Automated Discovery of Interacting Genomic Events That Impact Cancer Survival by Using Data Mining and Machine Learning Techniques

Richard Lupat

ORCID: 0000-0002-6435-7100

Master of Philosophy

December 2020

The Sir Peter MacCallum Department of Oncology

Faculty of Medicine, Dentistry and Health Sciences

The University of Melbourne

Submitted in total fulfilment for the degree of Master of Philosophy

Abstract

Rapid advancement in genomic technologies has driven down the cost of sequencing significantly. This efficiency has enabled large-scale cancer genomic studies to be conducted, generating a vast amount of data across different levels of omics variables. However, the tasks to extract new knowledge and information from this enormous volume of data present unique challenges. These analyses often require the application of specialised techniques for data mining, integration and interpretation to provide valuable insights. With the rise of machine learning adoption in recent decades, many advanced computational algorithms based on artificial intelligence techniques have also been proposed to analyse these genomics data. Although some of these applications have led to clinically relevant conclusions, many others are still relying on incomplete prior knowledge, or limited to only a selected number of features. These limitations raise the general question about the broader applicability of machine learning in the field of cancer genomics.

This research addresses this question by assessing the application of machine learning techniques in the context of breast cancer genomics data. This assessment includes a comprehensive evaluation of computational methods for predicting cancer driver genes and the development of a novel deep learning approach for identifying breast cancer subtypes. The evaluation result of driver gene prediction algorithms suggests that the selection of the best method to be applied to a dataset will primarily be driven by the objectives of the study and the characteristics of the dataset. All of the evaluated approaches could identify well-studied genes, but not all of them performed as well on smaller datasets, subtype-specific cohorts, and in discovering novel genes.

To examine the benefit of a more complex machine learning model, this thesis also presents a novel deep learning approach that integrates multi-omics data for predicting various breast cancer' biomarkers and molecular subtypes. This method combines a semi-supervised autoencoder for dimensionality reduction, and a supervised multitask learning setup for the classifications. Taking an input of gene expression, somatic point mutation and copy number data, the algorithm predicts the ER-Status, HER2-Status and molecular subtypes of breast cancer samples. Further survival analysis of the outputs from this deep learning approach indicates that the predicted subtypes show a stronger correlation with patient prognosis compared to the original PAM50 label.

While the outputs from machine learning algorithms still require further validation, the adoption of these complex computational methods in cancer genomics will become increasingly common. Collectively, the results from this thesis suggest that the machine learning analysis of 'omics data hold great potential in automating the discovery of clinically-relevant molecular features.

Declaration

This is to certify that:

1. This thesis comprises only my original work towards the MPhil
2. Due acknowledgement has been made in text to all other materials used,
3. This thesis is less than 50,000 words in length, exclusive of tables, figures, bibliographies and appendices

Signed:

Richard Lupat
20 December 2020

Preface

The following manuscript has been submitted for publication to Bioinformatics. The work of this publication arises from research completed for this thesis and is included in an article format for one of the result chapters.

Moanna: Multi-Omics Autoencoder-Based Neural Networks Algorithm for Predicting Breast Cancer Subtypes. Richard Lupat, Sherene Loi, Jason Li. *Submitted for publication to Bioinformatics on 20 December 2020 [Manuscript ID: BIOINF-2020-2670].*

Author contribution: Richard Lupat (RL) designed and implemented the neural network model with ongoing feedback and advice from Jason Li and Sherene Loi. The final draft of this manuscript was written by RL. All authors have read and approved this manuscript.

Acknowledgement

I would like to express my sincere gratitude to my supervisors Dr Jason Li and Professor Sherene Loi, my mentor Dr Maria Doyle, the thesis committee, research education team, fellow bioinformaticians and Peter Mac colleagues for their continuous support, technical expertise, and guidance in the completion of this thesis.

My sincere thanks extend to all the patients enrolled in The Cancer Genome Atlas (TCGA) and Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) studies. The results shown in this thesis are in part based upon data generated by these studies.

Above all, I would like to thank my family and friends, who have provided me with unwavering support and encouragement throughout my study.

Table of contents

List of Figures

List of Tables

Chapter 1 - Introduction

1.1. Overview

The first complete human genome was sequenced by The Human Genome Project in 2003, which took more than 10 years at the cost of almost $3 billion [1, 2]. Since then, new sequencing technologies have started to appear to deliver high throughput sequencing for a fraction of the cost. These next generation sequencing (NGS) technology can simultaneously sequence multiple whole-genomes and has helped drive down the cost of sequencing to almost $1,000 per human genome [2]. The high throughput and cost efficiency of NGS has enabled large genomic cohort studies to be conducted, such as 1000 Genomes Project [3], The Cancer Genome Atlas (TCGA) [4], and International Cancer Genome Consortium (ICGC) [5].

One of the primary objectives of large-scale cancer genomics studies is the identification of the main drivers of tumorigenesis. This process is onerous due to the genetic instability of cancer cells, where a large portion of the mutations are passengers that do not confer selective advantage [6]. The findings from these studies have also illustrated that different tumour types are likely to be driven by different driver genes [7, 8]. Moreover, even within the same cancer type, the mutation patterns often differ between patients and they are often studied as separate cancer subtypes [9]. Understanding the intertumour heterogeneity is essential, as some of these subtypes have been demonstrated to have different prognosis and response to treatments [10].

The advancement of computational methods is integral to the analysis of this data. Some of these projects studied multiple cancer types across thousands of patients, generating petabytes of data across the various level of omics that are often accompanied by long term clinical follow-up information. To extract new insights from these comprehensive datasets, it often involves the application of various skills, including data mining, machine learning algorithms and software engineering for data management. For example, cbioportal [11] provides a web interface for data exploration that facilitate the analysis of large-scale publicly available genomics data. Multiple machine learning algorithms have also been proposed for predicting cancer prognosis [12]. Although some of these computation methods are still in its infancy, many have been successfully applied to extract biologically relevant features from cancer omics data.

The growth in the size of genomic datasets is also encouraging for the applications of deep learning techniques in bioinformatics. The adoption of deep learning in healthcare has traditionally been hindered by the $n << p$ problem, where there are limited training data (n) in comparison to the number of genomic features (p). Although current dataset size is still not as big as other domains (such as medical imaging data), the continuous progression in computing power in recent years means it is now feasible to design a more complex model to handle high dimensional data. The combination of these developments, alongside cloud computing, have expanded the resources for machine-learning-based algorithms development in cancer genomics.

Various studies in genomics data interpretation have leveraged deep learning techniques to improve their accuracy further. Implementations such as DANN [13], and PrimateAI [14], apply artificial neural networks for variant interpretation. These deep learning methods are also being used to extract biologically relevant features from gene expression data [15] and for building cancer prognosis predictors [16]. Similarly, Google DeepVariant implements sequencing variant-calling approach using neural network algorithms that were originally developed for image classification [17]. It is likely that the adoption of deep learning approaches in cancer genomics will become increasingly common.

This thesis seeks to investigate the use of machine learning approaches in extracting valuable insights from large-scale breast cancer genomics studies. In particular, the main interest of this research is in the application of computational approaches to interrogate the main drivers of tumorigenesis and the adoption of deep learning in predicting clinically-relevant features. As the research of machine learning application in genomics is still actively progressing, it is important to study the current limitation and address how these methods can be further improved for other cancer research applications.

1.2. Research Aims

The overall objective of this study is to assess the application of machine learning techniques in predicting clinically-relevant molecular features from large genomic breast cancer datasets. The two broad aims of this research are:

Aim 1: To comprehensively evaluate driver gene prediction algorithms and their performance across different omics data

Aim 2: To develop a deep-learning-based classifier that integrates multi-omics data for categorising breast cancer samples into multiple subtypes

1.3. Thesis Outline

This thesis is structured into five chapters, including this introductory chapter, literature review, research results specifically for the degree, a planned publication and a final discussion chapter. The *submitted manuscript* is included in its final draft.

Chapter 2 presents the literature review that covers the biological and technical contexts of this thesis. This chapter is split into four sections. The first section (section 2.1) introduces the background of the machine learning principles applied in this study, with a particular focus on the deep neural network algorithms. The second section (section 2.2) covers cancer genomics, including the common genetic alterations and the sequencing technologies. The third section (section 2.3) reviews the main disease context of this thesis, breast cancer, and details the different breast cancer molecular subtypes. The last section (section 2.4) specifies the comprehensive cancer studies that were used for training and evaluating the methods outlined in this thesis.

Chapter 3 details the evaluation of popular driver gene predictions algorithms and the assessment of their predicted drivers in the context of breast cancer. Five algorithms were comprehensively reviewed, covering multiple aspects of driver gene characteristics and different setup of omics datasets.

Chapter 4 introduces the deep learning algorithm that is used to build the breast cancer subtype prediction model. This chapter comprises of the submitted manuscript and complementary sections covering the summary of the problem.

Chapter 5 provides the final conclusions of this thesis and observations on the possible future directions on the development and application of these deep learning techniques. This final chapter also outlines the limitations of the presented approaches and previews some ongoing initiatives stemming out from this research. The significance of the findings of this thesis in the context of cancer research is also reiterated.

Chapter 2 - Literature Review and Data Sources

This chapter consists of the literature review of four main sub-contexts used throughout this thesis. Section 2.1 details the background of the machine learning principles applied in this study, with a particular focus on deep neural network algorithms. Cancer genomics is described in section 2.2, and the main disease context, breast cancer, is introduced in section 2.3. Section 2.4 presents datasets from large cancer studies that were used for training and evaluating the methods outlined in this thesis.

2.1. Machine Learning

Machine learning is defined as algorithms that automatically discover how to accomplish tasks by learning and improving through experience with provided training data [18]. This learning experience is often measured by computing specific performance metrics, such as accuracy or error-rate, as the algorithm goes through some data points. These metrics vary depending on the tasks that the machine learning model is trying to undertake. Some of the most common machine learning tasks include classification, regression, denoising and clustering.

There are three broad categories of machine learning: supervised, unsupervised and reinforcement learning. Supervised learning is often applied to problems where training labels are available, such as classification tasks, and algorithms are trained to find a function to map training inputs into the provided targets. Unsupervised learning deals with training data that has no definitive outcome, such as clustering and denoising tasks. It is also valuable in discovering a simpler representation of input datasets by projecting them to lower-dimensional space. Meanwhile, reinforcement learning improves by interacting with its environment and acquiring feedback from previous experience to maximise rewards [19]. It is commonly applied on tasks that require decision making that can be perfected to trial and error, such as games [18].

Machine learning is a subset of Artificial Intelligence (AI), whose primary objective is to program machines for completing intellectually difficult tasks. While early AI techniques demonstrate values in solving tasks that are computationally difficult for a human, its application on *simpler* human problems, such as image or speech recognition, proves to be more challenging [18]. The human brain can naturally learn the patterns for these tasks, but they are difficult to be precisely defined. On the other hand, computers acquire their *intelligence* by learning intricate patterns of training data provided by users, and this often has to be represented in a very structured way [18, 20]. Thus, input feature engineering is a critical first step in developing machine learning models.

Manually designing machine learning features requires many domain experts and takes a long time [18]. It is often challenging to distinguish which features to extract, especially when applied on emerging domains like human genomics. Without a good feature engineering, machine learning models also suffered from the *curse of dimensionality*, where models struggled to learn patterns from input data with a high number of variables [21]. This setting is particularly problematic in a field where the amount of training samples is not proportional to the number of features, leading machine learning models to overfit [22].

Many variations of traditional machine learning have been developed and applied to address this problem. This includes a form of representation learning, such as *autoencoders*, where another layer of unsupervised machine learning is used to learn the representation of the input data [21]. Another variation that has been widely adopted to solve this problem is deep learning. Deep learning introduces multiple additional layers to learn various smaller abstract features defining the original representation of the raw data [23]. Although these extra layers come at an extra computational cost, it is also the one that set deep learning apart from traditional machine learning approaches in terms of performance in solving AI tasks.

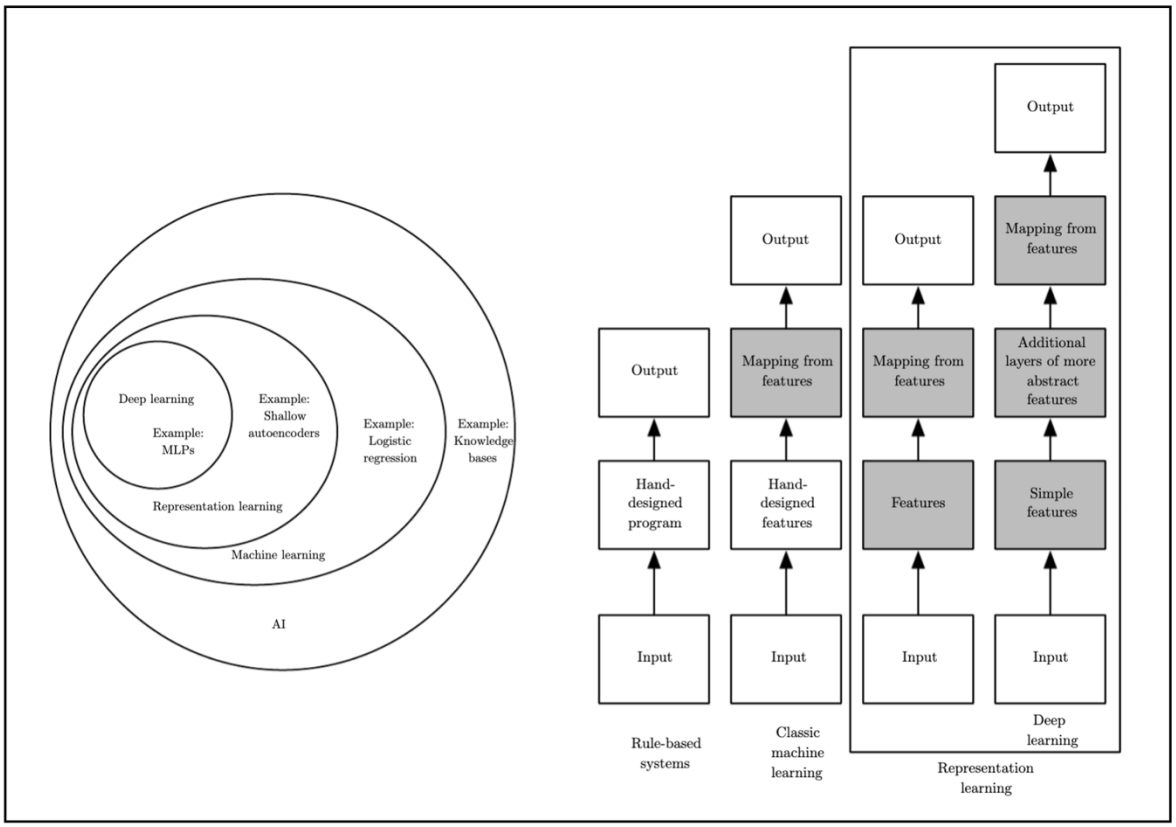


Figure 2.1 - (Left) Venn diagram showing the relationship between Artificial Intelligence (AI), machine learning and deep learning. (Right) Flowcharts describing how different AI systems learn *from their input data. Both diagrams are from [18].*

In the subsequent sections, varieties of machine learning techniques will be introduced. This includes deep learning, random forest, and examples of unsupervised machine learning algorithms.

2.1.1. Deep Learning

Deep learning is a type of machine learning that is characterised by its deep structure consisting of many layers of functions for transforming inputs to outputs. This structure allows it to address the main limitation of traditional machine learning of requiring hand-picked features. Despite its recent rise in popularity, deep learning has been around for decades. It was initially proposed in the 1940s [18] and has undergone many developments in the 1980s [24].

Multiple factors have driven the recent rise of deep learning popularity. First, the availability of more massive training datasets enables models to solve high-dimensional problems better and to generalise further [18]. Second, the advances in computing power have allowed models to increase their size to a state that was not computationally possible in the old days [18]. In turns, these *deeper* models can achieve higher accuracy in dealing with more complex real-world problems. Also, the availability of deep learning software libraries and infrastructures, such as Tensorflow [25] and PyTorch [26], facilitates more straightforward deep learning implementations. The continuous development of these frameworks has lowered the technical barrier for deep learning application across various domains.

Deep learning is commonly associated with artificial neural networks (ANNs), as neural network forms the basis of deep learning algorithms. The term *deep learning* refers to the depth of its neural network structure, where deep learning algorithms would have more layers than a single layer neural network model. This *neural network* terminology can be attributed to the early development of neural networks algorithms that was inspired by the human brain's ability to learn complex data through interconnected neurons [27]. Nevertheless, modern deep learning advancement is driven by applied mathematics, such as linear algebra, calculus and probability [18]. These fundamentals of mathematics form essential parts of deep learning components, especially in setting up objective function and training optimisation.

A typical deep learning algorithm development consists of at least a model, a dataset, a cost function and an optimisation procedure [18]. While these can vary on more complex deep learning implementations, some of these basic components are described below.

1. **Deep learning model.** A neural network model comprises of multiple layers. In its simplest form, there are three interconnected layers: input, hidden, and output. Every layer consists of multiple neurons, where each neuron in the input layer represents a feature from the input data. The number of hidden layers between input and output is the *depth* of this network architecture that is essential for learning different abstraction of input feature combinations. Each neuron in the hidden layers is a mathematical function that combines the values of input neurons from the previous layer, based on certain weights and biases that form the main attributes of the deep learning model. Output layer summarises the result of the network's learning for a given input. The number of neurons in this layer depends on the learning objective. For example, in multi-class classifications, the output layer will have as many neurons as the number of classes.

2. **Dataset.** There is usually more than one dataset employed to build a deep learning model. These include at least a training dataset and an independent test dataset, for training and measuring the generalisation performance of this trained model. A third dataset is usually introduced for tuning hyperparameters. This validation dataset is commonly derived from a fraction of training dataset and used for estimating models' generalisation error during training.

3. **Cost function.** A cost function, or objective function, is used to measure the training error from every training iteration, where this error is generally defined as the

difference between predicted and expected output. A deep learning model is trained to either minimise or maximise this function, depending on the learning objective and optimisation strategy. Some examples of cost functions are mean squared error and cross-entropy.

4. **Optimiser.** As a model is trained in regards to its objective function, another algorithm is engaged to optimise this function. The role of an optimiser is to efficiently update attributes of the model, such as weights, in order to minimise or maximise this cost function. An example of commonly used optimisers in deep learning is Stochastic Gradient Descent (SGD). SGD updates models' attributes by following the opposite direction of function's gradients to arrive at a local minimum. This gradient is calculated by a **backpropagation** algorithm, and the update follows a **learning rate** that controls the step size that SGD will take. More recent gradient descent algorithms, such as Adam [28], RMSProp [29], and Adagrad [30], introduce an adaptive learning rate to tune parameters' step size automatically.

Additional functions are also attached to complete each of these main elements. **Activation functions** are applied to hidden layers to handle the non-linear relationship in a deep learning model. Some examples of these functions include rectified linear unit (ReLU), hyperbolic tangent (tanh) and logistic sigmoid (sigmoid), and the choice of activation functions could affect the performance and training speed of the overall model [31]. These functions are also used on the output layer, and the selection of these units are tightly coupled with the cost function of the model. For example, to calculate the probability distribution of *n* classes in a classification task, one can use sigmoid function (for binary) or softmax (for multi-class) as the output unit.

There are also many strategies applied to prevent overfitting in a deep learning model. These **regularisation** strategies can be introduced to the model over multiples avenues. The most commonly used techniques are L2 and L1 regularisations, where additional *regularisation term* is added to the cost function [18]. Another widely use method is **dropout** regularisation, where a proportion of neurons are randomly switched off during training [32]. It is also possible to combine multiple methods and training to improve generalisation, such as multitask learning and semi-supervised learning [33].

