complicated modelling aspects and still produce useful information. Implementing key aspects of the developed methodology of this thesis as a Shiny application provides quick and easy cancer cell line evaluation. Preclinical cancer researchers can gain insight into suitability of available cancer cell lines, thus easing their decision making process–whether it is confirming their choice is right or guiding them towards potentially interesting cell line options.

To my knowledge, currently, there is only one comparable tool–CELLector (Najgebauer et al., 2018). While the core goal of my app and CELLector, evaluating cancer cell lines as models of patient disease through analysis of patient samples, is shared, there are several difference between the two.

CELLector subsets patient samples of interest through known cancer-related genomic features, with the user-defined restrictions on the features. This allows the user to focus in on the specific genomic alterations they are interested in. The patient set is subsetted hierarchically and cancer cell lines are then mapped to the corresponding subsets, depending on the matching genomic features of interest. This enables identification of patient sub-cohorts which lack a representative cancer cell line model.

While the Shiny application presented here lacks most of those options, it uses all of the available genomic features (time-complexity permitting), casting a wider net on the genomic landscape. While it does need to have a classification scheme, unlike CELLector, it does not rely on using only known oncogenic elements. This is an important distinction because often times the molecular drivers of a particular medically relevant characteristic, such as tumour aggressiveness, metastasis stage, or resistance to a particular treatment are not known. This classification scheme allows the user to gain subtype-specific information about the cell line without first knowing the molecular drivers.

Although not part of the app, an important property of my methodology, lacking in CELLector as it does not assign a subtype, is the ability to examine the importance of genomic features in the context of a subtype. For a newly discovered subtype, or a subtype which has seen an increase in the available data, this can lead to highlighting novel biomarkers. It will be important for future iterations of the app to include this functionality.

Other future additions that might be useful include the addition of further cancer types, the ability to include more than a single cancer type in an analysis (to allow the detection of cell lines assigned to the wrong cancer type) and the ability to update cell line expression, copy number and mutation data for cell lines beyond those included in CCLE. As it is I believe my app provides the first online web tool for selecting a cell line to represent an arbitrary subtype division of cancers, and as such should prove useful to the preclinical research community.

6 Discussion

Preclinical cancer research relies heavily on cancer cell lines to model patient disease and there is no sign this will change. Using cancer cell lines as a proxy for patient samples has a variety of problems which researchers have been aware for a long time. Systematic cell line scoring based on patient samples has seen few incarnations, none of which utilised classification methods. I have explored one way of doing this, and, although it has the standard classification limitations, providing context through classification scheme has strengths not found in other cell line evaluation efforts.

There have been several attempts to address the quality of cell lines, however,to my knowledge, there were few approaches (Najgebauer et al., 2018) that focused on a broad range of genomic features and used computational models to automate decision-making. These methods have tended to use features of predefined importance instead of allowing the data to speak for itself.

I set out to create a systematic evaluation methodology for cancer cell lines, through application of machine learning methods, based on patient cancer samples and their subtypes. Using TCGA data, I have shown capabilities of several machine learning methods. High accuracies of predicting cancer subtype in new patient samples were evidence of the suitability of my method choices. Although several methods performed similarly, I chose random forest as my focus due to its interpretability and ease of use. It is interesting to note that the method based on PCA worked almost as well, while providing additional information on patient sample to cell line distances. Although in both cases classifiers based on single data types have achieved high accuracy, a large subset of their features was not used by the combination classifiers (which were more accurate than the single data type ones), suggesting potential issues with feature redundancy, but also providing a way around that via data type combination classifiers. Thus, remarkable accuracy was achieved in predicting cancer subtype from the provided data, as measured by AUC (0.98 and 0.941 for weighted combination endometrial and weighted combination breast models) and F_1 (0.92 and 0.877 for the same models) metrics.

Applying trained models to cell lines, I generated scores for endometrial and breast cancer cell lines which, in general, matched the literature for the chosen subtypes, albeit with a few exceptions. Cell lines which had no notable literature or had ambiguous descriptions were placed on a score range, potentially helping their future use.

In addition to patient sample-trained classifiers, as a way of evaluating cancer cell lines, I explored tailored distance metrics. Using the variable importances from the random forest classifier as a filter in the distance calculation between cell lines and patient samples allowed me to tune out, for the defined subtype, irrelevant features.

Interesting patterns emerged. The classifiers built identified an E1 Cyclin containing amplification almost exclusively present in serous cancer patients. Only 5% of endometrioid samples were affected. It is also found in KLE–the cell line scoring as the most serous one.

Although no cell lines in the CCLE collection are annotated as serous, KLE might be one. It has amplifications characteristic of serous subtype (CCNE1) and is also close to patients when looking at the average distance. More KLE research in the context of these subtypes would be valuable to confirm it is not just a overmutated cell line accidentally exhibiting serous alterations.

Some cell lines have genomic profiles overall similar to both subtypes. Aside from T-47D, endometrial cell lines MFE-280 (high-grade endometrioid), SNU-1077 (an endometrial carcinosarcoma) and MFE-319 (low-grade endometrioid) cover a lot of serous and endometrioid patient samples, judging by genomic profile similarity.