There are many variations of deep learning implementations that adapt some of the concepts described above. The following sub-sections will introduce the specific deep learning techniques that are applied in this thesis, with implementation details and evaluations results further described in Chapter 4.

2.1.1.1. Feedforward neural networks

The most classic implementation of deep learning is deep feedforward neural network, or also commonly referred to as multilayer perceptron (MLP). A feedforward neural network model is structured to move inputs to output through a series of connected networks without any feedback connections, and hence the name *forward* [18]. A typical feedforward neural network consists of multiple layers representing chains of functions that transform input data into the task output. The structure and attributes of these model follow the general deep

learning setup, consisting of an input layer, hidden layers and output layer, as described in *section 2.1.1*.

This architecture design serves as an important basis for many other neural networks design. For example, a recurrent neural network (RNN) is a feedforward neural network with feedback [34]. Deep autoencoders [35] is a variation of multilayer perceptron with bottleneck layers (more details will be discussed in section 2.1.1.2), whereas convolution neural network (CNN) [36] is a feedforward neural network with *convolution* or pooling regularisations.

2.1.1.2. Autoencoders

Autoencoder is a type of unsupervised learning characterised by its bottleneck layer. The most basic autoencoders consist of two functions, an encoder and a decoder, both of which are a form of feedforward neural network [18]. The main idea of the algorithm is, given a set of data, there will be a function that can encode the input features into its compressed form and another separate function that can reconstruct these latent variables back into an approximate representation of its original form. The objective function for this model is minimising the difference between the original and reconstructed features, essentially comparing its own input features to its estimated replica. This technique is valuable when applied to dimensionality reduction, where the neurons in the autoencoder's bottleneck layer represent extracted features that have learnt the important representation of the full input features [35]. This application is similar to feature extractions with principal component analysis (PCA), where a large proportion of variance in high dimensional data can be represented by a few top principal components that were calculated from the eigenvalues and eigenvectors of its original data.

2.1.1.3. Semi-supervised learning

An unsupervised learning algorithm, such as autoencoder, can be applied for solving the dimensionality reduction problem. Its application in the form pre-trained model has been proposed to increase the efficiency of supervised deep learning [37]. However, recent development has pushed vanilla supervised learning to perform as good as without the pre-training step [38]. Valpola [39] suggested that this limitation of unsupervised learning can be attributed to how it was trying to learn all possible representation of input data, even though not all information is beneficial in supporting the supervised task. He proposed a joint supervised and unsupervised learning, called ladder network, to address this limitation by adding additional autoencoder connection to every layer of the supervised network [39]. This supervised component essentially helps the autoencoder to focus only on extracting necessary information. While adding additional connection result in more computations, Rasmus et al. [40] found that the overall training time of this semi-supervised learning architecture remains comparable, primarily due to a more efficient learning and faster generalisation.

2.1.1.4. Multitask learning

Multitask learning is characterised by multiple loss functions optimisations, where multiple tasks from the same datasets are being learned from the same training [33]. This approach is

one of deep learning regularisation techniques, implemented by teaching a model to generalise over the characteristics of multiple related tasks [41]. It is achieved through the sharing of hidden layers parameters across multiple tasks (figure 2.2). This parameter sharing helps the model to generalise faster, as noises and biases from several trainings are accumulated to assist model's decision making in discarding unrelated features. This setup is also indirectly augmenting additional training data [33], as variations learnt in the shared layers are targeted towards several tasks' labels.

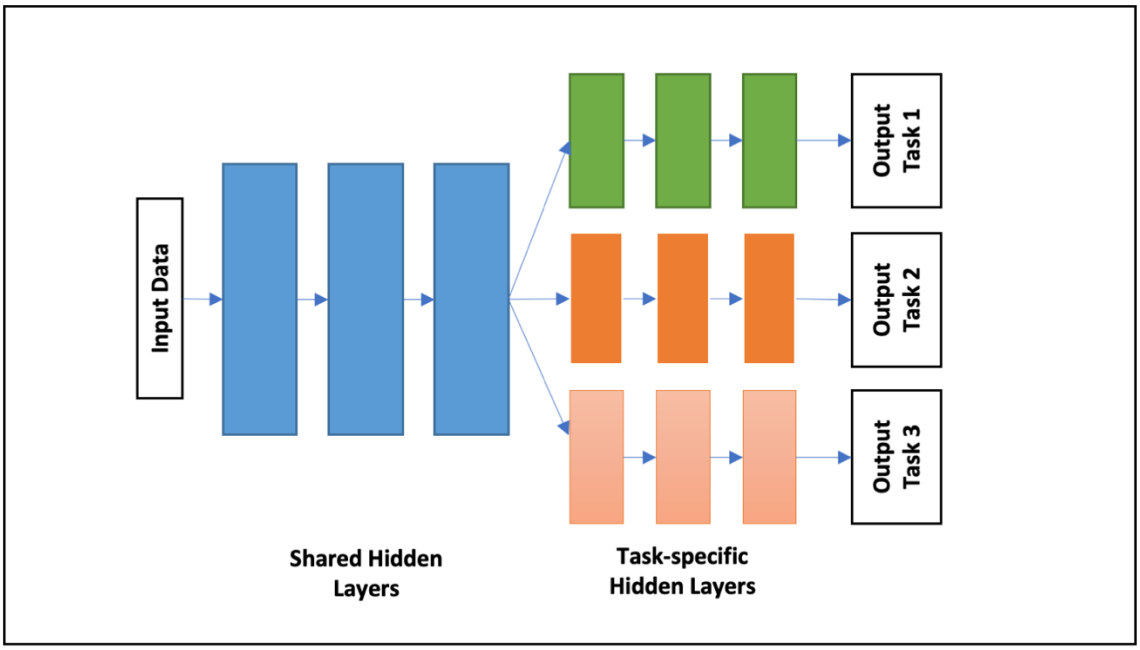


Figure 2.2 - An example of deep neural network setup with hard parameter sharing for multitask learning

2.1.2. Random Forest

Random forest is an ensemble machine learning technique that learns patterns of data through generating multiple decision trees and combining their individual learning [42]. These decision trees are algorithms that use a tree-like structure for representing the separation between different features and attributes of input data. While the application of decision tree benefits from its high interpretability, the individual tree is also prone to overfitting [43]. Random forest addresses this limitation by using the ensemble of multiple decision trees, essentially utilising the decision from a group of uncorrelated trees to correct each other mistakes.

Random forest uses bagging, or bootstrap aggregating, for generating training trees. This method was proposed by Breiman [42], where a set of training data is randomly sampled with replacement, to form each decision tree. This approach assists the building of a forest of uncorrelated trees, which has been described to help generalised the training better. In a classification task, majority voting between individual classifiers is often used to classify the final labels [44].

Random forest models offer various advantages compared to other more complicated machine learning techniques. On top of its overall good prediction accuracy, a random forest

is computationally cheaper and faster to train than neural networks model [45]. Moreover, a random forest model is also easier to interpret, given that the underlying structure is based on decision trees. The bagging and feature randomness approach provides a sort of internal cross-validation that improves models' generalisation [45].

2.1.3. Evaluation metrics

One of the most important final steps in machine learning model development is evaluating the effectiveness of the model in solving tasks that it was trained for. The choice of evaluation techniques differs according to the learning types and objectives. In supervised classification, the performance of a model is usually quantified by measuring the accuracy, precision and recall of the predictions against *gold* labels. As for regression problem, mean squared error and mean absolute error are commonly used as the performance metrics [46].

Various classification metrics focus on measuring a different aspect of a model. Classification accuracy evaluates the proportion of correct predictions to the total number of samples. Although it is the most intuitive, accuracy does not take into account false positive in its evaluation and is particularly problematic on unbalanced datasets [47]. For example, in a binary classification of class A (80% of datasets) and class B (20% of datasets), a model that classify all data to the majority (class A) will always yield high accuracy (80%) despite its poor learning. Therefore, accuracy report is often accompanied by other metrics such as precision, recall and f1-score.

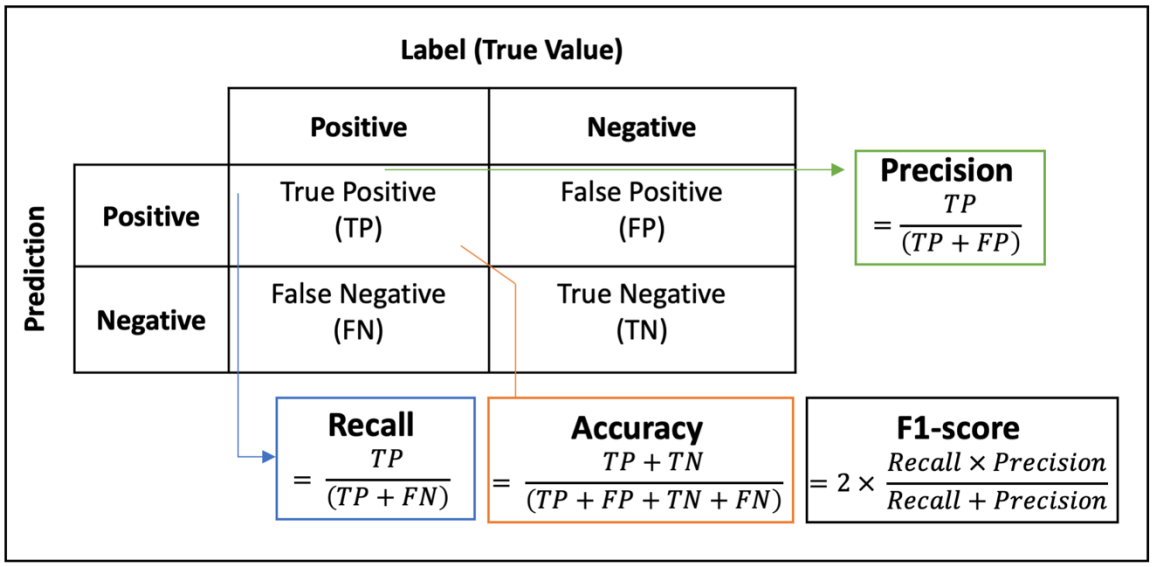


Figure 2.3 - An example of 2x2 confusion matrix and some of the evaluation metrics that could be calculated from this matrix (precision, recall, accuracy and f1-score)

A confusion matrix is also often used to further evaluate the mistakes from a model, especially for multi-class classification. Confusion matrix summarises the predictions from classifiers with regards to known labels [48]. The accuracy of a model can be calculated from the main diagonal of the matrix. Figure 2.3 shows an example of a confusion matrix and various metrics that can be extracted from it.

In the case of unsupervised learning, the effectiveness of a model can be more difficult to quantify due to the lack of ground truth. Unsupervised clustering algorithms' performance is often measured by calculating the distance of separation between predicted clusters, such as silhouette coefficients [49]. In the application for dimensionality reduction, the performance unsupervised learning method can be quantified by its utility in improving the evaluation metrics of downstream tasks. Dimensionality reduction can also be evaluated by its clustering quality [50]. This typically involved running additional clustering algorithm post feature extraction, followed by calculating clustering metrics, such as normalised mutual information (NMI) or adjusted rand index (ARI) [50, 51].

2.2. Cancer Genomics

Cancer is a collection of diseases that are characterised by abnormal cells proliferation. These atypical cells can appear in almost any tissue or organ as human cells progress through cell cycles. During the course of cancer development, neoplastic cells defy the orderly process of cells growth and apoptosis, which may eventually lead to tumour development. Although some of these tumours remain benign, cancerous tumours are malignant. Their invasive characteristic means they can invade the surrounding tissues and even spread beyond their primary sites [6]. This process is known as metastasis, where cancer cells journey through blood or lymphatic system in a body and successfully infiltrates a new location to form secondary tumours [52]. This metastatic cancer is a major cause of cancer deaths from solid tumours [53].

Cancer development complexity across numerous distinct types has led to the exploration of common distinguishable cancer characteristics. In the year 2000, Hanahan and Weinberg proposed six hallmarks of cancer [54], comprises of: self-sufficiency in proliferative signals, insensitivity to growth suppressors, resisting apoptosis, limitless replicative potential, inducing angiogenesis, and enabling invasion and metastasis. They updated this cancer hallmarks in 2014 [55], by introducing four more principles: reprogramming cellular energetics, evading immune destruction, instability and mutability of genome, and tumour-promoting inflammation. Collectively, these ten cancer hallmarks serve as a valuable framework for understanding the underlying complexities of cancer biology.

2.2.1. The genetics of cancer

Cancer is a genetic disease where alterations to genes' functions affect how cells in human body behave. A gene is a functional sequence of DNA that encode the instructions for the production of proteins and RNA molecules [6]. The flow of the genetic information largely follows the central dogma of molecular biology [56], in which DNA is copied into RNA (transcription), and then from RNA to protein (translation). For cancer cells, genetic variations disturb certain cancer genes' roles in production of proteins that are responsible for cells growth and repairs.

There are several types of mutations that alter DNA sequences. The simplest form of mutation is single nucleotide polymorphisms (SNPs), where a single nucleotide is substituted for one of the other 3 bases. Single base changes may also involve insertion or deletion of one

nucleotide, and these are collectively referred to as point mutation. Furthermore, insertions and deletions (indels) can span across multiple bases, where small indels of up to 30 base pairs and point mutations in DNA represent the majority of human genetic variations [57]. Other more complex mutations include a combination of substitutions, deletions and insertions (multi-nucleotide polymorphisms), repeated bases (duplication), and variants on longer stretches of DNA sequences. These larger scale variants may involve deletion and amplifications at regions in the size of an exon, a gene, or even a chromosome, and are commonly termed copy number variants (CNV). They could also contain structural variants (SV), such as chromosome rearrangements, translocations and inversions.

Human cells acquire DNA mutations over the lifetime of an individual. These alterations are an essential part of evolution, resulting in genetic diversity with different heritability traits and adaptability to a changing environment [58]. Mutations could start in normal cells that are repetitively damage by mutagens and failed to be fixed by DNA repair mechanisms [59]. Some of these damages are introduced by errors during DNA replication or external environmental agents. These external carcinogens, such as tobacco products, or radiation from ultraviolet lights, contribute to increase rates of mutations [60]. Accumulation of mutations increases the chance of a cell acquiring a selective advantage that allow it to proliferate faster with stronger survival mechanisms than its neighbours [6]. This natural selection within cells and their recurring mutations eventually give rise to cancer.

2.2.1.1. Germline and somatic mutations

Genetic mutations can be inherited from parents or acquired during one's lifetime. Inherited changes are present in germ cells and are called germline mutations. Germ cells propagate parents' genetic information to offspring and their mutations are found in every cell, including somatic cells that form the body of organisms [61]. Further mutations in somatic cells are acquired after conception and not passed to next offspring, but has the potential to turn into cancerous cells [6]. These mutations can only be found in certain cells and are called somatic variants. Both germline and somatic mutations may be associated with cancer initiation and progression.

The presence of germline mutations increases cancer susceptibility. Multiple genes with germline mutations, also known as cancer predisposition genes (CPG), have been investigated to be associated with multiple cancer types [62]. Following the two-hit hypothesis [63], these cancer-predisposing mutations are typically not fatal if the second allele is normal. However, when a subsequent somatic mutation occurs in the second allele, the predisposed genes is now totally inactivated and can potentially lead to tumorigenesis [64]. Examples of CPGs are *BRCA1* and *BRCA2* genes, where occurrence of germline mutations in these genes are associated with elevated risks of hereditary breast cancer [65]. The cancer risk assessment through genetic screening is critical for managing and preventing early-onset familial cancers.

Cancers emerge as a result of the accumulation of somatic mutations in the DNA. These mutations could be acquired during normal cell division, due to factor such as defective DNA repair process, or environmental exposures to carcinogenic mutagens. The effect of external agents, such as smoking, are not apparent until long period of exposures. This gradual process of tumour progression involve many years of natural selection cycles from randomly inherited

alterations in abnormal cells [6]. This phenomena could be the reason why the number of accumulated mutations show correlation with the age of patients [66]. Furthermore, age factor could also explain why most adults cancers have more mutations than paediatric tumours [67].

2.2.1.2. Single-nucleotide variation

Single-nucleotide variation (SNV) involves a single base changes in DNA that could occur in coding or non-coding regions of genes. Depending on their effects on protein sequence, SNVs within a coding region may be synonymous and non-synonymous. Non-synonymous variants result in a change in the translated amino acid, and it can be further divided into missense and nonsense mutations. In most solid tumours, majority of the somatic non-synonymous mutations are single-base substitutions, and 90% of these are missense mutations[67] . Figure 2.4 describes other variants consequences in the context of transcript structure [68].

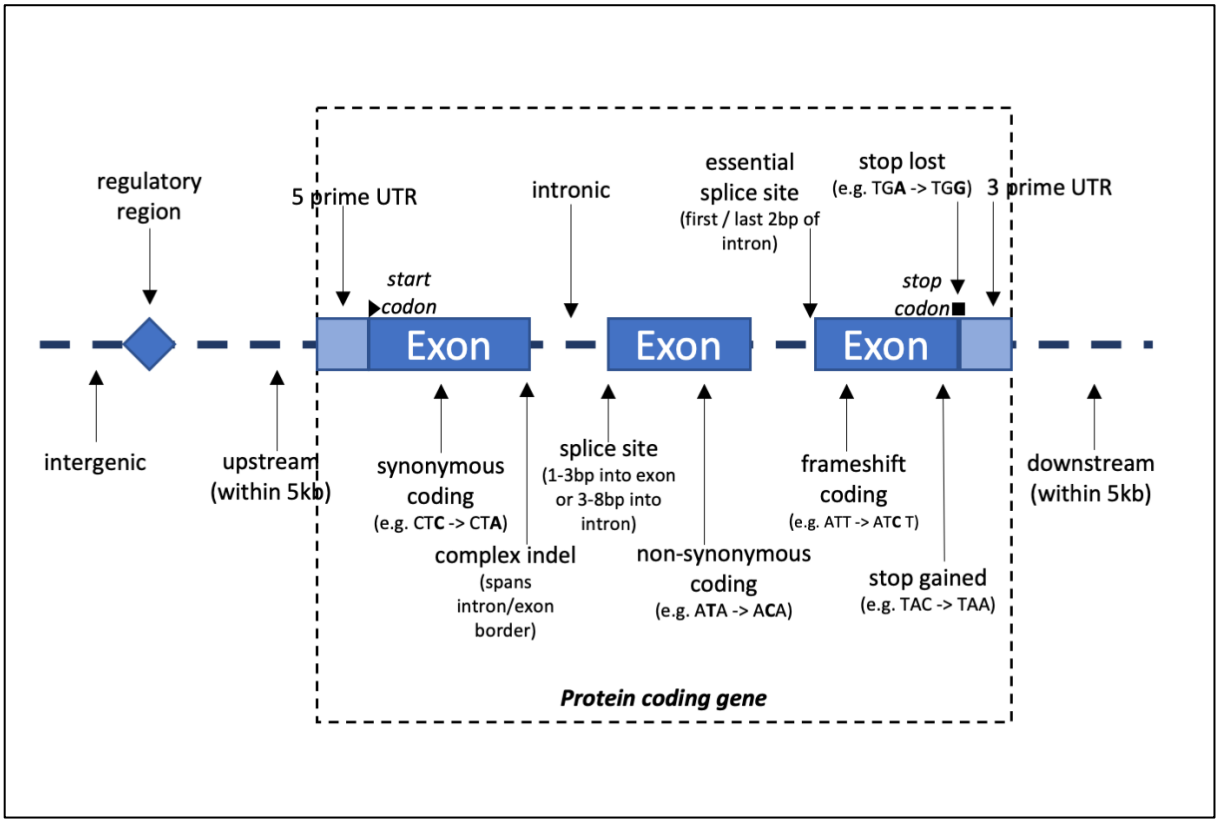


Figure 2.4 - Possible genomic variants consequences provided by Ensembl (adapted from [68])

2.2.1.3. Copy number variation

A copy number variation (CNV) is a large rearrangement that alter DNA diploid status. This variation includes amplification and deletion of a large chromosomal regions. An amplification event can lead to a particular gene being duplicated into multiple copies, resulting in an abnormally active protein product of the gene, whereas a deletion may result in inactivation of a tumour-suppressor gene [67]. Although the detail of CNVs is similar to small indels, the

main distinction is the size of the alterations, where CNVs represents changes of 1 kilobase or larger [57].

A gene copy-number variations is associated to the changes to the gene expression levels and the subsequent encoded proteins. A study by Zhao et al. [69] found that a copy number loss in tumour suppressor genes, *PTEN*, *MTAP*, *MCPH1* and *SMAD4*, are correlated with low gene expression of these genes. Conversely, an *ERBB2* amplification in breast cancer is associated with an upregulated expression of that particular gene [70]. This relationship highlights CNVs important role in the overall tumour progression.

2.2.1.4. Gene expression profiles

Cancer genes are commonly characterised by their differential expression. Gene expression can be defined as the process that transcribed DNA information into protein or RNA structures via messenger RNA (mRNA). Transcription level can be quantified by detecting the abundance of mRNA through sequencing methods [71]. This transcriptomes analysis assists in the identification of genes that are active or inactive in a cell, and can be applied for assessing functional impact of mutations and individual's response to treatments.

The gene expression levels differ between cells and individuals. Studies have found that point mutation and large chromosomal rearrangement play a role in the alterations of transcription level [67]. However, the number of genes that are differentially expressed in cancer are more than its mutated genes [72]. These gene expression changes could be due additional factors such as epigenetic changes. These modifications are driven by chromatin structure alteration, but do not modify cell's DNA sequences [6]. The variations in gene expression, whether it is overexpressed, underexpressed, or altered epigenetically, may initiate abnormal cells growth, enhanced tumour progression, or inhibit cells' DNA repair process.

The advent of sequencing has enabled the quantification of multiple genes expression at once. This gene expression profiling is applied to assess cells activities at any given time points. The fundamental analysis of expression profiles often involves the mRNA counts from more than one experimental or biological conditions [73]. Statistical analysis is then applied to test whether the differential expressions between the groups are significant [74]. Gene expression patterns have also been used to study the heterogeneity in cancers. Sorlie et al. [75] demonstrated that breast cancer can be categorised into multiple subtypes based on their gene expression patterns. These different subtypes have been studied to display different biological characteristics and prognosis profile [75-79]

2.2.2. Cancer genes

Cancer cells contain many somatic alternations, but not all of these somatic abnormalities are involved in tumorigenesis. Many of these *passenger* mutations do not directly contribute to tumour progression; they are rather accidental by-product from other genomic changes. The main *drivers* to the oncogenic events give selective growth advantage to the cancer cells. Differentiating these driver mutations from the passenger alterations is not straightforward

and has been one of the major goals of cancer studies. Despite the challenges, a large number of somatic mutations in human cancers have been comprehensively annotated [80, 81].

Over the years, many genes that contain driver mutations have been examined in regards to their role in cancer causation. These cancer critical genes, which are also referred to as *driver genes*, can be broadly categorised into two groups. The first group is called oncogenes, where a gain-of-function mutation on these genes promotes abnormal cell growth [6]. The other group is called tumour-suppressor genes, where a loss-of-function mutation in these genes results in unregulated proliferation and apoptosis [6]. In both cases, genetic changes in any of these two groups of genes have the potential to directly give rise to cancer.

Mutations that activate specific oncogenes stimulate added growth promoting effect. Activating mutations, such as missense mutations on oncogene *PIK3CA,* drive tumorigenesis through activating different molecular pathways[82, 83]. Similarly, amplification in *MYC* oncogene results in the gene being overproduced [6]. As a key regulator of up to 15% human genes, *MYC*-activation has big influence in tumorigenesis in numerous cancer types [84]. Some other examples of oncogenes that have been extensively studied in literature include *AKT1*, *BRAF*, *KRAS*, *JAK2*, and *MAPK1* [85].

Tumour suppressor genes guard against cancer and put a stop on unnecessary growth. For instance, loss-of-function mutation in a tumour suppressor gene *RB1* results in the disruption of normal cell-cycle control [6]. This tumour suppressor gene inactivation could also affect pathways that regulate stress and DNA damage response, such as *p53*. *TP53* is one of the most frequently mutated gene and have been observed to occur in nearly all human cancers [86]. Some other examples of tumour suppressor genes that have been extensively studied in literature include, *ARID1A, CDH1, BRCA1, BRCA2* and *PTEN* [85].

2.2.3. Genomics sequencing

DNA sequencing is a method to infer the exact order of nucleic acid sequence in a DNA, which consists of four bases: adenine (A), guanine (G), cytosine (C) and thymine (T) [2]. Human genome consists of more than 3 billion base pairs, organised into 23 pairs of chromosomes. It took more than 10 years for the first complete human genome to be sequenced by The Human Genome Project [1].

Sequencing technologies have gone through rapid evolution since its first introduction. In 1977, Frederick Sanger developed a sequencing method using DNA polymerase and chain-terminating nucleotides that transformed the study of genomics [6]. This sequencing technique, which is largely known as Sanger sequencing, was the most common method used by researchers and is classified as the first generation of DNA sequencing [87]. This sanger-based sequencing was the base of the sequencing approach use by The Human Genome Project that was started in 1990, involving researchers from multiple countries and cost almost three billion dollars on its completion in 2003 [2]. However, despite being revolutionary, sanger-based sequencing is expensive and generate low throughput.

In 2005, the second-generation sequencing technologies that supports massively parallel sequencing start to emerge. Commercial technology such as Roche's 454 promised to deliver high throughput sequencing for a fraction of the cost [87]. These new technologies, commonly refer to as Next Generation Sequencing (NGS), are able to sequence multiple of samples in parallel in a day and effectively bringing the cost of sequencing down to one thousand dollars per human genome [2]. Current leading NGS sequencing platforms relies on template enrichment, where DNA fragment libraries that are generated by polymerase chain reaction (PCR) amplification and sequenced in parallel [88, 89]. Presently, Illumina platforms are the most widely used NGS technology [90].

Since the introduction of NGS, it has become the preferred technology for whole genome sequencing (WGS). However, as interpretation of variants non-coding regions are still challenging, alternative approaches for sequencing smaller targeted regions are often considered for diagnostic and research purposes [91]. These approaches include whole-exome sequencing (WES), which aims at sequencing all the exons regions, and targeted sequencing that focusses on specific panels of genes of interest. By reducing the size of genomic regions offers greater sequencing depth and lowers the cost per sample [89]. This higher coverage sequencing is required for identification of low frequency or rare variants[92], that could require as deep as 10,000x coverage [89].

The rapid development of NGS platforms also benefits transcriptomic research. RNA-seq, which sequence the whole transcriptome using NGS technology, has replaced microarrays for gene expression studies [88]. In comparison to RNA-seq, microarrays have few disadvantages, such as cross-hybridization biases and limitations due to probes design [93]. Moreover, RNA-seq is also capable of detecting rare transcripts, splicing events and allele-specific expressions [93]. However, despite their differences, gene expression quantification from both platforms are concordant [94].

The advancement of sequencing technologies is still actively progressing. Third generation sequencing based on single-molecule real time sequencing addresses PCR amplification bias of NGS [95]. These technologies include platforms developed by Oxford Nanopore Technologies (Nanopore) [96] and Pacific BioSciences (PacBio) [97] , which both enable longer reads to be sequenced. These developments also further refined the traditional bulk-transcriptomic sequencing towards the analysis of individual cells. Popular single-cell RNA-seq technology development, such as the one developed by 10x Genomics, enables the analysis of rare cells populations and cellular heterogeneity in tumour [98]. This sequencing innovation has the potential to further improve our understanding in drug responses and tumour cell resistance.

2.3. Breast Cancer

Cancer is one of the leading causes of deaths in the world. It is estimated to be responsible for around 9.6 million deaths globally in 2018, the second-highest death-causing diseases [99]. There are approximately 17 million new cancer cases worldwide in 2018 [99]. In Australia, there are more than 1 million people currently suffering or have suffered from cancer, and it instigated 3 of every 10 Australian deaths in 2016 [100]

Cancer is a general term used for a wide range of related diseases. It is characterised by abnormal cells that are invasive, proliferate and grow out of control [6]. Cancer is traditionally named and classified by its site of origin, or according to the type of cells and tissue where they originated [6]. For example, breast cancer starts in the cell of the breast tissue. Overall, there are more than 100 different cancer types [101], and each of these cancer types has different characteristics, treatment options and survival rates [6]. The most diagnosed cancer types globally in 2018 are lung and breast cancer [99].

Breast cancer is the most common type of cancer for female globally [99]. It is estimated to be the second-highest cancer-related deaths for Australian women in 2019 [100]. Men can also be diagnosed with breast cancer, although the incidence rate is much lower than women at approximately 1% of the overall number of cases [102]. The relative survival rate for breast cancer has improved over the last 30 years, with a 5-year survival rate as high as 90% [100]. This progress has mainly been attributed to improved treatments and early detection [100]. For cases that have been detected at a later stage (stage IV), however, breast cancer has a reported 5-year survival rate of only around 32% [100]. There are also observed variations of incidence and mortality rate of breast cancer in higher-income countries as well as different age groups [103]. Thus, there is a fundamental need to improve treatment and diagnosis methods to conquer this disease, as it remains a significant health burden worldwide.

2.3.1. Breast Cancer Classifications

Breast cancer is a heterogeneous disease and has been classified to different subtypes to help identify the appropriate diagnosis and personalised treatment [104]. Multiples studies have looked at how groups of breast cancer with various pathological and molecular features show different characteristics despite being originated from the same organ [105]. They will often have distinctive clinical presentations, different risk factors and responses to particular treatments as well as different prognosis profile [104]. Some of the factors that have been used to define breast cancer subtypes include histological grade and type, hormone receptors status and molecular profile based on gene signature assays [106]. The recent advancement of cancer research and improved numbers of screening have allowed a better understanding of these different subtypes of breast cancer.

2.3.1.1. Subtypes based on Histopathology

Histologically, breast cancer can be broadly classified into two types. The first type, in-situ breast carcinoma, is non-invasive breast cancer that has not spread to other parts of breast tissue [107]. In contrast, invasive breast carcinoma is the type used to describe any malignant breast cancer with cancerous cells that have proliferated into surrounding breast tissues [107].

In-situ breast carcinoma can be further sub-categorised based on the tumours' originating site. Ductal carcinoma in-situ (DCIS) originates from the ducts and is more common than lobular carcinoma in-situ (LCIS) that begins from the lobules. DCIS is also more heterogeneous and can be further sub-divided into Comedo, Cribiform, Micropapillary, Papillary and Solid [104].

Similar to the non-invasive counterparts, invasive breast carcinoma also comprises of heterogeneous groups of tumours. The most common subtype from this group is infiltrating ductal carcinoma (IDC) that comprises about 80% of invasive breast cancer cases. Other subtypes in invasive breast cancer subtypes are Tubular, Ductal Lobular, Invasive Lobular, Mucinous (colloid), Medullary and Papillary [104]

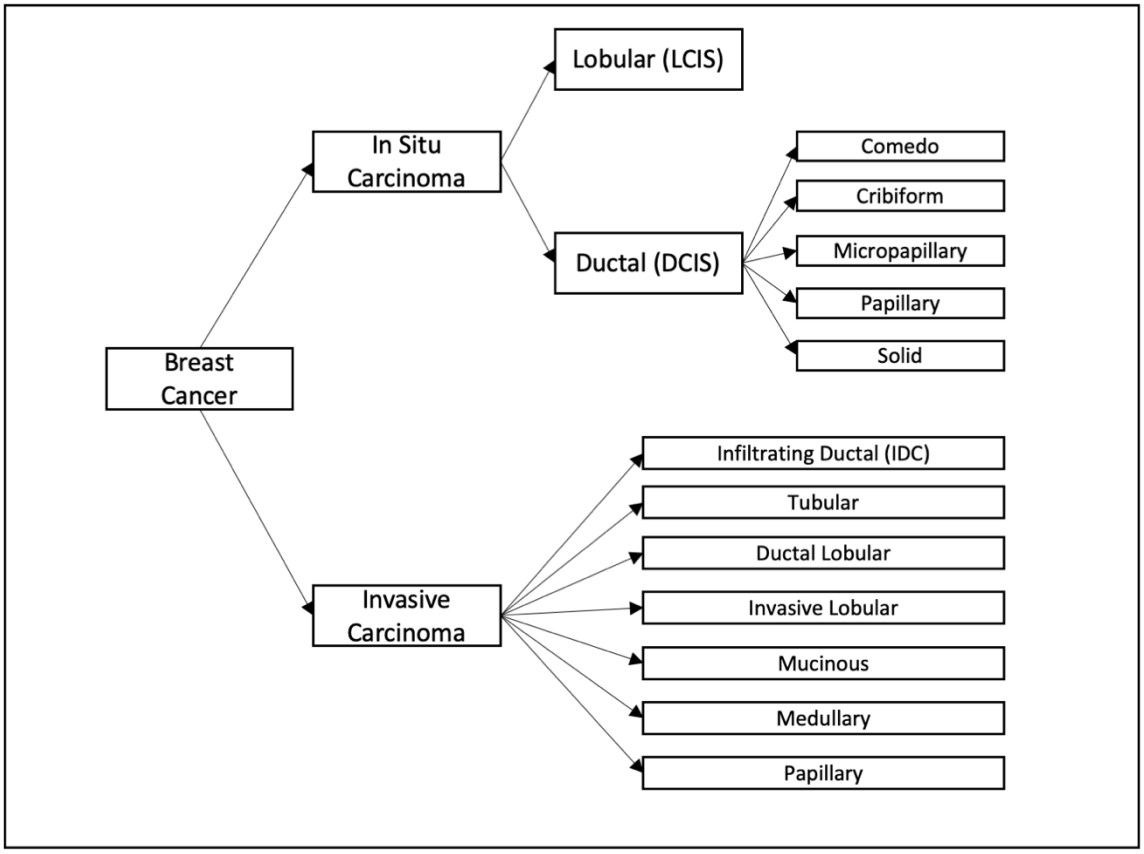


Figure 2.5 - Breast cancer subtypes based on histopathology

2.3.1.2. Subtypes based on Molecular Profile

The rise of large-scale genomic studies and gene expression profiling has enabled researchers to further analyse the molecular characteristics of different breast tumours. These studies had identified at least four main breast cancer intrinsic subtypes through unbiased hierarchical clustering of gene expression patterns among the samples [76]. The primary characteristics of the subtypes are based on the expression levels of estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2) and proliferation indicator Ki67 [75-77]. ER-positive (ER+) breast cancer, which commonly exhibits a high expression of luminal-enriched genes, are further divided into luminal A and luminal B based on the expression level of ER genes and proliferation indicator Ki67 [75, 77]. On the other hand, ER-negative (ER-) breast tumours are further sub-divided into two groups, HER2-enriched and basal-like. HER2+ tumours are characterised by the overexpression of *ERBB2*, while basal-like subtype is negative for all three receptors (ER, PR and HER2) [75-77]. A more recent study by Prat et al. in 2007 has further sub-classified the major breast cancer intrinsic subtype of ER- group into another distinct cluster, called claudin-low, where its profile is similar to triple-negative breast cancer [79].

Molecular subtypes of breast cancer have been studied to show different clinical characteristics, risk factors, treatments response and survival rates [104, 105]. The ER+ group displays a better response to hormonal therapy, such as tamoxifen, and in general, has higher 5-year overall survival than ER- tumours [104, 108]. Luminal A is the most common subtype of breast cancer, which are observed in about 40% of breast cancer cases, and has the best prognosis among the four subtypes [108]. HER2-enriched tumours proliferate faster with low survival rate and occur in 5%-15% of overall breast cancer cases. However, it is more likely to respond to HER2-targeted therapy, such as trastuzumab or lapatinib [108]. In contrast, triple-negative breast cancer (TNBC) still has a few efficient targeted treatments and is associated with the worst prognosis among all breast cancer subtypes [108]. TNBC, which includes most basal-like and claudin-low tumours, accounted for approximated 20% of all breast cancers, common in younger patients as well as women of African descent [108].

2.3.2 Molecular Subtypes of Breast Cancer

2.3.2.1 Luminal A

Luminal A breast cancer is characterised by relatively high expression of oestrogen and progesterone genes [106]. Tumour in this group is also observed not to have a high-level *ERRB2* gene amplification [76] but has elevated expression of luminal epithelial genes, such as *ESR1*, *GATA3* and, *FOXOA1* [77]. In terms of that mutational profiles, luminal A samples have the most potential driver genes, with *PIK3CA* being the most frequently mutated genes, followed by *MAP3K1*, *GATA3*, *TP53*, *CDH1* and *MAP2K4* [109]. In comparison to luminal B, this subtype also has a lower frequency of TP53 mutations [109]. Overall, the mutation rate in luminal A is the lowest among all breast cancer subtypes [109], and it is also the least likely to metastasize [110]. For its targeted therapy, anti-estrogen treatment such as tamoxifen or aromatase inhibitors were often used for luminal A patients [104]. Treatment option with tamoxifen has been studied to decrease 15-year risks of breast cancer relapse and mortality rate by about a third [111]. Among the subtypes, luminal A breast cancer has the best prognosis [108].

2.3.2.2. Luminal B

Similar characteristics to Luminal A, Luminal B breast cancer is characterised by luminal genes that are often overexpressed [77]. The main differences between these two luminal groups are the level of expression of ER genes, which tends to be low to moderate for luminal B tumours, having a higher level of proliferation indicator Ki67, and can be either HER2+ or HER2- [75, 108, 110]. Luminal B is also observed to show a higher frequency *TP53* mutation and *CCND1* amplification than luminal A [109]. Furthermore, luminal B tumours are often defined as the more aggressive subtype of ER+ breast cancer. It is associated with higher histological grade tumours and moderate prognosis that are worse than luminal A patients [108, 112].

2.3.2.3. HER2-enriched

HER2-enriched breast cancer tumours are characterised by *ERBB2* amplification and the absence of ER gene overexpression [108]. It is important to note that HER2-enriched is not

identical to the clinically defined HER2-positive breast cancer, although more than half of HER2-positive breast cancer are HER2-enriched [110]. Apart from HER2 amplification, *TP53* is also frequently mutated in HER2-enriched breast cancer [109]. Other potential abnormalities identified in this subtype, include higher expression of *EGFR*, loss of *PTEN* and *MED1* co-amplification [109, 113].

HER2-enriched breast cancer is more aggressive than luminal breast cancers and has higher rates of metastasis [108]. It has the highest somatic mutation rate [109] and also proliferates faster than ER+ tumours. There have been multiple HER2-targeted therapies introduced, and this has significantly enriched the outcomes of patients in this group [110]. This includes targeted treatments, such as trastuzumab (Herceptin) and lapatinib [104]. Overall, HER2-enriched breast cancers tend to have a worse prognosis compared to luminal breast cancers [114].

2.3.2.4. Triple Negative Breast Cancer and Basal-like

Triple-negative breast cancer (TNBC) tumours are characterised by the lack of expression of ER, PR and HER2 [76]. Although not all TNBC tumours are basal-like, a majority of basal-like tumours exhibits triple-negative phenotype [115]. Basal-like breast cancer is observed to have a high prevalence of *TP53* mutations as well as deletion of *RB1* and *BRCA1* [109]. In addition, germline mutations in cancer predisposition gene, BRCA1, are largely associated with the basal-like subgroup [77]. Other genomic aberrations that are linked to TNBC and basal-like breast cancer include frequent *MYC*-amplification [116], higher genomic copy number count of KRAS-linked genes [78], low expression of CDH1 [79] and high expression of *EGFR* [117]. Overall, TNBC displays high genomic instability with a higher mutation rate than other breast cancer subtypes [108].