The exceptions to the trend of classifier score matching literature assignment lead me to HER2 cancer cell lines with atypical pathway activation–HCC1569 and JIMT-1. JIMT-1 is a HER2-positive breast cancer cell line which was established from a patient resistant to Herceptin–drug used for HER2-positive breast cancer (Tanner et al., 2004). These cell lines were the closest cell lines for a subset of patients, suggesting that their profiles do represent a real cancer phenotype. Given the the similarity of these cell lines to triple negative cell lines, both in their expression profiles and their response to drug treatment they are a good candidate model for studying receptor positive cancer that are resistant to receptor targeting treatments.

MCF-7, one of the most commonly used breast cancer cell lines, was not the closest cancer cell line for any of the hundreds of breast cancer patient samples. The mutational profile of a commonly used, receptor-positive scoring, reportedly oestrogen and progesterone positive, cancer cell line–T-47D–is one that is the most similar for an overwhelming majority of triple-negative patient samples. Still, when it comes to expression and copy number, it is the most similar cell line to receptor-positive patient samples.

Similarities through copy number, mutation, expression or combination profiles are rarely consistent for the analysed cell lines, stressing the importance of -omics integration.

While the classifier scores had no spectacular outliers, this kind of background of some of the misclassified cell lines suggests the classifier approach is more effective than it might seem at first glance, especially when the background of the classification results are investigated.

Recognising the importance of research communication, I have implemented core aspects of my methodology as an easy-to-use, online tool–allowing researchers to score their cell lines and potentially making their decision-making process more informed. Since the tool has the ability to use user-define classifications, it has wider use in subtype-based scoring than the subtype-specific analyses I have conducted here.

Other tools aiming to evaluate cancer cell line using patient cancer samples exist. TumorComparer, a weighted approach (Sinha et al., 2015) directly comparing genomic profiles of cell lines and patient samples, not too unlike the distance metrics I have employed, showed promising results on a wide variety of cancer types, but is still unavailable to the public.

CELLector (Najgebauer et al., 2018) subsets, based on oncogenic genomic alterations, the patient cancer sample set in order to distinguish between different subcohorts, and afterwards match the cell lines to the subcohort of the same genomic profile. It is reliant on the known drivers, but it provides a quick overview of the landscape of the selected cancer type and identifies subcohorts lacking representative cell line models. It is highly accessible, since it is available as an online tool and an R package, and easy to use. Dividing patient samples, and later on cell lines, into many subcategories based on specific known drivers makes it a great tool for selecting a cell line for investigation of a specific genomic alteration.

However, it disregards the general genomic features which are available for both the patient samples and cancer cell lines. While these are less likely to be cancer-defining (since they are not already identified as drivers), they might play a role in a less studied or a new cohort and be disregarded by this kind of approach. Not using a classification approach allows CELLector to identify as many subcohorts as the user wants, giving it more flexibility in what the cell lines can be assigned to. If there are only two classes which a cell line can be assigned to, they could be assigned closer to one class only because they are far away from the other one. Selecting a classification approach might be desired in some cases–if there is a novel, unexplored cancer subtype, having it as one of the classes will inform the user which cell line is the closest fit. For a less explored cancer subtype, this could quickly identify potential preclinical models. Additionally, investigation of classification parameters of the computational model (e.g. variable importance values in the random forest models) allows for another venue of identification of causative elements. A classification approach can also help compare different classification schemes such as endometrioid/serous with the type 1/type 2 classification scheme in endometrial cancer. With a classification approach, deviations from the suspected class (i.e. subtype) are easier to detect–JIMT-1 mismatch between the classifier and the literature corresponds to it being HER2 positive, but treatment resistant cell line (Tanner et al., 2004). A non-classifying approach like the one utilised by CELLector has no tools to pick up that incongruence.

An alternative to immortalised cell lines for preclinical research is patient derived xenografts (PDXs) and other pre-clinical models. While these escape some of the restrictions associated with immortalised cancer cell lines, and are more likely to be representative of the cancer they are derived from it, they still suffer from some representativeness issues while being far more resource intense when compared to the cancer cell lines. If a systematic genomic analysis of a larger cohort of PDXs or other preclinical cancer models would be conducted, my approaches would be just as applicable to the new dataset, especially so considering the ability to unravel the bakcground drivers of subtype determination.

There are several ways to improve the methodology developed here. Methylation, histone modification and gene fusion data types are becoming more common and would provide an additional data type to tune the models while also allowing the models to determine the importance of their features. Expanding the classification to more than two classes would add much needed flexibility to the subtyping strategy. Expanding the analyses to between cancer types could show which cell lines might be related to a different tissue, or how cancer types might share some cancer cell lines as good models.

In conclusion, I have developed highly accurate models for cancer subtype predictions, leveraged them to assess the clinical relevance of cancer cell lines, thus helping the preclinical research choice of a good model. Additionally, I have identified cell lines which are less representative of the subtype they are commonly cited (e.g. JIMT-1, HER2+ cell line resistant to HER2-targeted treatment) as and identified a potentially new use for them (studying treatment resistant receptor-positive breast cancer). Finally, the core techniques I used have been implemented in an easy to use online tool to allow others to leverage my work.

References

Abdelmoula, W.M., Balluff, B., Englert, S., Dijkstra, J., Reinders, M.J.T., Walch, A., McDonnell, L.A., and Lelieveldt, B.P.F. (2016). Data-driven identification of prognostic tumor subpopulations using spatially mapped t-SNE of mass spectrometry imaging data. Proceedings of the National Academy of Sciences of the United States of America *113*, 12244–12249.

ACS medical and editorial team (2019). What Is Endometrial Cancer? American Cancer Society.

Akhbardeh, A., and Jacobs, M.A. (2012). Comparative analysis of nonlinear dimensionality reduction techniques for breast MRI segmentation. Medical Physics *39*, 2275–2289.

Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., Castro, E. de, Duvaud, S., Flegel, V., Fortier, A., Gasteiger, E., et al. (2012). ExPASy: SIB bioinformatics resource portal. Nucleic Acids Research *40*, W597–W603.

Aure, M.R., Vitelli, V., Jernström, S., Kumar, S., Krohn, M., Due, E.U., Haukaas, T.H., Leivonen, S.-K., Vollan, H.K.M., Lüders, T., et al. (2017). Integrative clustering reveals a novel split in the luminal A subtype of breast cancer with impact on outcome. Breast Cancer Research *19*, 44.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehar, J., Kryukov, G.V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603–607.

Barry, G.S., Cheang, M.C., Chang, H.L., and Kennecke, H.F. (2016). Genomic markers of panitumumab resistance including ERBB2/ HER2 in a phase II study of KRAS wild-type (wt) metastatic colorectal cancer (mCRC). Oncotarget *7*, 18953–18964.

Blondeau, J.J., Deng, M., Syring, I., Schrödter, S., Schmidt, D., Perner, S., Müller, S.C., and Ellinger, J. (2011). Suffocating cancer: hypoxia-associated epimutations as targets for cancer therapy.

Bowles, E., Corson, T.W., Bayani, J., Squire, J.A., Wong, N., Lai, P.B.-S., and Gallie, B.L. (2007). Profiling genomic copy number changes in retinoblastoma beyond loss ofRB1. Genes, Chromosomes and Cancer *46*, 118–129.

Breiman, L. (2001). Random Forests. Machine Learning *45*, 5–32.

Buettner, F., Natarajan, K.N., Casale, F.P., Proserpio, V., Scialdone, A., Theis, F.J., Teichmann, S.A., Marioni, J.C., and Stegie, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nature Biotechnology *33*, 155–160.

Capes-Davis, A., and Freshne, I. (2012). Database of Cross-Contaminated or Misidentified Cell Lines. 1–18.

Capes-Davis, A., Theodosopoulos, G., Atkin, I., Drexler, H.G., Kohara, A., MacLeod, R.A., Masters, J.R., Nakamura, Y., Reid, Y.A., Reddel, R.R., et al. (2010). Check your cultures! A list of cross-contaminated or misidentified cell lines. International Journal of Cancer *127*, 1–8.

Chang, S.-W., Abdul-Kareem, S., Merican, A., and Zain, R. (2013). Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods. BMC Bioinformatics *14*, 170.

Chang, W., Cheng, J., Allaire, J.J., Xie, Y., and McPherson, J. (2018). shiny: Web Application Framework for R. R Package Version 1.1.0.

Chen, Y.-C., Ke, W.-C., and Chiu, H.-W. (2014). Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. Computers in Biology and Medicine *48*, 1–7.

Choi, S.-y., Jang, J.H., and Kim, K.R. (2011). Analysis of differentially expressed genes in human rectal carcinoma using suppression subtractive hybridization. Clinical and Experimental Medicine *11*, 219–226.

Chu, W., Ghahramani, Z., Falciani, F., and Wild, D.L. (2005). Biomarker discovery in microarray gene expression data with Gaussian processes. Bioinformatics *21*, 3385–3393.

Collisson, E.A., Campbell, J.D., Brooks, A.N., Berger, A.H., Lee, W., Chmielecki, J., Beer, D.G., Cope, L., Creighton, C.J., Danilova, L., et al. (2014). Comprehensive molecular profiling of lung adenocarcinoma. Nature *511*, 543–550.

Combest, A.J., Sandison, K., Hanna, S.K., Zamboni, W.C., Habibi, S., Roberts, P.J., Dillon, P.M., Ross, C., Sharpless, N.E., Müller, M., et al. (2012). Genetically engineered cancer models, but not xenografts, faithfully predict anticancer drug exposure in melanoma tumors. Oncologist *17*, 1303–1316.

Cortes, C., and Vapnik, V. (1995). Support-vector networks. Machine Learning *20*, 273–297.

Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y.Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature *486*, 346–352.

Dai, X., Cheng, H., Bai, Z., and Li, J. (2017). Breast cancer cell line classification and Its relevance with breast tumor subtyping. Journal of Cancer *8*, 3131–3141.

Das, A.K. (2015). Anticancer Effect of AntiMalarial Artemisinin Compounds. Annals of Medical and Health Sciences Research *5*, 93–102.

Davis, J., and Goadrich, M. (2006). The Relationship Between Precision-Recall and ROC Curves. In Proceedings of the 23rd International Conference on Machine Learning, (New York, NY, USA: Association for Computing Machinery), pp. 233–240.

Delen, D., Walker, G., and Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine *34*, 113–127.

Domcke, S., Sinha, R., Levine, D.A., Sander, C., and Schultz, N. (2013). Evaluating cell lines as tumour models by comparison of genomic profiles. Nature Communications *4*, 10.

Dunbier, A.K., Anderson, H., Ghazoui, Z., Lopez-Knowles, E., and Pancholi, S. (2011). ESR1 Is Co-Expressed with Closely Adjacent Uncharacterised Genes Spanning a Breast Cancer Susceptibility Locus at 6q25.1. PLoS Genet *7*, 1001382.