TNBC is one of the most aggressive subtypes of breast cancer. It exhibits a higher rate of relapse, as well as frequent brain and lung metastases [110, 118]. There is a limited number of targeted therapies available for TNBC patients, and conventional chemotherapy remains as the most standard treatments [108]. However, recent studies have suggested that the use of PARP inhibitors, such as olaparib and talazoparib, shows good potential for BRCA-mutations carrier TNBC patients [119]. The presence of high tumour-infiltrating lymphocytes (TIL) is associated with a better prognosis for early TNBCs patients [120], indicating the critical role of the immune system for the outcome of this subtype. In general, TNBC and basal-like have one of the worst survival rates among all the breast cancer subtypes [77].

2.3.3. Molecular Profiling Assays in Breast Cancer

There are two main methods for classifying breast cancer patients into subtypes for guided treatments: immunohistochemistry (IHC)-based markers and gene-based assays [106]. The IHC-based subtype has been adopted for measuring clinicopathological criteria, including the levels of ER, PR, HER2, and Ki-67 index [121]. On the other hand, gene-based assays analyse the expression of several genes to determine the intrinsic subtypes of breast cancer [122]. IHC-based has been applied more frequently for clinical use, as the cost of running gene expression arrays are still cost-prohibitive [122].

There are multiple molecular assays available, and each is targeting a different set of genes. Some examples of these frequently used gene expression assays are OncotypeDX (21-gene recurrence score), PAM50 (50-gene signature), MammaPrint (70-gene signature) and BluePrint (80-gene signature) [106, 114, 123, 124]. Subtypes identified by these gene expression assays do not always conform with IHC. Between IHC-based method and MammaPrint/BluePrint, it has been observed that there is more than 25% discordance [106], while as much as 38.4% identified subtypes do not agree between IHC-based subtype and PAM50 [125]. Some samples could also comprise of more than one subtypes characteristics, and this intra-tumour heterogeneity contributes to the inconsistencies between subtyping system [126-128]. In addition, there are also limitations when they are used solely as a classifier. For example, PAM50-classifier do not work well if datasets ER-status is imbalance [129]. However, these molecular profiling assays are largely prognostic and predictive, which may contribute to benefit patients in selecting potential targeted treatments [106, 114].

2.4. Large-Scale Breast Cancer Genomics Datasets

This section describes the data sources and data pre-processing methods used in this project.

2.4.1. The Cancer Genome Atlas

The Cancer Genome Atlas (TCGA) [4] is a comprehensive cancer study that uses the advances of genomics to extend our knowledge of human cancer. TCGA studied 33 different cancer types from 11000 patients, generating 2.5 petabytes of data that was made publicly available to the research community. This data includes information from various types of analysis including results from DNA and RNA sequencing, copy number, array-based expression, DNA methylation, SNP array, reverse phase protein assays as well as samples' clinical information. For breast cancer, there are a total of 817 samples available in TCGA [109, 130]. This includes 127 invasive lobular carcinomas (ILC), 490 invasive ductal carcinomas (IDC) and 88 mixed IDC / ILC. These samples cover all PAM50 breast cancer intrinsic subtypes.

The TCGA breast cancer dataset used for this thesis were downloaded from cBioPortal [11, 131]. The following section details the type of data used for this thesis.

2.4.1.1. TCGA gene expression data

The raw mRNA expression data was derived from RNA sequencing that was processed on Illumina HiSeq. TCGA [130] aligned these sequencing reads to hg19 reference genome using MapSplice[132], followed by gene expression quantifications and normalisation with RSEM [133].

2.4.1.2. TCGA somatic mutation data

The somatic mutation data was derived from whole-exome sequencing that was processed on Illumina Hi-Seq 2000 with Agilent SureSelect All Exome v2.0 kit or Nimblegen SeqCap EZ Human Exome v2.0 [109]. The analysis pipeline used to generate this data is described to be as follow [130]:

- Sequencing reads were aligned to hg19 with BWA and pre-processed with Picard and samtools.
- Somatic variants were called using multiple variants callers, including samtools, somaticSniper, Strelka, VarScan, GATK and Pindel.
- Variants were annotated with GENCODE from Ensembl 69
- Variants were filtered with various rules, including minimum coverage, variant allele fraction, and removing recurrent artifacts and common germline variants.
- Final somatic mutation data is stored as maf file.

2.4.1.3. TCGA somatic copy number data

The somatic copy number data was derived from Affymetrix SNP 6.0 arrays [130]. The author described that individual copy number calls were obtained through Circular Binary Segmentation, and GISTIC 2.0 algorithm was used to identify significant putative copy number alterations. The resulting CNV data used for analysis has five values: -2 (homozygous deletion), -1 (hemizygous deletion), 0 (neutral), 1 (gain), 2 (high level amplification).

2.4.2. Molecular Taxonomy of Breast Cancer International Consortium

Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) [134] is a comprehensive breast cancer study from over 2000 primary tumours. In 2012, this study analysed gene expression profiles and copy number changes from patients in the UK and Canada [134]. In 2016, METABRIC published a further study on these samples through targeted sequencing of 173 frequently mutated breast cancer genes to understand the somatic mutation profiles of this cohort [9] (the list of these 173 genes is available in the Appendix). METABRIC comprises of breast cancer samples from different molecular subtypes where clinical data and long-term follow-up information were mostly available, including morphological assessment, hormone receptor status, age and prognosis information.

For the analysis specified in Chapter 3 and Chapter 4, all the genomic, transcriptomic, and clinical data from METABRIC were downloaded from cBioPortal[11, 131]. The following section provides more details about the origins of these omics data.

2.4.2.1. METABRIC gene expression data

The transcriptomic profiles were obtained through microarray data on Illumina HT-12 v3 platform. It has been pre-processed and normalised by METABRIC with custom R script and limma Bioconductor package [134].

2.4.2.2. METABRIC somatic mutation data

The somatic mutation data was obtained from targeted sequencing on Illumina HiSeq 2000. METABRIC aligned the reads from this sequencing with Novoalign, pre-processed with GATK and variants were identified with MuTect and HaplotypeCaller, filtered with custom rules (mapping quality, read direction bias, variant allele frequency and variant caller's quality scores) and are then annotated with ANNOVAR. Final mutation data has also been further

filtered with the use COSMIC, 1000 Genomes database and normal samples in the dataset to minimise false-positive coming from sequencing bias and germline variants [9].

2.4.2.3. METABRIC somatic copy number data

The Somatic copy number data was obtained through genomic profiling on Affymetrix SNP 6.0. METABRIC used ASCAT to call copy number segments and GISTIC 2.0 to analyse recurrent somatic copy number aberrations. Thresholds for copy number gains and losses were set at log2 ratio of 0.4 and 0.5 [134].

2.4.3. Data pre-processing

TCGA and METABRIC datasets that are retrieved from cBioPortal are further filtered to ensure consistencies between analysis. Samples that do not have all three genomic and transcriptomic profiles (gene expression, copy number and somatic mutation) are excluded. For subtype-specific analysis, only samples that belong to the four intrinsic subtypes by PAM50 (Basal-like, HER2-enriched, Luminal A and Luminal B) will be included.

Gene expression profiles from METABRIC and TCGA were obtained from different sequencing platforms. To ensure the assessment between the two profiles are comparable, the relative expression (z-score transformed) is used as the final inputs to the machine learning algorithms. This z-score is provided by cBioPortal, where the distribution of the diploid samples in the datasets is used to further normalised the RSEM-calculated gene expression estimates.

Chapter 3 - Evaluation of Driver Gene Prediction Algorithms

3.1. Summary

Large-scale sequencing of cancer patients has enabled researchers to study more in-depth into the relationship between different genomics variables. One of the primary objectives of these studies is to identify the main drivers of tumorigenesis that is critical for finding potential treatments for cancer patients. Multiple computational methods for predicting driver genes have been proposed and claimed to predict crucial genes that are involved in promoting oncogenic activities. However, a thorough evaluation of these algorithms' performance is still challenging as the complete cataloguing of all cancer driver genes is still an ongoing effort.

This chapter focuses on assessing various driver gene prediction algorithms in the context of breast cancer. The final detailed evaluation includes a total of five computational methods that target different characteristics of driver genes. All algorithms analyse the same sets of breast cancer public datasets, and their performance was assessed on the predicted driver genes. These performance metrics include comparisons to cancer gene census (CGC), concordance between the algorithms, consistency of prediction across different datasets and number of samples, and driver genes identified on breast cancer subtype-specific datasets.

The evaluation results described herein suggest the best algorithm for driver gene identification are likely to be driven by the aims and datasets of the study. Network-based algorithms are better at identifying driver genes that have a low frequency of point-mutations, and this might benefit the discovery of novel driver genes. However, this could also lead to a higher number of false-positive predictions, as evidenced by their lower number of overlaps with CGC. On the other hand, methods that analyse only somatic mutations data, require a larger number of samples to be explored. Although these algorithms excel in identifying established cancer genes, their predictions were inconsistent across different datasets. They also could not analyse some breast cancer subtype-specific datasets reliably due to the smaller sample size. Nevertheless, genes that are identified by at least three methods are all well-studied genes, suggesting that a consensus-based prediction could theoretically be an excellent approach for high sensitivity prediction. Overall, the assessment presented in this thesis demonstrated that all five evaluated methods are capable of identifying important cancer genes when applied to a suitable dataset, and their application in future cancer studies could potentially help the discovery of novel driver genes.

3.2. Introduction

Rapid genomics technologies advancement in the last decade has produced an enormous amount of clinical omics data across different levels of variables. These unprecedented amounts of data require techniques for data mining, integration and interpretation to provide more insights into cancer genomics fields. One of the main goals of studying these data is to identify the main drivers of cancer progression to differentiate them from the sea of passenger mutations [135]. Tumour growth is often associated with abnormal function of proteins and changes in gene expression, which can often be due to many types of genetic

and epigenetic variations. Gene copy number changes, DNA methylation, single nucleotide variation, small insertions and deletions are some of the main example of variations that can impact gene function [136].

There are two main types of mutated genes that drive tumorigenesis. The first type is called *oncogenes*, where a gain-of-function mutation on these genes leads to the cancer cell proliferating out of control [6]. On the other hand, a loss-of-function mutation in *tumour-suppressor genes* results in unregulated cell growth and apoptosis [6]. Apart from directly impacting the expression of the mutated genes, these mutations can also contribute indirectly by causing genomic instability [6]. Collectively, these mutations are referred to as driver mutations.

The definition of driver genes, however, is slightly more complicated. It is generally accepted that a driver gene contains one or more driver mutations, and potentially also harbour many passenger mutations [67]. Though this definition distinguishes mutations that cause positive selection, it does not explain other well-studied driver genes that have been observed to have no or very few driver gene mutations[67]. This includes genes with epigenetic changes, which have been studied to show different genes expression patterns in tumours [137]. Both genetic mutations and epigenetic alterations contribute to cancer initiation and progression, and therefore, it is essential to consider both driver events.

Numerous computational algorithms for detecting driver genes have been proposed. These methods are commonly grouped into different categories, depending on the targeted driver mutations' features [138], the number of data integrated [136], or their design principles [139]. The most common approaches were looking at driver genes as the significantly mutated genes that have a higher mutation rate than their background frequency. This includes widely used methods such as MuSiC [140] and MutSigCV [141]. However, this approach has limitations in difficulty of estimating the background mutation rate accurately due to the heterogeneity of tumours. Various improvements were proposed to address this limitation, including the use of prior knowledge. Algorithms, such as ActiveDriver [142], OncodriveFM [143], and OncodriveCLUST [144], incorporate functional genomic positions into their predictions. As functional importance is not available from the raw mutation data, they are often retrieved from external databases, or through observing positional clustering patterns [138]. Other alternative strategies employ a gene interaction network to improve the prediction of less frequently mutated genes. These network-based methods, such as DriverNet [135], DawnRank [145], and OncoImpact [146], assess not only somatic mutation information but also gene expression and copy number variations data. Integrating multiple omics data allow these algorithms to pick up other driver events, such as amplification, deletion or epigenetic drives. Recent developments further refine some of these ideas through supervised machine learning approaches. For example, 20/20+ [147] combines multiple signals of positive selection on its random forest models for predicting driver genes and deepDriver [148] applies convolutional neural network to integrate gene expression patterns and somatic mutations information to detect potential driver genes.

Table 3.1 summarises recent driver gene predictions tools available in the literature. From these algorithms, this thesis selected five methods for final evaluation, aimed at covering multiple different driver gene prediction methods and principles. Other selection criteria

include the type of integrated data and accessibility of the tools, such as software availability, installation of dependency packages, provided test data and runtime. The final five algorithms that were evaluated in great details are DriverNet, DawnRank, OncodriveFML, OncodriveCLUSTL and 20/20+.

Table 3.1 - Summary of the recently published driver gene predictions tools

Name	Strategy	Data	Methods' key principle
MuSiC [140]	Significantly mutated genes	- Somatic mutations	Estimates background mutation rate through statistical tests (significantly mutated gene test)
MutSigCV [141]	Significantly mutated genes	- Somatic mutations - Gene expression	Estimates background mutation rate from covariates data, obtained through gene expression and DNA replication timing data
ActiveDriver [142]	Functional genomic positions	- Somatic mutations	Estimates background mutation rate on protein-phosphorylation regions
OncodriveCLUST [149]	Functional genomic positions	- Somatic mutations	Detects abnormal clustering of mutations within protein-sequence sites
OncodriveCLUSTL [144]			
OncodriveFM [143]	Functional genomic positions	- Somatic mutations	Detects bias towards high functional impact mutations
OncodriveFML [150]			
DriverNet [135]	Network-based	- Somatic mutations - Gene expression - Copy number variations	Constructs bipartite graph to assess the impact of mutated genes in the gene interaction network
DawnRank [145]	Network-based	- Somatic mutations - Gene expression - Copy number variations	Adopts PageRank algorithm for identifying genes that regulates expression of other genes. Driver genes are then selected based on pairwise comparison/voting method based on the mutation status.
PARADIGM-SHIFT [151]	Network-based	- Somatic mutations - Gene expression - Copy number variations	Utilises pathway information to assess mutational event by detecting the difference in gene activity from the downstream and upstream neighbourhood (belief-propagation algorithm)
OncoImpact [146]	Network-based	- Somatic mutations - Gene expression - Copy number variations	Groups genes into expression modules based on gene interaction networks. Driver mutated genes are selected by assessing the impact score of these expression modules.
CHASM [152]	Machine learning	- Somatic mutations	Identifies driver events through a random forest classifier, trained on curated and estimated sets of driver and passenger mutations
CanDrA [153]	Machine learning	- Somatic mutations	Performs supervised training using support vector machine (SVM) for classifying driver mutations
20/20+ [147]	Machine learning	- Somatic mutations	Predicts driver genes through a random forest model, trained on curated sets of oncogenes and tumour suppressor genes
deepDriver [148]	Machine learning	- Somatic mutations - Gene expression	Applies convolutional neural network (CNN) on features derived from mutation and gene expression data

3.3. Methods

3.3.1. Driver Gene Prediction Algorithms

3.3.1.1. DriverNet

DriverNet [135] is a network-based driver gene prediction algorithm that integrates genome and transcriptome data from large-cohort tumour studies by utilising known biological pathway information. Its main algorithm evaluates how tumour gene expression networks will be affected by somatic aberrations. The method is formulated along with the assumption that driver genes will have strong effects on the expression of many genes. DriverNet accepts somatic aberrations data in the form of somatic mutation or copy number variations, gene expression profile and integrating them through a predefined gene interaction network. This biological pathway information is based on prior knowledge of gene pathways that can be obtained from sources such as Reactome[154] and KEGG [155]. Somatic mutations are defined as a binary matrix of all possible genes against samples in the datasets. For example, Gene A in Sample 1 will be assigned value 1 if there is a mutation or copy number changes for that particular gene or value 0 when there is no variation. Interaction between somatic aberrations and expression modules were defined using bipartite graph where a node on the left side represents a mutation status of a gene and gene with outlying expression status is represented by a different node on the right side of the graph. Edges between two nodes from different sides are connected if (1) left node is a gene with a mutation (2) right node is a gene that shows outlying expression, and (3) if both genes are known to interact according to the prior knowledge graph. Driver genes will be selected by using a greedy optimisation approach to find mutated genes that have the highest number of interactions with differentially expressed genes [2].

In this thesis' evaluation, DriverNet R package (version 1.28.0) was installed through Bioconductor and used to evaluate somatic mutation, copy number and gene expression data of TCGA and METABRIC breast cancer datasets, as specified in Chapter 2. Mutations data were converted into absolute binary values as DriverNet does not handle directionality and different copy number status. Gene expression data were transformed into patient outlier matrix through DriverNet's provided function (getPatientOutlierMatrix). The gene interaction network used for DriverNet's assessment is retrieved from DawnRank's paper, ensuring a consistent review across both network-based methods. This gene interaction network consisted of 8726 genes post-filtering. The final predicted driver genes are defined as genes with p-values less than 0.05.

3.3.1.2. DawnRank

DawnRank [145] integrates multi-omics data and ranks genes based on its gene interaction network for identifying driver genes. Dawnrank's general principle is similar to DriverNet that the predicted driver genes will most likely cause many downstream genes to be differentially expressed. The main difference, however, DawnRank also classifies driver genes for individual patients in addition to population-level driver genes, which will be useful in the future of clinical personalised medicine or treatment. This network analysis-based algorithm works by integrating somatic alterations in the form of a gene-patient binary matrix, where 1 is for a mutated gene and 0 for non-mutated genes, and differential gene expression data between

cancer and normal samples. The primary method of DawnRank is adopted from the PageRank algorithm to rank genes based on its degree connections on the gene network. To determine the driver genes in a population, DawnRank uses a modified condorcet voting method to predict best gene driver candidates. This voting method makes pairwise comparisons on all possible combination pairs with penalty heuristics introduced to increase the ranking of mutated genes when compared to the non-mutated ones. The gene interaction network matrix provided was built using a variety of resources (MEMo, Reactome, NCI-Nature Curated PID and KEGG).

For this thesis, DawnRank R package (version 1.2) was installed through GitHub (MartinFXP/Dawnrank) and used to evaluate somatic mutation, copy number and gene expression data of TCGA and METABRIC breast cancer datasets, as specified in Chapter 2. Similar to DriverNet's pre-processing step, mutations data were converted into absolute binary values. Gene interaction network matrix applied was obtained from the author supplementary website, which consists of 8726 genes and 155900 edges (post-filtering). As DawnRank does not provide any statistical tests for the population-level driver genes, a DawnRank score threshold of 0.95 was applied to obtain the final driver gene list.

3.3.1.3. OncodriveFML

OncodriveFML [150] is a computational method for identifying cancer driver genes by assessing the pattern of functional impact bias in a cohort of somatic mutations. The rationale behind this approach is based on an established theory that driver genes have high impact functional mutations that often associated with tumorigenesis. OncodriveFML starts by calculating the somatic mutations' average Functional Impact (FI) score across the samples in the datasets. The scoring depends on an independently maintained functional impact prediction framework, such as CADD (Combined Annotation-Dependent Depletion) [156]. Once the average has been computed, an iteration of random sampling on sets of mutations was performed to estimate local expected average FI scores across all target genes. This step is required for computing the empirical p-values by comparing the observed average FI scores against the simulated random sampling FI scores. In the final stage, the p-values are adjusted with the Benjamini-Hochberg procedure to control the false discovery rate.