Ellison, G., Klinowska, T., Westwood, R.F.R., Docter, E., French, T., and Fox, J.C. (2002). Further evidence to support the melanocytic origin of MDA-MB-435. Molecular Pathology *55*, 294–299.

Ertel, A., Verghese, A., Byers, S.W., Ochs, M., and Tozeren, A. (2006). Pathway-specific differences between tumor cell lines and normal and tumor tissue cells. Molecular Cancer *5*, 55.

European Collection of Authenticated Cell Cultures (2018). ECACC General Cell Collection: COLO 685 (Public Health England).

Exarchos, K.P., Goletsis, Y., and Fotiadis, D.I. (2012). Multiparametric Decision Support System for the Prediction of Oral Cancer Reoccurrence. IEEE Transactions on Information Technology in Biomedicine *16*, 1127–1134.

Fantini, D. (2017). easyPubMed: Search and Retrieve Scientific Publication Records from PubMed. R Package Version 2.3.

Fiegl, H., Millinger, S., Goebel, G., Müller-Holzner, E., Marth, C., Laird, P.W., and Widschwendter, M. (2006). Breast Cancer DNA Methylation Profiles in Cancer Cells and Tumor Stroma: Association with HER-2/neu Status in Primary Breast Cancer. Cancer Res *66*, 29–33.

Gazdar, A.F., Kurvari, V., Virmani, A., Gollahon, L., Sakaguchi, M., Westerfield, M., Kodagoda, D., Stasny, V., Cunningham, H.T., Wistuba, I.I., et al. (1998). Characterization of paired tumor and non-tumor cell lines established from patients with breast cancer. International Journal of Cancer *78*, 766–774.

Gazdar, A.F., Girard, L., Lockwood, W.W., Lam, W.L., and Minna, J.D. (2010). Lung Cancer Cell Lines as Tools for Biomedical Discovery and Research. Jnci-Journal of the National Cancer Institute *102*, 1310–1321.

Getz, G., Gabriel, S.B., Cibulskis, K., Lander, E., Sivachenko, A., Sougnez, C., Lawrence, M., Kandoth, C., Dooling, D., Fulton, R., et al. (2013). Integrated genomic characterization of endometrial carcinoma. Nature *497*, 67–73.

Gevaert, O., Smet, F.D., Timmerman, D., Moreau, Y., and Moor, B.D. (2006). Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. Bioinformatics *22*, e184–e190.

Gillet, J.-P., Calcagno, A.M., Varma, S., Marino, M., Green, L.J., Vora, M.I., Patel, C., Orina, J.N., Eliseeva, T.A., Singal, V., et al. (2011). Redefining the relevance of established cancer cell lines to the study of mechanisms of clinical anti-cancer drug resistance. Proceedings of the National Academy of Sciences of the United States of America *108*, 18708–18713.

Goodarzi, H., Liu, X., Nguyen, H.C., Zhang, S., Fish, L., and Tavazoie, S.F. (2015). Endogenous tRNA-Derived Fragments Suppress Breast Cancer Progression via YBX1 Displacement. Cell *161*, 790–802.

Gozgit, J.M., Wong, M.J., Moran, L., Wardwell, S., Mohemmad, Q.K., Narasimhan, N.I., Shakespeare, W.C., Wang, F., Clackson, T., and Rivera, V.M. (2012). Preclinical Development Ponatinib (AP24534), a Multitargeted Pan-FGFR Inhibitor with Activity in Multiple FGFR-Amplified or Mutated Cancer Models. Mol Cancer Ther *11*.

Grigoriadis, A., Mackay, A., Noel, E., Wu, P.J., Natrajan, R., Frankum, J., Reis-Filho, J.S., and Tutt, A. (2012). Molecular characterisation of cell line models for triple-negative breast cancers. BMC Genomics *13*, 619.

Gunn, S. (1998). Support Vector Machines for classification and regression. Technical Report at University of Southampton.

Hackenberg, R., Hawighorst, T., Hild, F., and Schulz, K.D. (1997). Establishment of new epithelial carcinoma cell lines by blocking monolayer formation. Journal of

Cancer Research and Clinical Oncology *123*, 669–673.

Haiman, C.A., Han, Y., Feng, Y., Xia, L., Hsu, C., Sheng, X., Pooler, L.C., Patel, Y., Kolonel, L.N., Carter, E., et al. (2013). Genome-Wide Testing of Putative Functional Exonic Variants in Relationship with Breast and Prostate Cancer Risk in a Multiethnic Population. PLoS Genetics *9*, e1003419.

Hajisharifi, Z., Piryaiee, M., Mohammad Beigi, M., Behbahani, M., and Mohabatkar, H. (2014). Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. Journal of Theoretical Biology *341*, 34–40.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: The next generation. Cell *144*, 646–674.

Hartigan, J.A., and Wong, M.A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics) *28*, 100–108.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning (New York, NY: Springer New York).

Hay, D.L., Garelja, M.L., Poyner, D.R., and Walker, C.S. (2018). Update on the pharmacology of calcitonin/CGRP family of peptides: IUPHAR Review 25. British Journal of Pharmacology *175*, 3–17.

Hay, M., Thomas, D.W., Craighead, J.L., Economides, C., and Rosenthal, J. (2014). Clinical development success rates for investigational drugs. Nature Biotechnology *32*, 40–51.

Heinz, R.E., Rudolph, M.C., Ramanathan, P., Spoelstra, N.S., Butterfield, K.T., Webb, P.G., Babbs, B.L., Gao, H., Chen, S., Gordon, M.A., et al. (2016). Constitutive expression of microRNA-150 in mammary epithelium suppresses secretory activation and impairs de novo lipogenesis. Development *143*, 4236–4248.

Hellton, K.H., and Thoresen, M. (2016). Integrative clustering of high-dimensional data with joint and individual clusters. Biostatistics 1–14.

Hensman, J., Fusi, N., and Lawrence, N.D. (2013). Gaussian Processes for Big Data. In Uncertainty in Artificial Intelligence - Proceedings of the 29th Conference, Uai 2013, pp. 282–290.

Hofmann, T., Scholkopf, B., and Smola, A.J. (2008). Kernel Methods in Machine Learning. The Annals of Statistics *36*, 1171–1220.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Research *37*, 1–13.

Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B.S., Bravo, H.C., Davis, S., Gatto, L., Girke, T., et al. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. Nature Methods *12*, 115–121.

Hunter, J.D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering *9*, 90–95.

Jamieson, A.R., Giger, M.L., Drukker, K., Li, H., Yuan, Y., and Bhooshan, N. (2010). Exploring nonlinear feature space dimension reduction and data representation in breast Cadx with Laplacian eigenmaps and t-SNE. Medical Physics *37*, 339–351.

Jiang, G., Zhang, S., Yazdanparast, A., Li, M., Pawar, A.V., Liu, Y., Inavolu, S.M., and Cheng, L. (2016). Comprehensive comparison of molecular portraits between cell lines and tumors in breast cancer. BMC Genomics *17*, 525.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., and Madden, T.L. (2008). NCBI BLAST: a better web interface. Nucleic Acids Research *36*, W5–W9.

Kalogera, E., Roy, D., Khurana, A., Mondal, S., Weaver, A.L., He, X., Dowdy, S.C., and Shridhar, V. (2017). Quinacrine in endometrial cancer: Repurposing an old antimalarial drug. Gynecologic Oncology *146*, 187–195.

Kandoth, S., C. (2013). Integrated genomic characterization of endometrial carcinoma. Nature *497(7447)*, 67–73.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. Nucleic Acids Research *44*, D457–D462.

Kao, J., Salari, K., Bocanegra, M., Choi, Y.-L., Girard, L., Gandhi, J., Kwei, K.A., Hernandez-Boussard, T., Wang, P., Gazdar, A.F., et al. (2009). Molecular profiling of breast cancer cell lines defines relevant tumor models and provides a resource for cancer gene discovery. PloS One *4*, e6146.

Kim, J., and Shin, H. (2013). Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data. Journal of the American Medical Informatics Association *20*, 613–618.

Kim, W., Kim, K.S., Lee, J.E., Noh, D.-Y., Kim, S.-W., Jung, Y.S., Park, M.Y., and Park, R.W. (2012). Development of novel breast cancer recurrence prediction model using support vector machine. Journal of Breast Cancer *15*, 230–238.

Koboldt, D.C., Fulton, R.S., McLellan, M.D., Schmidt, H., Kalicki-Veizer, J., McMichael, J.F., Fulton, L.L., Dooling, D.J., Ding, L., Mardis, E.R., et al. (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61–70.

Koestler, D.C., Marsit, C.J., Christensen, B.C., Karagas, M.R., Bueno, R., Sugarbaker, D.J., Kelsey, K.T., and Houseman, E.A. (2010). Semi-supervised recursively partitioned mixture models for identifying cancer subtypes. Bioinformatics *26*, 2578–2585.

Koo, B.-H., Hurskainen, T., Mielke, K., Aung, P.P., Casey, G., Autio-Harmainen, H., and Apte, S.S. (2007). ADAMTSL3/punctin-2, a gene frequently mutated in colorectal tumors, is widely expressed in normal and malignant epithelial cells, vascular endothelial cells and other cell types, and its mRNA is reduced in colon cancer. International Journal of Cancer *121*, 1710–1716.

Korch, C., Spillman, M.A., Jackson, T.A., Jacobsen, B.M., Murphy, S.K., Lessey, B.A., Jordan, V.C., and Bradford, A.P. (2012). DNA profiling analysis of endometrial and ovarian cell lines reveals misidentification, redundancy and contamination. Gynecologic Oncology *127*, 241–248.

Kovalchik, S. (2017). RISmed: Download Content from NCBI Databases. R Package Version 2.1.7.

Krause, S., Maffini, M.V., Soto, A.M., and Sonnenschein, C. (2010). The microenvironment determines the breast cancer cells' phenotype: organization of MCF7 cells in 3D cultures. Bmc Cancer *10*.

Kuramoto, H., and Nishida, M. (2003). Cell and Molecular Biology of Endometrial Carcinoma.

Lavecchia, A. (2015). Machine-learning approaches in drug discovery: methods and applications. Drug Discovery Today *20*, 318–331.

Längkvist, M., Karlsson, L., and Loutfi, A. (2014). A review of unsupervised feature learning and deep learning for time-series modeling. Pattern Recognition Letters *42*, 11–24.

Ledig, S., Röpke, A., and Wieacker, P. (2010). Copy Number Variants in Premature Ovarian Failure and Ovarian Dysgenesis. Sexual Development *4*, 225–232.