For this evaluation, OncodriveFML (version 2.3.0) was installed from pip, and the overall analysis was run on Python 3.6.8. We applied this algorithm on both TCGA and METABRIC breast cancer somatic mutations datasets, as specified in Chapter 2. It was run with hg19 genome build and CADD v1.0 for the functional impact scoring, and both were downloaded through the utility (bgdata) provided by the author. The number of sampling was set to 100000, and the signature method was set to *complement*. For this evaluation, a driver gene is defined as having adjusted p-value (after Benjamini-Hochberg correction) of q <= 0.1.

3.3.1.4. OncodriveCLUSTL

OncodriveCLUSTL [144] detects genes with mutations that significantly clustered towards specific protein regions. This approach follows the observations of oncogenic genes that often has hotspot mutations that grouped together, providing the cancer cells a significant advantage in their tumour progression. This unsupervised clustering-based method applies a

smoothing function on the somatic variants provided in the input datasets prior to identifying clusters and scoring them based on the distributions and frequency of their mutations. Also, OncodriveCLUSTL estimates background mutation rate by calculating the mutational profile through random sampling of local mutations for multiple iterations. This analysis is done to assess the significance of the identified clusters, and the resulting p-values are adjusted for multiple hypotheses testing with the Benjamini-Hochberg method.

For this thesis, OncodriveCLUSTL (version 1.1.3) was installed from pip and environment was set to Python 3.6.8. The analysis was done on hg19 genome build downloaded via *bgdata*, default 1000 simulations, a cut-off of 2 for element and cluster mutations, and 11 for smoothing and cluster window. This algorithm was on both TCGA and METABRIC breast cancer datasets with samples as specified in Chapter 2. The element file was to bed file obtained from Ensembl version 95, and all elements are set to the gene symbol. For this evaluation, the driver gene is defined as elements with analytical adjusted p-value (after Benjamini-Hochberg correction) of q <= 0.1.

3.3.1.5. 20/20+

Tokheim et al. [147] proposed a machine learning method for predicting cancer driver genes based on the random forest that extends the original 20/20 rule for classifying oncogenes and tumour suppressor genes. This ratiometric-based method combines many predictive features of mutations that are likely to play essential parts in tumorigenesis, including functional impact, positional clustering and pathogenicity. A total of 24 features are included in a three-class random forest model that is used to vote and score whether a gene is a passenger, oncogene (OG) or tumour suppressor gene (TSG). In the subsequent steps, a driver score is calculated by taking the sum of OG and TSG scores and statistical significance are measured through Monte-Carlo simulations. The resulting p-value is adjusted for multiple testing correction using the Benjamini-Hochberg method.

In this evaluation, a copy of 20/20+ package (version 1.2.3) was obtained from the author's GitHub repository. All the dependencies were installed through Conda on Python 3.6.8 environment, including probabilistic2020 package (version 1.2.3) and snakemake (version 3.13.3). Driver gene predictions on TCGA and METABRIC breast cancer datasets were analysed using provided pre-trained classifiers (2020plus_10K) on 10000 simulations, 10 iterations and 200 trees. The remaining parameters were set to default, using provided data files, gene lists and reference transcripts from SNVBox. For this evaluation, driver genes are defined as genes with adjusted driver p-value (after Benjamini-Hochberg correction) of q <= 0.1.

3.3.2. Evaluation

Selecting the most appropriate algorithms for any prediction-related tasks require thorough neutral evaluations, including assessment of the type of datasets that they can be applied to and limitations of the approaches. In most cases, prediction models are typically benchmarked by comparing their results to validated ground truths. In regards to cancer driver genes, the search for complete findings is still a work-in-progress, with many genes still require further investigation. Computational methods for predicting such genes are predominantly benchmarked against a list of cancer driver genes from Cancer Gene Census

[85]. However, comparison to manually curated genes is not extensive, especially in discovering novel genes, and more comprehensive evaluation strategies have been proposed [147]. For the evaluation protocol presented in this chapter, a combination of assessment approaches from the literature will be adopted, and the overall workflow is summarised in Figure 3.1.

3.3.2.1. Cancer Gene Census

The Cancer Gene Census (CGC) [85] is an ongoing initiative for manually curating a list of established cancer genes, as part of the Catalogue of Somatic Mutations in Cancer (COSMIC) [81]. CGC provides comprehensive detail of various genes' contributions to the cause of cancer and its progression. As of August 2020, CGC has catalogue a total of 723 genes, that are split into Tier 1 (576 genes) and Tier 2 (147 genes). Tier 1 comprises of well-documented cancer genes that have clear evidence in its role in oncogenic activities. Tier 2 is a recently added section that contains genes with less substantial evidence, but have shown strong signs of involvement in cancer-causing events. For this thesis' evaluation, a separate list of breast cancer CGC genes (will be referred to as CGC-breast-genes) were extracted from the full list (will be referred to as CGC-genes). CGC-breast-genes contains a total of 58 genes, where the tumour type for somatic or germline in CGC-gene is categorised as *breast* or *other*.

3.3.2.2. Evaluation framework

The algorithms were evaluated based on six different criteria, as follow:

1. **The number of significant driver genes** from each method was assessed as an early indicator of the algorithms' performance. Whilst a very low number of predictions might result in high precision for other criteria, it could also imply potential overfitting towards established genes used in the prior knowledge database. On the other hand, an increased number of predictions might suggest a better chance of detecting novel drivers as well as a higher rate of false positives.
2. The predicted driver genes were compared to the list of genes previously identified in **Cancer Gene Census** (CGC-gene and CGC-breast-genes) to measure each method's fractions of predictions overlapped with established cancer gene list. Precision, recall and f1-score were also computed.
3. All the predicted driver genes from different approaches were combined to find a **consensus** call (from at least two other methods), and each algorithm was compared against the consensus calls. Precision, recall and f1-score were also computed.
4. **The consistency of the predictions** from each method was assessed on different datasets from the same tumour type (breast cancer). Further consistency evaluation was performed on smaller datasets consisting of randomly selected set of samples from the original data. This random sub-sampling approach was used to investigate the approximate **minimum number of samples** required by each algorithm.
5. While copy number aberration events have been observed to dominate breast-cancer genome, most driver gene prediction methods are focusing on driver mutations, particularly on point mutations. As part of this evaluation, algorithms were also assessed on their ability to pick up additional potential drivers that are not frequently mutated, including **rare and copy number driver** events.

6. As breast cancer has been widely studied as multiple independent diseases, algorithms were also evaluated in their capacity to **predict subtype-specific breast cancer driver genes**.

The following equations are used for calculating precision, recall, and f1- score, following one of the evaluation methods from DawnRank [145]. Gold standard genes used for comparison are one of the CGC-genes, CGC-breast-genes or the computed common drivers of the consensus from multiple algorithms.

$$precision = \frac{\#\ overlapped(gold\ standard\ genes,\ \ predicted\ genes)}{\#\ predicted\ genes}$$

$$recall\ = \frac{\#\ overlapped(gold\ standard\ genes,\ \ predicted\ genes)}{\#\ gold\ standard\ genes}$$

$$f1 - score = \frac{2\ \times\ precision\ \times\ recall}{precision + recall}$$

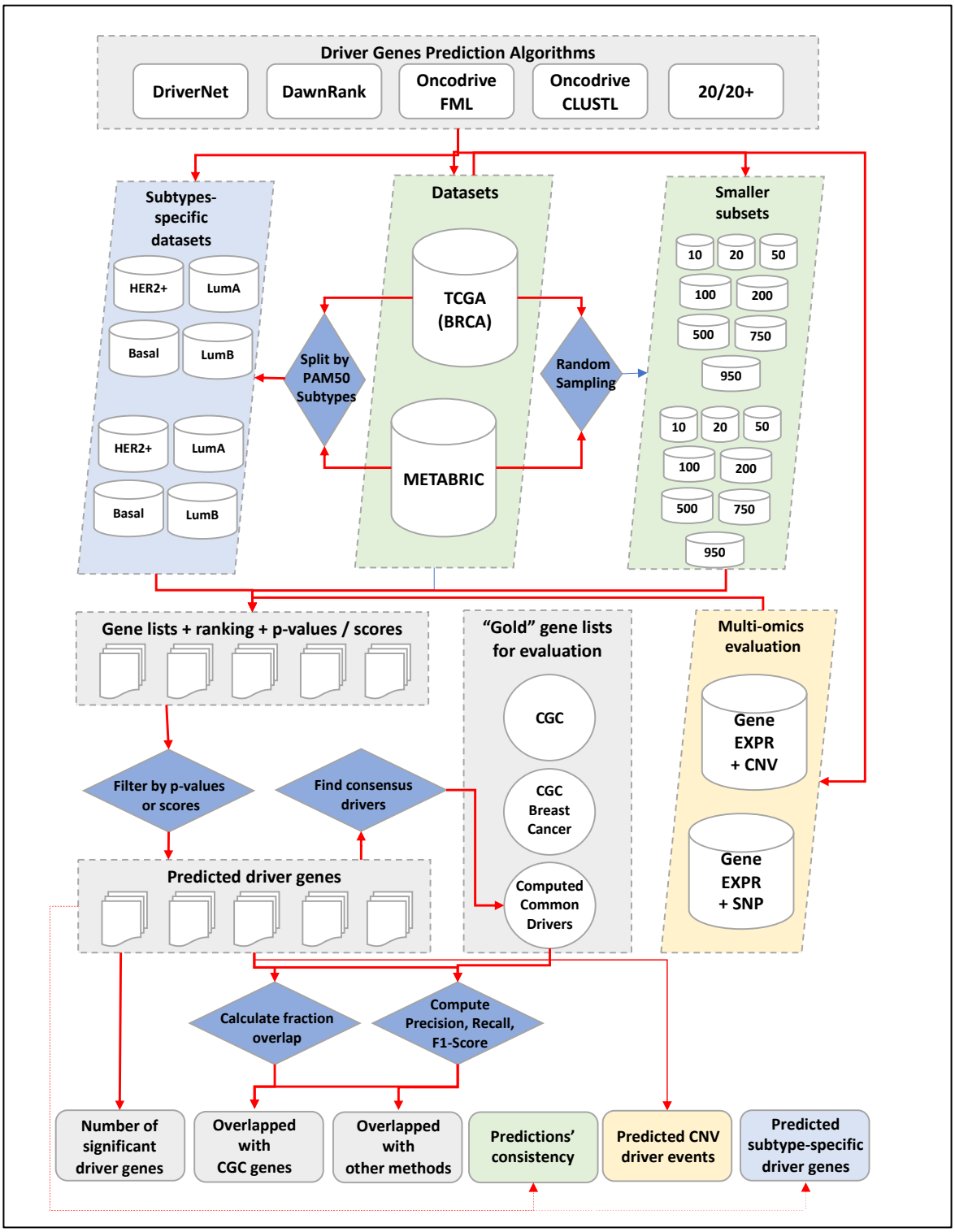


Figure 3.1 - Flowchart of strategies for assessing driver gene predictions methods. Five algorithms (DriverNet, DawnRank, OncodriveFML, OncodriveCLUSTL and 20/20+) are evaluated based on their predictions in TCGA Breast Cancer and METABRIC datasets, including their subtype-specific and randomly sampled smaller datasets. Lists of significant driver genes are selected based on the p-values or driver-scores threshold. These significant drivers are then assessed against the list of critical cancer genes from Cancer Gene Census (CGC), CGC breast cancer genes and common drivers identified by multiple methods. This evaluation also includes assessing the number of predicted driver genes, predictions consistency, predicted copy-number and subtype-specific driver events

3.4. Evaluation Results

This thesis evaluated five different computational methods for predicting driver genes from breast cancer datasets. Two of these algorithms, DriverNet and DawnRank, integrate more than one type of omics data. The remaining three methods, OncodriveFML, OncodriveCLUSTL and 20/20+, focus on the analysis of somatic single point mutations through observing various signals of positive selection. Identical datasets, as well as their subsets, were put through these five algorithms to measure the performance of these methods to identify cancer genes that are crucial for tumorigenesis.

3.4.1. Predicted Driver Genes

All five algorithms identified crucial breast cancer genes that have been previously studied. Tumour suppressor genes *TP53* and *CDH1* were identified by every algorithm on both datasets, while oncogene *PIK3CA* was always among the top-ranked genes for METABRIC (Figure 3.2). These predictions are consistent with multiple independent analysis of breast cancer cohorts, which have confirmed that *TP53* and *PIK3CA* are two of the most of significantly mutated genes [9, 109, 130, 157]. Meanwhile, inactivating mutations on *CDH1* have been frequently observed in invasive lobular breast cancers [130, 158].

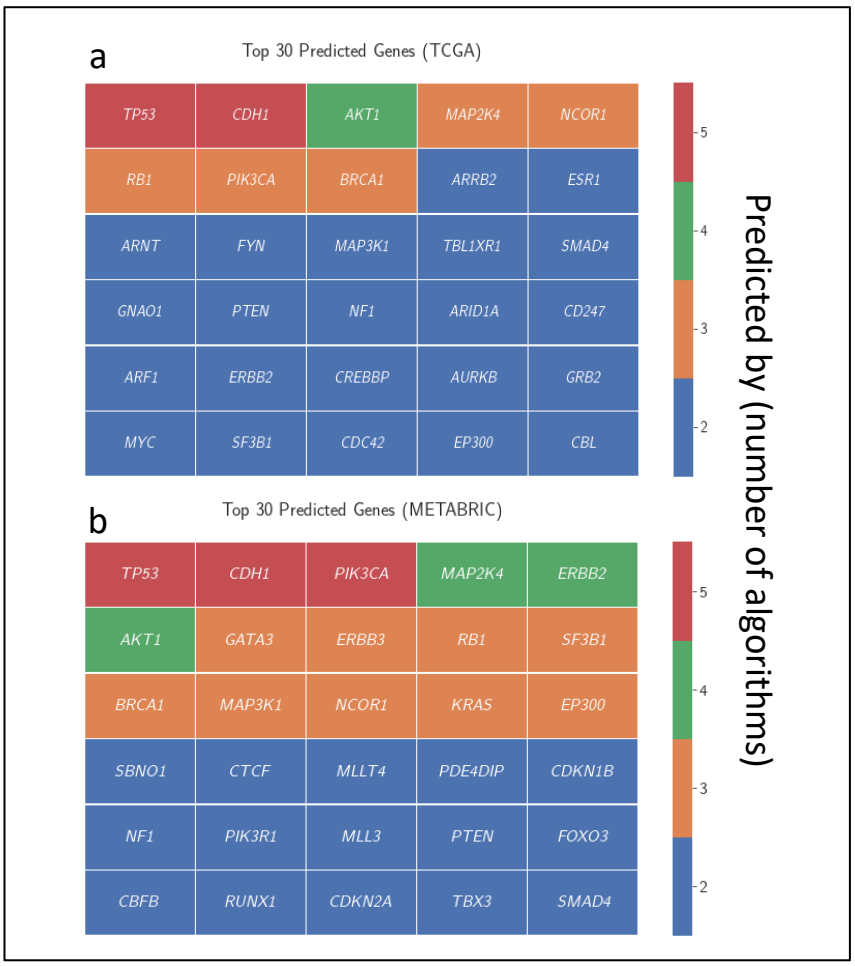
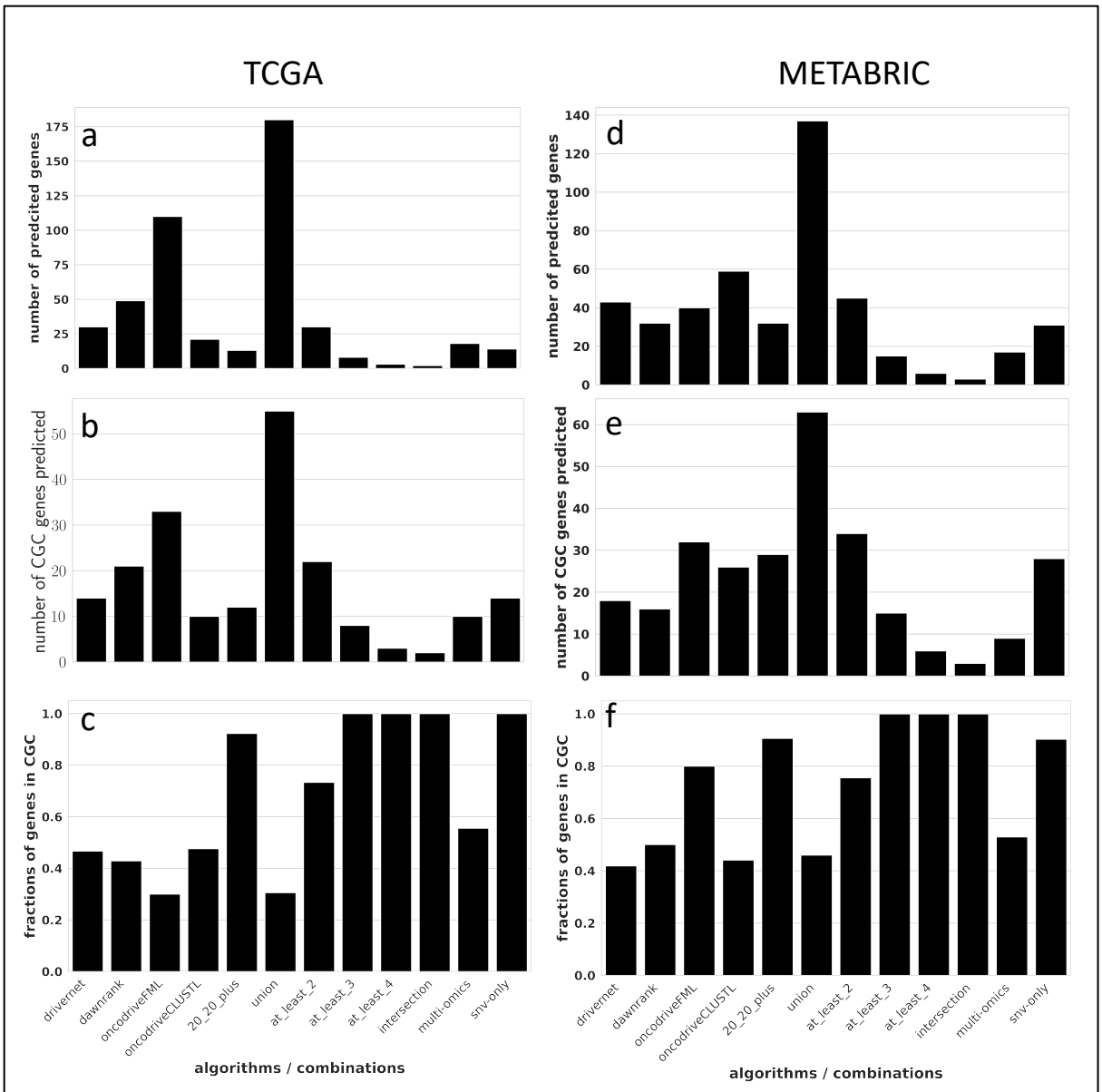


Figure 3.2 - Top 30 genes predicted by more than one algorithm on a) TCGA breast cancer dataset, and b) METABRIC dataset



union	union of genes predicted by all five algorithms
at_least_2	driver genes predicted by at least 2 algorithms
at_least_3	driver genes predicted by at least 3 algorithms
at_least_4	driver genes predicted by at least 4 algorithms
intersection	intersection of driver genes predicted by all five algorithms
multi-omics	intersection of driver genes predicted by DriverNet and DawnRank
snv-only	intersection of driver genes predicted by OncodriveFML, OncodriveCLUSTL and 20/20+

Figure 3.3 - Summary of predicted driver gene comparisons across different evaluated algorithms and their combinations. Plots on the left (a-c) are from TCGA breast cancer dataset and plots on the right (d-f) are from METABRIC dataset. Top plots (a, d) indicates the raw count of significant driver genes. Middle plots (b, e) displays the number of identified genes that overlapped with the Cancer Gene Census (CGC). Bottom plots (c, f) shows the fraction of predicted driver genes in CGC.