Lee, D.-F., Kuo, H.-P., Chen, C.-T., Hsu, J.-M., Chou, C.-K., Wei, Y., Sun, H.-L., Li, L.-Y., Ping, B., Huang, W.-C., et al. (2007). IKKβ Suppression of TSC1 Links Inflammation and Tumor Angiogenesis via the mTOR Pathway. Cell *130*, 440–455.

Lee, G., Rodriguez, C., and Madabhushi, A. (2008). Investigating the efficacy of nonlinear dimensionality reduction schemes in classifying gene and protein expression studies. IEEE/ACM Transactions on Computational Biology and Bioinformatics *5*, 368–384.

Lehmann, B.D., Bauer, J.A., Chen, X., Sanders, M.E., Chakravarthy, A.B., Shyr, Y., and Pietenpol, J.A. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. The Journal of Clinical Investigation *121*, 2750–2767.

Ley, T.J., Miller, C., Ding, L., Raphael, B.J., Mungall, A.J., Robertson, A.G., Hoadley, K., Triche, T.J.J., Laird, P.W., Baty, J.D., et al. (2013). Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. The New England Journal of Medicine *368*, 2059–2074.

Liaw, A., and Wiener, M. (2002). Classification and Regression by randomForest. R News *2*, 18–22.

Liu, Y., Gu, Q., Hou, J.P., Han, J., and Ma, J. (2014). A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression. Bmc Bioinformatics *15*.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology *15*, 550.

Luo, D., Wilson, J.M., Harvel, N., Liu, J., Pei, L., Huang, S., Hawthorn, L., and Shi, H. (2013). A systematic evaluation of miRNA:mRNA interactions involved in the migration and invasion of breast cancer cells. Journal of Translational Medicine *11*, 57.

Maclin, P.S., Dempsey, J., Brooks, J., and Rand, J. (1991). Using neural networks to diagnose cancer. Journal of Medical Systems *15*, 11–19.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations (University of California Press).

Mayakonda, A., and Koeffler, P.H. (2016). Maftools: Efficient analysis, visualization and summarization of MAF files from large-scale cohort based cancer studies. BioRxiv.

McGranahan, N., and Swanton, C. (2017). Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. Cell *168*, 613–628.

Medico, E., Russo, M., Picco, G., Cancelliere, C., Valtorta, E., Corti, G., Buscarino, M., Isella, C., Lamba, S., Martinoglio, B., et al. (2015). The molecular landscape of colorectal cancer cell lines unveils clinically actionable kinase targets. Nature

Communications *6*, 7002.

Menden, M.P., Iorio, F., Garnett, M., McDermott, U., Benes, C.H., Ballester, P.J., and Saez-Rodriguez, J. (2013). Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. Plos One *8*.

Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biology *12*.

Mo, B., Vendrov, A.E., Palomino, W.A., DuPont, B.R., Apparao, K.B.C., and Lessey, B.A. (2006). ECC-1 cells: A well-differentiated steroid-responsive endometrial cell line with characteristics of luminal epithelium. Biology of Reproduction *75*, 387–394.

Molina-Pinelo, S., Gutiérrez, G., Pastor, M.D., Hergueta, M., Moreno-Bueno, G., García-Carbonero, R., Nogal, A., Suárez, R., Salinas, A., Pozo-Rodríguez, F., et al. (2014). MicroRNA-Dependent Regulation of Transcription in Non-Small Cell Lung Cancer. PLoS ONE *9*, e90524.

Molstad, A.J., Hsu, L., and Sun, W. (2019). Gaussian process regression for survival time prediction with genome-wide gene expression. Biostatistics.

Muzny, D.M., Bainbridge, M.N., Chang, K., Dinh, H.H., Drummond, J.A., Fowler, G., Kovar, C.L., Lewis, L.R., Morgan, M.B., Newsham, I.F., et al. (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature *487*, 330.

Nagamani, M., and Stuart, C.A. (1998). Specific binding and growth-promoting activity of insulin in endometrial cancer cells in culture. American Journal of Obstetrics and Gynecology *179*, 6–12.

Najgebauer, H., Yang, M., Francies, H., Stronach, E.A., Garnett, M.J., Saez-Rodriguez, J., and Iorio, F. (2018). CELLector: Genomics Guided Selection of Cancer in vitro Models. bioRxiv 275032.

Nakayama, K., Nakayama, N., Ishikawa, M., and Miyazaki, K. (2012). Endometrial serous carcinoma: Its molecular characteristics and histology-specific treatment strategies. Cancers *4*, 799–807.

Neve, R.M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F.L., Fevr, T., Clark, L., Bayani, N., Coppe, J.-P., Tong, F., et al. (2006). A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. Cancer Cell *10*, 515–527.

Newlands, E.S., Stevens, M.F.G., Wedge, S.R., Wheelhouse, R.T., and Brock, C. (1997). Temozolomide: A review of its discovery, chemical properties, pre-clinical development and clinical trials. Cancer Treatment Reviews *23*, 35–61.

Niu, N., and Wang, L. (2015). In vitro human cell line models to predict clinical response to anticancer drugs. Pharmacogenomics *16*, 273–285.

Noske, A., Brandt, S., Valtcheva, N., Wagner, U., Zhong, Q., Bellini, E., Fink, D., Obermann, E.C., Moch, H., and Wild, P.J. (2017). Detection of CCNE1/URI (19q12) amplification by in situ hybridisation is common in high grade and type II endometrial cancer. Oncotarget *8*, 14794–14805.

Parise, C.A., and Caggiano, V. (2014). Breast Cancer Survival Defined by the ER/PR/HER2 Subtypes and a Surrogate Classification according to Tumor Grade and Immunohistochemical Biomarkers. Journal of Cancer Epidemiology *2014*, 469251.

Park, C., Ahn, J., Kim, H., and Park, S. (2014). Integrative gene network construction to analyze cancer recurrence using semi-supervised learning. PloS One *9*, e86309.

Park, K., Ali, A., Kim, D., An, Y., Kim, M., and Shin, H. (2013). Robust predictive model for evaluating breast cancer survivability. Engineering Applications of Artificial Intelligence *26*, 2194–2205.

Pascolo, S. (2016). Time to use a dose of Chloroquine as an adjuvant to anti-cancer chemotherapies. European Journal of Pharmacology *771*, 139–144.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science *2*, 559–572.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research *12*, 2825–2830.

Pranckevičius, T., and Marcinkevičius, V. (2017). Comparison of Naïve Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification. Baltic J. Modern Computing *5*, 221–232.

Qiu, Z., Zou, K., Zhuang, L., Qin, J., Li, H., Li, C., Zhang, Z., Chen, X., Cen, J., Meng, Z., et al. (2016). Hepatocellular carcinoma cell lines retain the genomic and transcriptomic landscapes of primary human cancers. Scientific Reports *6*, 27411.

Rae, J.M., Creighton, C.J., Meck, J.M., Haddad, B.R., and Johnson, M.D. (2007). MDA-MB-435 cells are derived from M14 melanoma cells–a loss for breast cancer, but a boon for melanoma research. Breast Cancer Research and Treatment *104*, 13–19.

Rahman, R., Matlock, K., Ghosh, S., and Pal, R. (2017). Heterogeneity Aware Random Forest for Drug Sensitivity Prediction. Scientific Reports *7*, 11347.

Ram, M., Najafi, A., and Shakeri, M.T. (2017). Classification and Biomarker Genes Selection for Cancer Gene Expression Data Using Random Forest. Iranian Journal of Pathology *12*, 339–347.

Rasmussen, C.E., and Williams, C.K.I. (2005). Gaussian Processes for Machine Learning. Gaussian Processes for Machine Learning 1–247.

R Core Team (2018). R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing).

Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P., and Mesirov, J.P. (2006). GenePattern 2.0. Nature Genetics *38*, 500.

Riaz, M., Jaarsveld, M.T. van, Hollestelle, A., Prager-van der Smissen, W.J., Heine, A.A., Boersma, A.W., Liu, J., Helmijr, J., Ozturk, B., Smid, M., et al. (2013). miRNA expression profiling of 51 human breast cancer cell lines reveals subtype and driver mutation-specific miRNAs. Breast Cancer Research *15*, R33.

Rizk, N.P., Ishwaran, H., Rice, T.W., Chen, L.-Q., Schipper, P.H., Kesler, K.A., Law, S., Lerut, T.E.M.R., Reed, C.E., Salo, J.A., et al. (2010). Optimum Lymphadenectomy for Esophageal Cancer. Annals of Surgery *251*, 46–50.

Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics *20*, 53–65.

Saadatpour, A., Lai, S., Guo, G., and Yuan, G.-C. (2015). Single-Cell Analysis in Cancer Genomics. Trends in Genetics *31*, 576–586.

Sabatier, R., Finetti, P., Cervera, N., Lambaudie, E., Esterni, B., Mamessier, E., Tallet, A., Chabannon, C., Extra, J.-M., Jacquemier, J., et al. (2011). A gene expression signature identifies two prognostic subgroups of basal breast cancer. Breast Cancer Research and Treatment *126*, 407–420.

Sarica, A., Cerasa, A., and Quattrone, A. (2017). Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review. Frontiers in Aging Neuroscience *9*, 329.

Saso, S., Chatterjee, J., Georgiou, E., Ditri, A.M., Smith, J.R., and Ghaem-Maghami, S. (2011). Endometrial cancer. BMJ *343*.

Schmidt, M., Böhm, D., Törne, C. von, Steiner, E., Puhl, A., Pilch, H., Lehr, H.-A., Hengstler, J.G., Kölbl, H., and Gehrmann, M. (2008). The humoral immune system has a key prognostic impact in node-negative breast cancer. Cancer Research *68*, 5405–5413.

Sertkaya, A., and Birkenbach, A. (2011). Examination of clinical trial costs and barriers for drug development (U.S. Department of Health & Human Services).

Shen, R., Olshen, A.B., and Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics *25*, 2906–2912.

Shepherd, T., and Owenius, R. (2012). Gaussian Process Models of Dynamic PET for Functional Volume Definition in Radiation Oncology. Ieee Transactions on Medical Imaging *31*, 1542–1556.

Sia, D., Hoshida, Y., Villanueva, A., Roayaie, S., Ferrer, J., Tabak, B., Peix, J., Sole, M., Tovar, V., Alsinet, C., et al. (2013). Integrative Molecular Analysis of Intrahepatic Cholangiocarcinoma Reveals 2 Classes That Have Different Outcomes. Gastroenterology *144*, 829–840.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. Bioinformatics *21*, 7881.

Sinha, R., Schultz, N., and Sander, C. (2015). Comparing cancer cell lines and tumor samples by genomic profiles. bioRxiv 28159.