A total of 251 driver genes were predicted to be significant (criteria as described in the method section) across both datasets. Network-based algorithms, which integrated multi-omics data, predicted 30 to 49 driver genes, while the remaining methods that only analysed somatic mutations data have predictions ranging from 13 to 110 genes (Figure 3.3). The two Oncodrive packages tend to classify more genes, while 20/20+ has consistently identified the least number of genes (full list of predicted genes are listed in Appendix). From 180 driver genes predicted on TCGA breast cancer dataset, 83% (n = 150) are unique to individual algorithms, while only 30 genes are shared by more than one algorithm (Figure 3.3). A similar pattern can also be observed on METABRIC dataset, where only 45 out of 137 identified genes are shared among the algorithms (Figure 3.3). These numbers could potentially suggest that these different computation algorithms are complementary in search of novel driver genes, considering their methods' definition is targeting different principles of cancer drivers. However, it is also worth noting that a high percentage of unique genes may also lead to higher false positive predictions.

Apart from successfully identifying known driver genes, all algorithms also exclude potential false-positive findings from their final predictions. These include a list of long genes, such as *GPR98, HMCN1, OBSCN, RYR2, RYR3, TTN*, that have been observed to have a high frequency of mutations on tumour samples [141, 144]. None of these six genes was identified as drivers by any of the algorithms across both datasets (Appendix) despite being on the list of top 50 frequently mutated genes (Table 3.2).

Table 3.2 - Top 50 frequently mutated genes from the datasets. Each gene frequency was counted per sample from the binary somatic mutation matrix.

TCGA - Breast Cancer

GeneSymbol	Frequency	GeneSymbol	Frequency
PIK3CA	347	*NCOR1*	52
TP53	345	*MUC5B*	51
TTN	225	*CSMD3*	51
MUC16	138	*DNAH11*	50
CDH1	130	*RYR1*	50
GATA3	127	*LRP1B*	49
MAP3K1	92	*FAT3*	48
RYR2	88	*CACNA1E*	48
FLG	80	*RUNX1*	47
SYNE1	77	*PKHD1L1*	47
HMCN1	74	*HYDIN*	46
USH2A	72	*APOB*	46
SPTA1	68	*HUWE1*	45
RYR3	68	*CSMD2*	45
OBSCN	66	*XIRP2*	45
ZFHX4	62	*NF1*	45
CSMD1	60	*DNAH17*	45
DST	59	*GPR98*	44
DMD	59	*LRP2*	44
PTEN	58	*DNAH9*	43
NEB	58	*PCLO*	43
SYNE2	57	*MACF1*	43
ABCA13	56	*MAP2K4*	42
MUC4	54	*ARID1A*	42
MUC17	54	*VPS13B*	41

METABRIC

GeneSymbol	Frequency	GeneSymbol	Frequency
PIK3CA	800	*COL22A1*	110
TP53	664	*STAB2*	109
AHNAK2	420	*COL12A1*	105
MUC16	415	*THADA*	100
SYNE1	316	*MYH9*	97
GATA3	246	*ATR*	95
MAP3K1	223	*CBFB*	94
AHNAK	219	*SHANK2*	94
DNAH2	213	*AKT1*	92
HERC2	199	*ARID1B*	91
USH2A	185	*NCOR1*	89
RYR2	184	*EP300*	89
CDH1	181	*PTPRD*	88
NOTCH1	170	*NF1*	85
DNAH5	169	*LAMB3*	84
PDE4DIP	152	*SETD2*	81
TG	150	*PTEN*	80
COL6A3	148	*ERBB2*	78
BIRC6	147	*RUNX1*	78
NCOR2	143	*SETD1A*	76
AKAP9	140	*SF3B1*	75
UTRN	134	*ROS1*	73
ARID1A	128	*THSD7A*	73
LAMA2	126	*ALK*	73
TBX3	112	*UBR5*	72

3.4.2. Overlapped with CGC Genes

Predicted driver genes from every evaluated algorithm were assessed against the gene list from CGC. All algorithms predicted at least 10 genes that overlapped with CGC-genes. From the evaluation on TCGA breast cancer dataset, OncodriveFML has the highest number of predicted genes in CGC list (n = 33), followed by DawnRank (n=21), DriverNet (n=14), 20/20+ (n=12) and OncodriveCLUSTL (n=10). This result could be attributed to the greater number of identified genes in OncodriveFML. When fractions are used for ranking the methods, 20/20+ has considerably better performance than the rest. On the other hand, OncodriveFML has only 30% of its predicted genes in CGC, while the remaining methods are at 40% or more (Figure 3.3). As the number of predicted genes vary between algorithms, their performance was further assessed based on their precision, recall, F1-score against CGC-genes as the gold standard. The result is summarised on Table 3.3, which shows a reasonably balanced f1-score across all algorithms, with OncodriveFML, DawnRank and DriverNet being the top 3 methods.

This same comparison was also re-assessed against CGC-breast-genes. The ranking remains identical in the number of overlapped genes (OncodriveFML, DawnRank, DriverNet, 20/20+, OncodriveCLUSTL). The f1-score is also following similar balanced pattern, except for 20/20+ that is now marginally better due to its higher precision [Table 3.3].

This evaluation was then reanalysed with the predicted genes from METABRIC dataset to gauge algorithms' predictions consistency. Fractions and the number of predicted genes in CGC are similar for multi-omics methods while the other three mutation-data-only methods performance show slight variations. OncodriveFML still has the highest number of overlapped genes with CGC, but the number of predicted genes is only halved to that of TCGA data (Figure 3.3). On the other hand, 20/20+ has more than doubled the number of genes classified as drivers (n = 32) despite still having the best f1-score among all the evaluated algorithms (Table 3.3). It is important to note that METABRIC dataset has approximately twice the number of samples that TCGA breast cancer dataset has, and only sequenced 173 frequently mutated breast cancer genes. These differences will be explored further in the consistency evaluation result section.

Genes identified by more than one algorithm have a higher percentage of predicted genes overlapping with CGC-genes. Driver genes classified by at least two different algorithms have at least 75% of its predicted genes in CGC, while every single gene indicated by three or more algorithms are always one of the CGC genes. This result is consistent across both TCGA and METABRIC datasets (Figure 3.3).

3.4.3. Consensus between algorithms

Motivated by the high number of overlaps between CGC-genes and majority voting method, gene list generated from this approach was then used as the next evaluation's gold standard. The majority voting here is defined as genes that are identified by at least two other algorithms, excluding the evaluated method. On TCGA data, DawnRank has the highest recall among all the evaluated algorithms with the highest number of overlaps (n=8), while 20/20+ has the best precision and f1-score. On METABRIC dataset, all the mutation-data-only

methods have higher number of overlaps, with 20/20+ has significantly better performance on all criteria (Table 3.3).

Table 3.3 - The number of predicted genes (n), precision, recall, and f1-score of evaluated algorithms across different comparisons.

TCGA		DriverNet	DawnRank	Oncodrive -FML	Oncodrive -CLUSTL	20/20+
CGC	n	14	21	**33**	10	12
	precision	0.47	0.43	0.3	0.48	**0.92**
	recall	0.02	0.03	**0.05**	0.01	0.02
	f1-score	0.04	0.05	**0.08**	0.03	0.03
CGC (breast)	n	8	11	**15**	6	8
	precision	0.27	0.22	0.14	0.29	**0.62**
	recall	0.14	0.19	**0.26**	0.1	0.14
	f1-score	0.18	0.21	0.18	0.15	**0.23**
Predicted by >=2 other algoithms	n	4	**8**	7	4	6
	precision	0.13	0.16	0.06	0.19	**0.46**
	recall	0.25	**0.53**	0.32	0.14	0.24
	f1-score	0.17	0.25	0.11	0.16	**0.32**

METABRIC		DriverNet	DawnRank	Oncodrive -FML	Oncodrive -CLUSTL	20/20+
CGC	n	18	16	**32**	26	29
	precision	0.42	0.5	0.8	0.44	**0.91**
	recall	0.02	0.02	**0.04**	**0.04**	**0.04**
	f1-score	0.04	0.04	**0.08**	0.07	**0.08**
CGC (breast)	n	9	10	**18**	12	17
	precision	0.21	0.31	0.45	0.2	**0.53**
	recall	0.16	0.17	**0.31**	0.21	0.29
	f1-score	0.18	0.22	0.37	0.21	**0.38**
Predicted by >=2 other algoithms	n	6	9	13	12	**14**
	precision	0.14	0.28	0.33	0.2	**0.44**
	recall	0.18	0.27	0.45	0.3	**0.47**
	f1-score	0.16	0.28	0.38	0.24	**0.45**

Bolded texts represent the highest values in the category

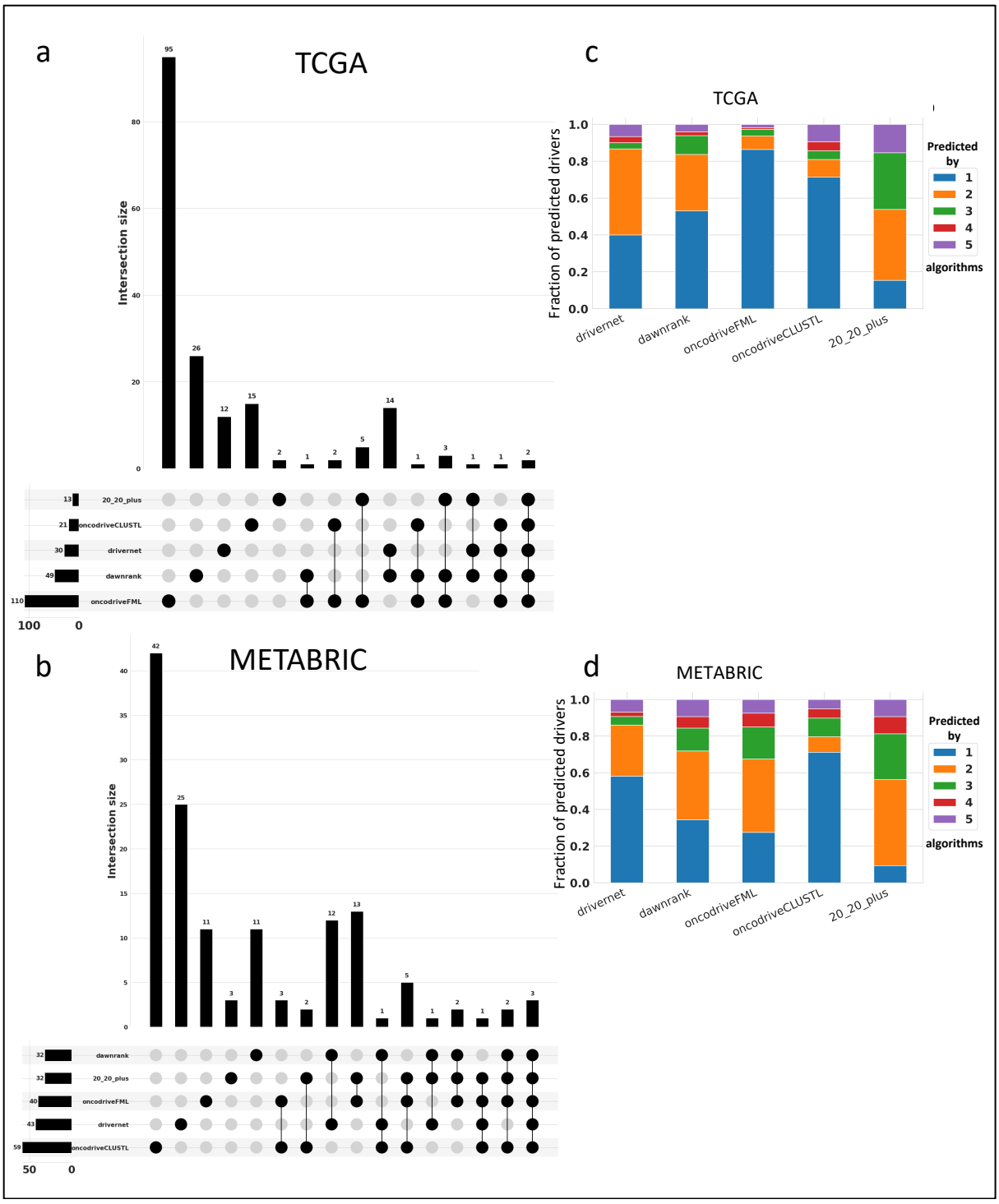


Figure 3.4 - Agreement of predictions between different algorithms. (a-b) The UpSet plots summarise the number of shared predictions between algorithms. (c-d) The fraction of predicted genes in each algorithm appearing in the predicted lists from other methods.

Multiples studies have suggested that computationally predicted genes are more likely to be true positive if they are identified by multiple algorithms [147, 159]. However, the result of this evaluation indicates that the majority of the predictions are unique to particular methods. To further assess how each method agrees with its peers, fractions of drivers were also calculated on individual processes to measure how many identified genes from each algorithm appear on another method's predictions. Figure 3.4 summarises the result of this assessment. Across both datasets, 20/20+ has the largest proportions of predicted genes that agree with the other algorithms, where it shared the most significant number of predicted genes with OncodriveFML. This result is not unexpected, considering 20/20+ is a random forest method consisting of different features for capturing multiple signals of positive selections. These features are likely to cover similar domains as OncodriveFML's functional impact bias method and OncodriveCLUSTL' clustering-based method. Similarly, DriverNet and DawnRank share many of the same predicted driver genes, and both algorithms are integrating multi-omics data through the same gene interaction network matrix. More than half of DriverNet's predicted driver genes were also identified by DawnRank on TCGA dataset, while only sharing three genes with 20/20+. In general, less than 10% of the predicted genes from each algorithm were also identified by all the other methods. These genes are *TP53*, *CDH1* and *PIK3CA*, which are well-studied driver genes, as described in the earlier section.

3.4.4. Prediction consistency

To measure the consistency of each method in predicting driver genes, all algorithms were run through two different datasets and multiple smaller subsets derived from them. As described in chapter 2, these two independent datasets have different number of samples and genes, and their gene expression data were generated from different sequencing technology. For the second consistency assessment, smaller datasets were generated by randomly sampling 10, 20, 50, 100, 200, 500, 750 and 950 samples across both TCGA and METABRIC datasets. The result of this consistency evaluation is summarised in Figure 3.5.

The first approach of using independent dataset assessed the consistency of each algorithm's predictions when applied to different shape of data. DriverNet and DawnRank are the most consistent among all the algorithms, predicting a relatively similar number of driver genes, with more than half of the identified drivers appearing across the results on both datasets. OncodriveCLUSTL and 20/20+ predicted more driver genes when run on METABRIC data, doubling the number of predictions when compared to the TCGA equivalent analysis. Their predictions are also mostly inconsistent, with more than two-third of the significant driver genes unique to individual datasets. Likewise, OncodriveFML's predictions are also fundamentally different between datasets, where the number of predictions on TCGA is almost three times more than the predicted genes on METABRIC. This comparison suggests that the number of samples and genes included in the dataset affected network-based algorithms less than methods that rely solely on somatic mutations data.

For the second consistency evaluation, datasets with a smaller number of samples were randomly sampled to assess the performance of algorithms in dealing with smaller cohorts. F1-score was also calculated based on the comparison with prediction from the complete datasets. Overall, DriverNet and DawnRank were notably better than OncodriveFML,

OncodriveCLUSTL and 20/20+ in dealing with different datasets size. These network-based algorithms' predictions stay reasonably consistent and have higher f1-score across different size of datasets. On the other hand, OncodriveFML, OncodriveCLUSTL and 20/20+ reported no significant driver genes when run on a small cohort. The number of driver genes predicted by these mutation-frequency-based methods improved as the number of samples increased. Nonetheless, all methods are observed to have lower performance as the number of samples dropped below 50. This consistency patterns could explain the prediction differences between the datasets, as METABRIC has considerably more samples than TCGA.

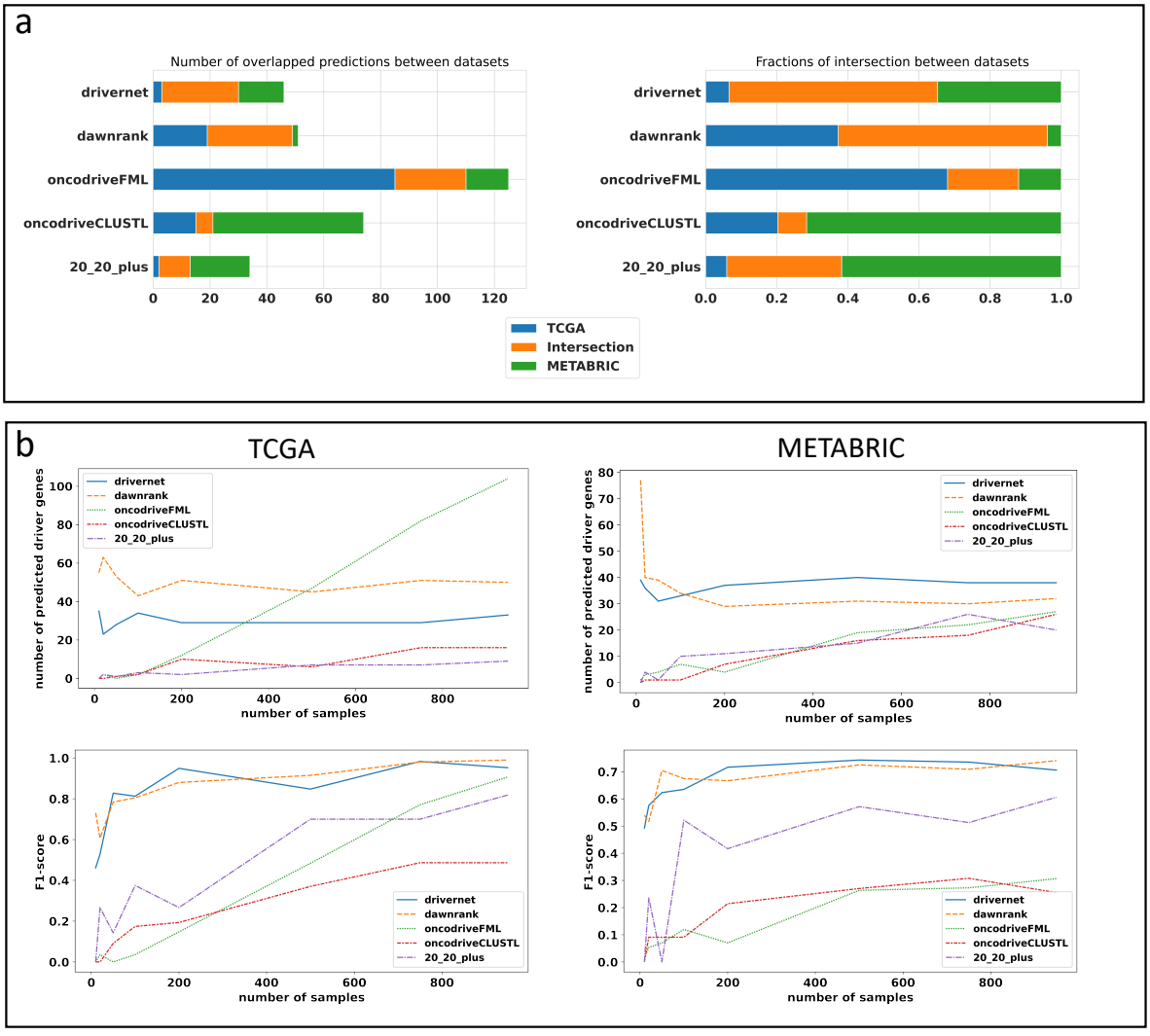


Figure 3.5 - Consistency assessment summary of all evaluated algorithms. Top plots (a) display the number of overlapped driver genes (and the fraction) when methods are used to analyse different datasets. Bottom plots (b) show the pattern of predictions as the number of samples in the datasets increase

3.4.5. Identification of copy number drivers

Breast cancer has been studied to be dominated by genomic copy number aberrations [9]. As OncodriveFML, OncodriveCLUSTL and 20/20+ are only analysing the impact of somatic mutations for identifying driver genes; they could potentially miss genes that were driven by copy-number or epigenetics changes. Network-based algorithms, such as DriverNet and DawnRank, were in theory supposed to address this limitation. However, both algorithms require to have the somatic single point mutations (SNP) data to be combined with the copy number variations (CNV) data in a single binary matrix, rendering it challenging to differentiate the type of mutations that determine the predictions.

To assess the added value of integrating multiple types of omics data, DriverNet and DawnRank were also evaluated using the unmerged genomic aberrations data. By combining CNV information, DriverNet identified additional 11 driver genes that would not be otherwise picked up by SNP and gene expression integration alone. Similarly, DawnRank classified a further 39 genes as potential drivers when copy number data is integrated into the overall analysis. Collectively, these genes include *AURKA*, *AURKB*, *CDC42*, *CDK7*, *MYC* and *SMAD4*. The complete list of these additional genes is listed in Table 3.4, with a full list of predicted genes available in the Appendix.

One of these essential CNV-related genes predicted by DriverNet and DawnRank is *MYC*, which was not in the driver gene list considered as significant by the other three algorithms. Samples with *MYC* copy number events in METABRIC datasets have a worse 5-year disease-free survival ($p < 0.05$) [Figure 3.6]. Literature has described the role of *MYC* oncogene in the development of cancer, where *MYC* amplification is associated with aggressive breast cancer subtypes and poor prognosis [84, 116, 160]. Overall, this result further highlights the importance of integrating copy number variants in predicting driver genes.

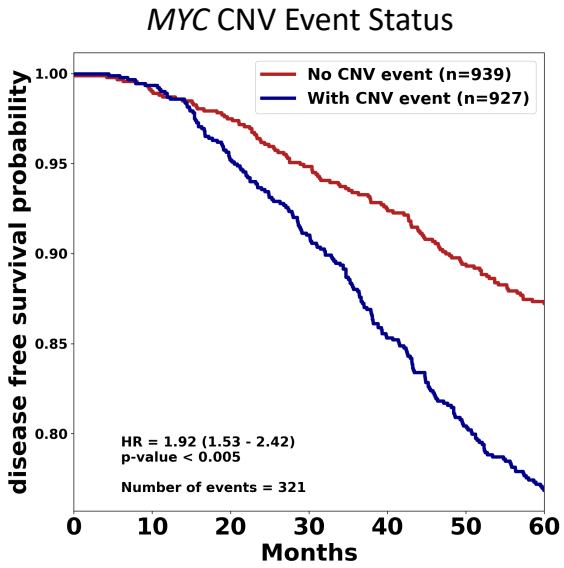


Figure 3.6 - Kaplan-Meier curves according to MYC copy-number variation (CNV) status. Samples with no MYC CNV event have better 5-year disease-free survival (HR = 1.92; CI = 1.53 – 2.42; p < 0.005). Cox proportional hazard ratio analysis was performed using the Python package lifelines (https://doi.org/10.5281/zenodo.3267531).

Table 3.4 - Additional copy-number-drivers predicted by DriverNet and DawnRank

DriverNet	DawnRank		
ARF1	*ACTA1*	*DLG4*	*LYN*
ARRB2	*ARF1*	*DVL2*	*MAPK1*
AURKB	*ARNT*	*E2F1*	*MAPK3*
CDC42	*ARRB2*	*E2F4*	*MAX*
CDK7	*AURKA*	*ESR1*	*MYC*
FYN	*AURKB*	*ETS1*	*NCOA2*
GNAO1	*AXIN1*	*FOS*	*NFATC2*
GRB2	*BCAR1*	*FYN*	*PAFAH1B1*
HSP90AA1	*CBL*	*GNAL*	*PLK1*
MYC	*CD247*	*GNAO1*	*POU2F1*
RPL13	*CDC42*	*GNB1*	*SHC1*
	CEBPB	*GRB2*	*SMAD4*
	CRK	*HDAC2*	*SRC*

3.4.6. Evaluation on subtype-specific datasets

Breast cancer is a heterogeneous disease that consists of multiple subtypes with different molecular characteristics. The four widely studied breast cancer molecular subtypes are basal-like, HER2-enriched, luminal A and luminal B, and they have been described to exhibit different mutation profiles [9, 109, 130, 134]. In this evaluation, we run all algorithms against a subtype-specific subset of the original datasets and assess the performance of each method in predicting potential subtype-specific driver genes. The number of predicted driver genes per subtypes and their top 40 driver genes are summarised in figure 3.7.

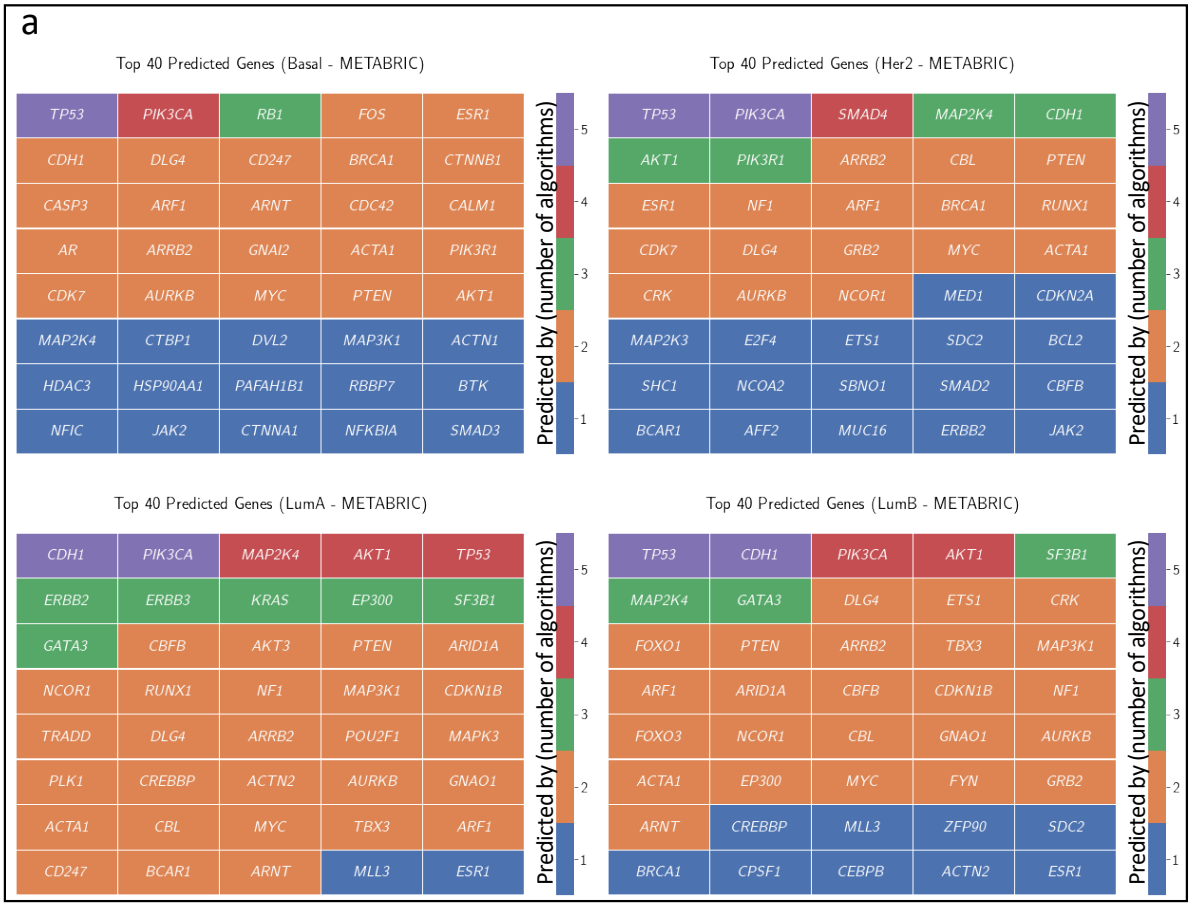
As expected, the top predicted genes on subtype-specific datasets differ slightly from the full datasets. For this criteria, top predicted genes are ranked by the number of algorithms that nominated the particular gene as a driver. The genes that were identified by all five algorithms in the full METABRIC datasets, *TP53*, *CDH1* and *PIK3CA*, were not the most significant for every single subtype. *TP53* ranked highest on all subtypes except for luminal A, while *PIK3CA* was the top for HER2-enriched and luminal A samples. *CDH1* was more commonly observed as a top driver on ER+ subtypes. Whilst *GATA3* was predicted as a main driver only in luminal A and luminal B samples, *RB1* was assessed as the top drivers only in basal-like datasets.

The number of subtype-specific driver genes predicted by each algorithm follows similar consistency patterns in the previous section. DriverNet's number of predicted drivers is mostly consistent across all subtypes (n = 26 – 34). DawnRank predicted considerably fewer drivers for luminal A subtypes (n = 36) compared to Basal-like, HER2-enriched, or Luminal B (n = 82, 61, 75 respectively). Except for luminal A, the remaining three algorithms classify a relatively lower number of driver genes across all subtypes. For example, OncodriveCLUSTL and 20/20+ did not identify any driver genes for HER2-enriched TCGA dataset. This could be attributed to the availability of samples per subtype, where TCGA has a relatively smaller number of HER2-enriched samples, and this is one of the main limitations of mutation-frequency-based methods.

Some of these algorithms also predicted additional driver genes that were not previously identified in the full datasets. DriverNet reported *AKT3* and *MAPK3* as potential drivers for

luminal A samples, as well as *PIK3R1* and *CASP3* for basal-like samples. 20/20+ identified *ATR* as a likely driver for HER2-enriched subtype and *USP28* for luminal B samples. On the other hand, DawnRank proposed *CCND1* as a potential driver gene for luminal B, as well as multiple immune-related genes as possible drivers for basal-like subtype, including *JAK2*, *CD4*, *CDK7* and *CASP3*. Overall, DawnRank nominated the highest number of additional subtype-specific drivers.

a



b

TCGA	DriverNet	DawnRank	OncodriveFML	OncodriveCLUSTL	20/20+
Basal	26 (6)	82 (45)	6 (0)	8 (6)	5 (1)
Her2	27 (10)	61 (21)	3 (0)	0 (0)	0 (0)
LumA	34 (10)	36 (5)	36 (7)	11 (5)	9 (1)
LumB	26 (5)	72 (25)	7 (1)	2 (1)	5 (0)

METABRIC	DriverNet	DawnRank	OncodriveFML	OncodriveCLUSTL	20/20+
Basal	32 (6)	58 (35)	4 (0)	3 (0)	6 (0)
Her2	30 (1)	44 (17)	11 (0)	5 (0)	17 (1)
LumA	32 (6)	36 (9)	25 (0)	17 (0)	21 (0)
LumB	35 (3)	41 (13)	19 (0)	11 (1)	19 (1)

Figure 3.7 - a) Top 40 driver genes predicted per-subtype, by more than one algorithm. b) Number of driver genes identified by each algorithms per-subtype. Numbers in bracket represent the amount additional genes identified that were not in the initial full datasets' predictions.

3.4.7. Overall performance

Overall, all five evaluated methods identified well-studied driver genes from a cohort of breast cancer genomics data. Despite differences in the algorithms' principles, each approach brings diverse advantages to the field. It is also important to note various limitations that come with each method, especially around sample sizes. The summary of this chapter's evaluations is presented in Table 3.5. In conclusion, the evaluation results suggest that there is no single *best* algorithm for all use-cases, and the most appropriate method selection depends on datasets characteristics and the objectives of a study.

Table 3.5 - The evaluation summary of driver gene prediction algorithms

	DriverNet	DawnRank	OncodriveFML	OncodriceCLUSTL	20/20+
Number of significant driver genes	**Small** to **medium** number of predicted driver genes	**Small** to **medium** number of predicted driver genes Provide no statistical test. Significant genes were calculated based on the internal scoring system	The number of predicted genes **varies** Predicted significantly more drivers on mutation data from whole exome sequencing (TCGA)	The number of predicted genes **varies** Predicted the **most** number of drivers on a larger dataset (METABRIC), but **lowest** on a smaller dataset (TCGA)	The **smallest** number of predictions among all algorithms
Overlapped with CGC	**Low** f1-score across both datasets	**Medium** f1-score across both datasets	The **highest** number of overlapped with CGC genes **Low** f1-score due to higher number of significant driver genes	**Worst** f1-score on TCGA; Medium f1-score on METABRIC	**Best** f1-score across both datasets
Overlapped with other methods	**Low** f1-score across both datasets	**Medium** f1-score across both datasets	**Worst** f1-score on TCGA due to significantly more genes predicted when compared to other methods	**Low** f1-score across both datasets	**Best** f1-score across both datasets
Predictions' consistency	**Best** consistency High fraction of overlap between predicted genes from both datasets Similar number of predictions across both datasets.	**Good** consistency High fraction of overlap between predicted genes from both datasets Predicted fewer driver genes on METABRIC due to fewer genes being sequenced for mutation	**Poor** consistency of predictions Predicted a significantly higher number of potential drivers on WES mode than the targetted mode	**Worst** consistency among the evaluated algorithms Minimal overlap between genes predicted on TCGA and METABRIC datasets	**Poor** consistency of predictions Predicted twice the number of drivers when run on a larger dataset
Number of samples required	**Worked on a smaller dataset**, but best with n > 50	**Worked on a smaller dataset**, but best with n > 50 Provides single sample predictions	Predicted **no / very few driver genes** when the number of samples is less than 100. Number of predictions increases as the number of samples increases	Predicted **no / very few driver genes** when the number of samples is less than 100. Number of predictions increases as the number of samples increases	Predicted **no / very few driver genes** when the number of samples is less than 100. Number of predictions increases as the number of samples increases
Predicting driver CNV events	**Predicted** a few potential CNV driver genes	**Predicted** a few potential CNV driver genes	**Not supported**	**Not supported**	**Not supported**
Predicting subtype-specific driver genes	**Predicted** a few potential subtype-specific driver genes **Provided** driver genes predictions on all subtypes **consistently** **Predicted some additional genes** that were not previously identified in the full datasets	**Predicted** a few potential subtype-specific driver genes **Provided** driver genes predictions on all subtypes **consistently** **Predicted many additional genes** that were not previously identified in the full datasets	**Predicted** a few potential subtype-specific driver genes **Could not** be applied reliably on all subtypes due to the lower number of samples on specific subtypes **Predicted very few additional genes** that were not previously identified in the full datasets (except TCGA LumA)	**Predicted** a few potential subtype-specific driver genes **Could not** be applied reliably on all subtypes due to the lower number of samples on specific subtypes **Predicted very few additional genes** that were not previously identified in the full datasets (except TCGA LumA & Basal-like)	**Predicted** a few potential subtype-specific driver genes **Could not** be applied reliably on all subtypes due to the lower number of samples on specific subtypes **Predicted very few additional genes** that were not previously identified in the full datasets

3.5. Discussion

Cancer is a complex genetic disease, developed from various somatic mutations that alter the cellular functions. Cancer cells accumulate a large number of genomic aberrations, but the majority of these alterations are not directly involved in tumorigenesis. Identifying the main drivers of cancer initiation and progression has been a significant objective of cancer genomics studies. This goal has been further supported by the advancement of sequencing technology over the past decades that have driven the cost of sequencing down, enabling many extensive cancer studies to be conducted. The vast amount of comprehensive genomics data generated is crucial in assisting the search of cancer driver genes.

Many computational methods have been developed to identify critical cancer genes. These algorithms are based on various commonly observed characteristics of driver genes such as 1) mutations that occurred in positions with high functional importance that affect the protein function 2) frequently mutated across many samples, and 3) mutations that affect the expressions of many different genes. Although many of these methods have successfully identified some cancer-associated genes, there are little overlaps in the predicted genes across other methods. Measuring the performance of these algorithms is also challenging in the absence of ground truth. In this chapter, an evaluation strategy consisted of six criteria was proposed to assess five driver gene predictions tools. These criteria include: 1) the number of predicted driver genes, 2) comparison to established cancer genes from CGC, 3) consensus between methods, 4) prediction consistency across different datasets and number of samples. 5) identification of rare and copy number events, and 6) identification of subtype-specific driver genes. Overall, each method demonstrates its strengths and limitations, and the choice of the best algorithm depends on the main objective of the study and the characteristics of the datasets.

Computational methods can identify novel and established driver genes

Genes identified in CGC is the *de facto* standard for evaluating newly proposed driver gene methods, used exclusively by almost every proposed implementation. CGC catalogues a comprehensive list of cancer genes, including the type of mutations and tumour type commonly observed, from a vast range of sources. This provides a good platform for proposed methods to demonstrate their abilities in identifying established cancer genes, crucial for quantifying their algorithms' sensitivity. However, for a study that aims to discover novel driver genes, algorithms with very high fractions of predictions in CGC do not serve many purposes. An excessive number of novel predictions could also be attributed to higher false positives, but it is also important to consider that CGC is an ongoing effort that continuously introduce many new genes every year.

An alternative gold standard for assessing the sensitivity of driver gene prediction is through comparisons with complementary methods. The principle behind this approach states that genes identified by multiple algorithms that detect different signals of positive selection are more likely to be driver genes [159]. The result from this thesis' evaluation matches this theory, where predictions identified by at least two methods have 75% overlapped with CGC genes. Adding more algorithms into the consensus method increases the fractions of overlapped, with genes identified by three or more algorithms are all well-studied genes.

These established genes include *TP53, CDH1, PIK3CA, AKT1, MAP2K4, ERBB2, NCOR1, RB1, BRCA1, GATA3, ERBB3, SF3B1, MAP3K1, KRAS* and *EP300*.

20/20+ outperformed other methods evaluated in this thesis with its precision in identifying established genes. It consistently has the highest fractions of predicted genes in the different proposed *ground truth*, including CGC, CGC-breast and computed common drivers across multiple methods. Conversely, 20/20+ is also the only algorithm that has less than ten unique drivers, with only three genes that were not predicted by other methods. The possible reason for this observation can be attributed to the design of the 20/20+ algorithm. 20/20+ is a machine-learning-based method that builds a random forest with multiple features that target different signals of positive selection. Some of these features would have overlapped with the principle of other methods. For example, OncodriveFML evaluates the functional impact of mutations in the cohort for predicting driver genes, and these functional impacts are some of the features in 20/20+ model. This will also explain the high number of overlapped predictions between OncodriveFML and 20/20+, where the overlapped fractions are considerably higher than the two network-based methods.

Tumorigenesis is not only driven by the presence of somatic point mutations. Multiple studies have presented evidence that copy number or epigenetics changes could directly affect the expression of the genes [69, 137, 161]. Genes such as *MYC* (gain) and *MAP2K4* (loss) do not see as many point mutations as some other driver genes but are more frequently amplified or deleted [67]. Unfortunately, some of these drivers are often not picked by algorithms that do not support the evaluation of more than one type of omics data. In this chapter evaluation, it is evident that network-based algorithms (DriverNet and DawnRank) could predict some rarely point-mutated breast cancer drivers better than the remaining algorithms. For instance, DriverNet identified *AURKB*, *EGFR* and *MYC* as potential drivers and these genes were missed by 20/20+, OncodriveFML, and OncodriveCLUSTL. Overexpression of *AURKB* is correlated with poor breast cancer survival [162], while *MYC*-amplification and high expression of *EGFR* are observed on a more aggressive breast cancer subtype [116, 117]. This additional advantage indicates that integrating multi-omics data, such as transcriptional profiling, can further improve the search of novel driver genes.