Smittenaar, C.R., Petersen, K.A., Stewart, K., and Moitt, N. (2016). Cancer incidence and mortality projections in the UK until 2035. British Journal of Cancer *115*, 1147–1155.

Soliman, P.T., and Lu, K.H. (2013). Neoplastic Diseases of the Uterus Endometrial Hyperplasia, Endometrial Carcinoma, and Sarcoma: Diagnosis and Management.

Stojadinovic, A., Nissan, A., Eberhardt, J., Chua, T.C., Pelz, J.O.W., and Esquivel, J. (2011). Development of a Bayesian Belief Network Model for personalized prognostic risk assessment in colon carcinomatosis. The American Surgeon *77*, 221–230.

Suzuki, R., and Shimodaira, H. (2015). pvclust: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling. R Package Version 2.0-0.

Tan, H., Bao, J., and Zhou, X. (2015). Genome-wide mutational spectra analysis reveals significant cancer-specific heterogeneity. Scientific Reports *5*, 12566.

Tanner, M., Kapanen, A.I., Junttila, T., Raheem, O., Grenman, S., Elo, J., Elenius, K., and Isola, J. (2004). Characterization of a novel cell line established from a patient with Herceptin-resistant breast cancer. Molecular Cancer Therapeutics *3*, 1585–1592.

Tarca, A.L., Kathri, P., and Draghici, S. (2013). SPIA: Signaling Pathway Impact Analysis (SPIA) using combined evidence of pathway over-representation and unusual signaling perturbations. R Package Version 2.28.0.

Theisen, E.R., Gajiwala, S., Bearss, J., Sorna, V., Sharma, S., and Janat-Amsbury, M. (2014). Reversible inhibition of lysine specific demethylase 1 is a novel anti-tumor strategy for poorly differentiated endometrial carcinoma. BMC Cancer *14*, 752.

Tseng, C.-J., Lu, C.-J., Chang, C.-C., and Chen, G.-D. (2014). Application of machine learning to predict the recurrence-proneness for cervical cancer. Neural Computing and Applications *24*, 1311–1316.

Verbaanderd, C., Maes, H., Schaaf, M.B., Sukhatme, V.P., Pantziarka, P., Sukhatme, V., Agostinis, P., and Bouche, G. (2017). Repurposing Drugs in Oncology (ReDO)-chloroquine and hydroxychloroquine as anti-cancer agents. Ecancermedicalscience *11*, 781.

Waddell, M., Page, D., and Shaughnessy, J. (2005). Predicting cancer susceptibility from single-nucleotide polymorphism data. In Proceedings of the 5th International Workshop on Bioinformatics - Biokdd '05, (New York, New York, USA: ACM Press), p. 21.

Wang, X., Osada, T., Wang, Y., Yu, L., Sakakura, K., Katayama, A., McCarthy, J.B., Brufsky, A., Chivukula, M., Khoury, T., et al. (2010). CSPG4 Protein as a New Target for the Antibody-Based Immunotherapy of Triple-Negative Breast Cancer. JNCI: Journal of the National Cancer Institute *102*, 1496–1512.

Wassermann, D., Bloy, L., Kanterakis, E., Verma, R., and Deriche, R. (2010). Unsupervised white matter fiber clustering and tract probability map generation: Applications of a Gaussian process framework for white matter fibers. Neuroimage *51*, 228–241.

Weigelt, B., Warne, P.H., Lambros, M.B., Reis-Filho, J.S., and Downward, J. (2013). PI3K Pathway Dependencies in Endometrioid Endometrial CancWeigelt, B. et al. (2013) PI3K Pathway Dependencies in Endometrioid Endometrial Cancer Cell Lines, Clinical Cancer Research, 19(13), pp. 3533–3544. doi: 10.1158/1078-0432.ccr-12-3815.er Cell Line. Clinical Cancer Research *19*, 3533–3544.

Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M., and Canc Genome Atlas Res, N. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics

45, 1113–1120.

Xiong, S., Klausen, C., Cheng, J.-C., Zhu, H., and Leung, P.C.K. (2015). Activin B induces human endometrial cancer cell adhesion, migration and invasion by up-regulating integrin beta3 via SMAD2/3 signaling. Oncotarget *6*, 31659–31673.

Yang, W., Soares, J., Greninger, P., Edelman, E.J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J.A., Thompson, I.R., et al. (2013). Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Research *41*, D955–D961.

Yang, Z., Zhuang, L., Szatmary, P., Wen, L., Sun, H., Lu, Y., Xu, Q., and Chen, X. (2015). Upregulation of Heat Shock Proteins (HSPA12A, HSP90B1, HSPA4, HSPA5 and HSPA6) in Tumour Tissues Is Associated with Poor Outcomes from HBV-Related Early-Stage Hepatocellular Carcinoma. Int. J. Med. Sci *12*, 256–263.

Zack, T.I., Schumacher, S.E., Carter, S.L., Cherniack, A.D., Saksena, G., Tabak, B., Lawrence, M.S., Zhang, C.-Z., Wala, J., Mermel, C.H., et al. (2013). Pan-cancer patterns of somatic copy number alteration. Nature Publishing Group.

Zhou, X., Wang, Z., Zhao, Y., Podratz, K., and Jiang, S. (2014). Characterization of sixteen endometrial cancer cell lines. Cancer Research *67*, 3870 LP–3870.