Limitations of current approaches in predicting driver genes

Despite the efficacy of the evaluated algorithms in identifying critical cancer genes, there are some ongoing limitations with these methods. Computational approaches have a strong dependency on the completeness of the prior knowledge that they employed. Network-based methods, DriverNet and DawnRank, require a gene interaction network (GIN) for assessing the importance of the predicted genes. This GIN is often derived from other databases, such as KEGG and REACTOME, which are still incomplete. In its paper, DriverNet noted that a well-studied gene *ZNF703* was missing from their GIN and thus will never appear as a significant driver gene on its prediction. Likewise, OncodriveFML relies on the functional impact score from other databases, such as CADD, which is an independently maintained machine-learning-based method for variant prioritisation. This limitation is also observed on 20/20+ where a set of predefined oncogenes and tumour suppressor genes are required for training a new random forest model. However, from the results of this thesis' evaluation, it could also be hypothesised that the consensus-based approach of integrating the predictions from

multiple methods can partly address this limitation, especially in predicting established driver genes.

Aside from prior knowledge, algorithms are also dependent on third-party computations for generating the input *'omics* data to their methods. Genomic sequencing data are often analysed through bioinformatics pipelines consisting of multiple tools that are still being actively developed. Filtering methods and thresholds used to select true somatic mutations can also potentially affect the number of mutations in the datasets, and thus affect the final driver gene predictions. In the case of DriverNet and DawnRank, additional considerations are necessary for interpreting driver copy number events. Amplifications and deletions can be detected to span across multiple genes, depending on the copy number detection methods. The use of binary matrix for representing somatic mutations will also mean that extreme copy number events were treated similarly to any other mutations. Both algorithms are further limited by its implementation that did not consider the directionality of gene expression profiles, as well as dependency towards the normalisation methods applied to the raw gene counts.

Driver gene predictions from these computational methods are also found to be largely inconsistent. Evaluating them through different datasets yield different sets of driver genes despite both were from the same tumour type. Whilst OncodriveCLUSTL performed poorly in this evaluation criterion, network-based methods demonstrated better stability than their counterparts with similar number of predictions and plenty of overlaps across both datasets. The different number of genes and samples between TCGA and METABRIC datasets could be the main reasons that explain these differences. METABRIC only sequenced 173 frequently mutated breast cancer genes compared to the whole-exome sequencing data from TCGA. Algorithms that rely solely upon point mutation data would have significantly fewer data points to work with, in comparison to DriverNet and DawnRank that include copy number and gene expression data in their equations. The evaluation results on randomly sampled datasets also suggest that OncodriveFML, OncodriveCLUSTL and 20/20+ require a large samples cohort to predict driver genes. This observation is another possible reason for the discrepancies between predictions when run on the larger METABRIC dataset.

The utility of these methods in the current diagnostics settings has remained restricted. Although the identification of new novel driver genes could lead to the research of new treatment and drugs, its application for personalised care remains challenging. Most algorithms require a large cohort of data to make predictions, and the frequently mutated variants do not necessarily present in every single patient. Out of the five evaluated algorithms, DawnRank is the only method that identified drivers from individual samples. However, obtaining gene expression and whole exome data for diagnostics purposes are still likely to be cost-prohibitive.

Tumour heterogeneity affects driver gene predictions

Pan-cancer analysis of genomics data from an enormous cohort of patients reveals different oncogenic drivers between tumour types [163]. Heterogeneity is also observed within breast cancer, and this intertumour heterogeneity is often illustrated through various molecular or clinical features, including staging, molecular subtyping or biomarkers such as hormone

receptors [10]. These clinical variables often dictate the targeted treatment available to patients and show differences in the overall prognosis.

Analysing the algorithms' predictions on subsets of TCGA & METABRIC datasets illustrate some potential subtype-specific driver genes. OncodriveFML, OncodriveCLUSTL and 20/20+ predicted *GATA3* to be a driver for luminal breast cancer subtypes, but not ER- samples. This is accordant with the literature, where Pereira et al. [9] observed that *GATA3* is more frequently mutated in ER+ breast cancer, and Yoon et al. [164] correlated *GATA3* low expression to estrogen negativity and worse survival. Also, DawnRank and 20/20+ identified *BRCA1* as one of the significant drivers for basal-like datasets, which is in line with previously published reviews of higher *BRCA1* mutation frequencies in basal-like and triple-negative breast cancer [110, 165].

Analysing the whole breast cancer cohort as one large dataset could potentially add noises to the overall prediction. This is evident with the additional driver genes identified when algorithms were run on subtype-specific samples, where these genes were initially deemed to be insignificant in the full dataset. DriverNet and DawnRank predicted *AKT3* as a driver for luminal A subtype and this gene did not appear as significant in the all-subtypes analysis. Although the role of this gene in luminal A breast cancer need further evaluation, a recurrent translocation of *AKT3* has been observed in triple-negative breast cancer [157]. Other additional subtypes driver genes identified by DawnRank are *CCND1* and *PAK1* (luminal B), *MED1* (HER2-enriched) and *JAK2* (basal-like). Shrestha et al. described *PAK1* amplification involvement in MAPK and MET signalling [166], while *CCND1*-amplified samples are linked to worse survival [167]. Yang et al. studied *MED1* co-amplification with HER2 and suggested its critical role in HER2-enriched tumours [113]. Similarly, *JAK2* amplification is commonly detected in triple-negative breast cancers and is associated with drug resistance [168]. Therefore, taking all of these into account, it is crucial to consider tumour heterogeneity in predicting driver genes.

Conclusion

In summary, selecting the best algorithm to predict driver genes remains challenging in the absence of a complete gold standard driver list. The overall evaluation results presented in this chapter suggest that datasets characteristics and objective of one's study will dictate the algorithm selection. These evaluated methods also have their limitations. Their dependencies on other databases, upstream data analysis and inconsistencies across the different size of datasets are likely to be an ongoing challenge. Moreover, heterogeneity within a tumour type means algorithms need to be more precise in handling smaller datasets and the presence of subtypes within the cohort. However, despite these limitations, the evaluated computational methods have demonstrated that they could identify established critical cancer genes when applied to an appropriate dataset. These computational approaches have the potential to uncover more knowledge from many comprehensive cancer studies conducted, elevating researchers' understanding of cancer genomics that will hopefully lead to better personalised cancer treatments in the future.

Chapter 4 - Deep Learning Application for Breast Cancer Subtyping

4.1. Summary

The increasing size and complexity of publicly available genomics and transcriptomics data have motivated many deep learning applications in extracting valuable insights from these studies. As multi-omics data are heterogeneous, neural networks algorithms become particularly attractive because they can be set up in many different ways to support data integration. However, deep learning techniques require a large number of training samples to deal with high-dimensional data effectively. Despite the recent growth in available datasets, the number of omics variables is still considerably larger than the availability of training samples, and this often leads to overfitting. Thus, deep learning models usually adopt dimensionality reduction and regularisation steps into their neural network design to handle this challenge.

This thesis presents a novel deep learning implementation for integrating multi-omics data, called Moanna, that combines a semi-supervised autoencoder with a multitask learning. The autoencoder's role in this design is to reduce the dimensionality of the integrated input features consisting of gene expression, copy number and somatic mutation data. The extracted features from autoencoder are then piped into multiple feed-forward neural networks (FFNN) for supervised classifications of multiple related biological labels. This setup follows a hard parameter sharing of a multitask learning algorithm, acting as the regularisation for the overall network. This whole deep learning architecture is then trained together to optimised for a combined objective function from the individual network. The significant advantage of this setup is that it allows the network to be extended to include other tasks in the overall learning. This future application will be discussed further in Chapter 5.

For this thesis' implementation, Moanna is being developed in the context of predicting breast cancer subtypes and hormone receptor status. As discussed in the previous chapters, breast cancer is a heterogeneous disease consisting of various well-studied molecular subtypes that exhibits different characteristics and mutation profile. Multiple methodologies have been used to predict breast cancer subtypes in research and clinical, but their classifications often vary between methods. These differences could suggest that other biological variances are still not captured by the current approaches, as some of the most popular methods are only analysing gene expression profiles of a few critical genes. With the availability of other genomics information from datasets like TCGA, it motivates this study to develop a deep-learning-based method (Moanna) for integrating multi-omics data to investigate the subtypes classification further.

This chapter presents the details of this algorithm in a manuscript format (*material has been submitted for publication to Bioinformatics journal*). This manuscript also describes the background of this method, evaluation of its application on breast cancer data, and a discussion about the predicted subtypes. In brief, the clustering analysis of the dimensionality reduction step demonstrates the efficacy of Moanna's autoencoder in dealing with high-dimensional multi-omics data. Clusters identified by the extracted features have already grouped the samples based on their subtypes even prior to the supervised classification steps.

Evaluating the whole classification model, Moanna attains a high accuracy in predicting PAM50 subtypes (85%), differentiating basal-like samples (98%), and determining ER-status (96%). Overall, in comparison to the original PAM50 classifications, breast cancer subtypes predicted by Moanna show a stronger correlation with patient survival.

4.2. Manuscript submitted for review (Bioinformatics): *Moanna: Multi-Omics Autoencoder-based Neural Networks Algorithm for Predicting Breast Cancer Subtypes*

The following manuscript describes the deep learning model developed for this research and has been submitted for publication. This work is included in an article format.

Moanna: Multi-Omics Autoencoder-based Neural Networks Algorithm for Predicting Breast Cancer Subtypes

Richard Lupat[1, 2, *], Sherene Loi[1, 2], Jason Li[1, 2]

[1] Peter MacCallum Cancer Centre, Melbourne, Australia

[2] The Sir Peter MacCallum Department of Oncology, University of Melbourne, Parkville, Australia

Abstract

Motivation: Cancer subtyping delivers valuable insights into the study of cancer heterogeneity and fulfils an essential step towards personalised medicine. Breast cancer, for example, has been studied to show multiple molecular subgroups with distinct characteristics that are associated with patient survival and treatment responses. Despite being widely adopted, there are a variety of methodologies used to predict breast cancer subtypes, with each defining different classifications. These differences suggest that some biological information has not been integrated into the current methods. Moreover, recent computational approaches to address this challenge have been limited by their dependency on incomplete prior knowledge and difficulties handling high dimensional data beyond gene expression.

Results: We propose a novel deep-learning-based algorithm, Moanna, that is trained to integrate multi-omics data for predicting breast cancer subtypes. Moanna's architecture consists of a semi-supervised autoencoder attached to a multi-task learning network for generalising the combination of gene expression, copy number and somatic mutation

data. We trained Moanna on a subset of METABRIC breast cancer dataset and evaluated the performance on the remaining hold-out METABRIC samples and a fully independent cohort of TCGA samples. We evaluated our use of autoencoder against other dimensionality reduction technique and demonstrated its superiority in learning patterns associated with breast cancer subtypes. The overall Moanna model also achieved high accuracy in predicting samples' ER-status (96%), differentiating basal-like samples (98%), and classifying samples into PAM50 subtypes (85%). Moreover, Moanna's predicted subtypes show a stronger correlation with patient survival when compared to the original PAM50 subtypes. In summary, Moanna provides a novel approach in integrating multi-omics data for predicting different breast cancer biomarkers, and the extensible neural network architecture allows future application to other cancer studies.

Availability and Implementation: Moanna is an open-source deep-learning implementation in PyTorch. Source code, containerised package, training data, and the final model are available for download at https://github.com/rlupat/moanna.

Contact: richard.lupat@petermac.org

Supplementary Information:

Introduction

Cancer is characterised by abnormal cells that are invasive and growing out of control [1]. Each cancer type, such as breast cancer, can be further categorised into multiple subtypes through histopathological and clinical characteristics, and more recently, through molecular profiling of the primary tumour [2-6].

Cancer subtyping provides valuable molecular insights that help achieve personalised treatments. In breast cancer, multiple studies have demonstrated that tumours with different pathological and molecular features display different biological characteristics despite being originated from the same site [2-6]. These studies have identified four main primary breast cancer intrinsic subtypes, namely luminal A, luminal B, HER2-enriched and basal-like subtypes, through unbiased hierarchical clustering of gene expression patterns among the samples [2-6]. The primary characteristics of the subtypes are based on the expression levels of estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2) and proliferation indicator Ki67 [2-6].

Breast cancer subtypes have been associated with distinctive clinical presentations, risk factors, responses to treatments and prognosis profile [7, 8]. The ER-positive group has higher 5-year overall survival and relapse-free survival than the ER-negative tumours, and better response to hormonal therapy such as tamoxifen [7, 9]. Luminal A is the most common subtype of breast cancer and has a better prognosis compared to luminal B, which occurs in 10%-20% of breast cancer cases [9]. HER2-enriched group, which happens in 5%-15% of breast cancer, proliferates faster with worse prognosis but is more likely to respond to HER2-targeted therapy, such as trastuzumab or lapatinib [9].

Triple-negative breast cancer (TNBC), which includes most basal-like tumours, tends to be more aggressive and has the worst prognosis among all other subtypes with few targeted therapy available [9].

There are many published methodologies to identify the intrinsic subtypes of breast cancer. Two of the most frequently used methods in clinical setting are either immunohistochemistry (IHC)-based markers or gene expression-based assays. PAM50 (50-gene signature), MammaPrint (70-gene signature) and BluePrint (80-gene signature) are examples of assays based on gene expression [10-12]. Subtypes identified by these methods are able to predict prognosis and potential targeted therapies that benefit patients [13]. However, multiple studies have shown that breast cancer subtypes identified by these methodologies do not always align, with as high as 25% discordance rate between IHC-based method and MammaPrint/ BluePrint [11] and 38.4% between IHC-based subtype and PAM50 [14]. The inconsistencies could also be attributed to intra-tumour heterogeneity, where samples are composed of multiple subtypes [15-17]. In addition, the PAM50-classifier has been demonstrated to have limitations if ER-status is not balanced within the dataset [18]. Therefore, there is a scope to further improve the precision of the methodologies used to identify subtypes.

In this study, we introduce a neural network algorithm for predicting breast cancer subtypes using the combination of gene expression, copy number variation and somatic mutation data. Apart from gene expression profiles, studies have shown that breast cancer subtypes show different patterns of mutations and copy number aberrations [19-22]. Basal-like breast cancer is characterised by a high prevalence of *TP53* mutations, and deletion of *RB1* and *BRCA1,* while *ERBB2* amplification is often associated with

HER2-enriched subtypes [19]. On the other hand, the two luminal subtypes are frequently observed with *PIK3CA* mutations, with luminal B also shows a higher frequency of mutated *TP53* gene than luminal A [19]. Therefore, we hypothesise that integrating these different sources of omics data through a deep learning model will improve prediction for subtype classification.

Recent advances in the field of machine learning have enabled deep learning algorithms to be applied more widely on cancer data. Specifically, innovations in computer vision artificial intelligence have assisted developments in radiographic imaging and digital pathology[23-26]. For instance, deep learning techniques have been applied to diagnose metastasis in lymph nodes of breast cancer patients from whole-slide pathology images [24] and to automatically classify lung cancer tissue into its specific lung cancer subtypes [25]. Algorithms such as DeepSurv [27] and Cox-nnet [28] built prognosis predictors using artificial neural network extension of the Cox regression model. Other deep learning-based methods such as Tybalt uses autoencoders, an unsupervised neural network approach, to extract biologically relevant features from gene expression data [29].

One of the difficulties of deep learning application in genomics is its high-dimensional data. The number of genes available is significantly larger than the availability of training data, leading model to often overfits. Deep learning implementation, such as DeepCC [30], uses functional pathways to transform input gene expression data, while DeepTRIAGE [31] converts its input features through Gene Ontology (GO). These prior-knowledge-based dimensionality reduction techniques have an excellent advantage in its interpretability [32-34]. However, they have also been described to

have some limitations, particularly around bias on the knowledge that is still incomplete, as well as the inability to include all genes in the datasets [32, 33]. Moreover, they often only work for a single point of data, in this case, only gene expression data [32, 33], and thus not applicable to multiple omics data integration.

Our proposed solution in this paper is to develop a multi-omics neural network-based algorithm (Moanna) to classify molecular breast cancer subtypes using a semi-supervised autoencoder layer that is jointly trained with supervised feed-forward neural network multi-task classification layers. It is important to note that the main aim of this study is not to identify new clusters, but rather to further refine current methodology with the help of state-of-the-art neural network models in integrating copy number and somatic mutation data on top of the well-evaluated gene expression data. The employed dimensionality reduction technique is designed to computationally generalise the high-dimensional multi-omics data, away from the limitation of prior knowledge method. Thus, the implementation will then serve as a proof of concept for future Moanna's application in predicting other breast cancer biomarkers, such as the percentage of Tumour Infiltrating Lymphocytes (TILs) and for building a deep-learning-based prognosis model.

Materials & methods

In this section, we outline the detailed description of our proposed deep neural network architecture, Moanna, as well as the datasets used to train, validate and test the breast cancer subtyping model.

Overall design of Moanna's neutral-network architecture

Moanna is a deep learning framework that combines multiple supervised and unsupervised neural network architecture. This setup is adapted from the idea of semi-supervised autoencoders, or also known as ladder network, where a supervised learning method is attached to a deep autoencoder to assist in filtering irrelevant features [35]. This allows both networks to be jointly trained, instead of only utilising the autoencoder as a separate pre-training model for dimensionality reduction [35-37]. In our implementation, we build Moanna with two major components (Figure 1):

1) Semi-supervised autoencoder layer

   Each of the samples in the datasets consists of approximately 47,000 features, containing the details of gene expression, copy number and somatic mutation profiles from over 15,000 genes. Small datasets with a large number of features (large p; small n problems) is a common obstacle of deep learning application, where feature engineering step is required to prevent over-fitting [32]. Moanna employs an autoencoder to discover the latent representation of the input variables, extracting the non-linear relationship between the original data. Autoencoder is an unsupervised machine learning technique consisting of an encoder function that map input features to internal representation and a decoder function that tried to recreate the original form of these encoded features [38]. The model trained itself by optimising the loss function of the difference between its original and reconstructed input.

2) Multi-task classifications layers

   Latent features from the bottleneck layer above are carried into several feed-forward neural networks for supervised classification. For this study, we are using multiple breast cancer biomarkers, including ER status, HER2 status and

PAM50 subtypes, as our training labels. This setup is an implementation of deep neural network multitask learning which has been described to be useful in improving independent multi-class classifications by reducing overfitting in general [39]. In addition, breast cancer samples' hormone receptor status and subtypes have been studied to be correlated and it is therefore intuitive that the classification neural network should share common variables. This leads to the design of Moanna, where classification tasks are sharing some mutual hidden layers and parameters.

Neural network training

A joint supervised and unsupervised neural network training allows better generalisation in data learning [36]. The sum of loss functions from the two components becomes the objective function that is used to train this model. For semi-supervised autoencoder, Moanna measures the mean-squared error between the input and reconstructed layer. On the other hand, cross-entropy loss between training and predicted classification labels were calculated for the classification tasks. This objective function was jointly optimised with a single backpropagation using stochastic gradient descent algorithm, eliminating the necessity to set up multiple independent sets of training. Therefore, apart from better generalisation, this neural network architecture is also more efficient computationally [36].

Neural network parameters estimation

We performed a grid search to select the optimum Moanna parameters with the highest classification accuracy on our hold-out validation data (Figure 2). The final designed model consists of 1 autoencoder and 5 supervised classifiers. The encoder part of this autoencoder was designed with 2 hidden layers of 256 and 128 neurons, a representation layer of 64 encoded neurons and Tanh activation function. The decoder

part of the model mirrored the encoder setup on the other side. The classifiers took these 64 encoded features through a hidden layer of 40 neurons. Moanna used Stochastic Gradient Descent (SGD) as its optimiser for backpropagation with a learning rate of 0.005 and momentum of 0.9, over 100 epochs.

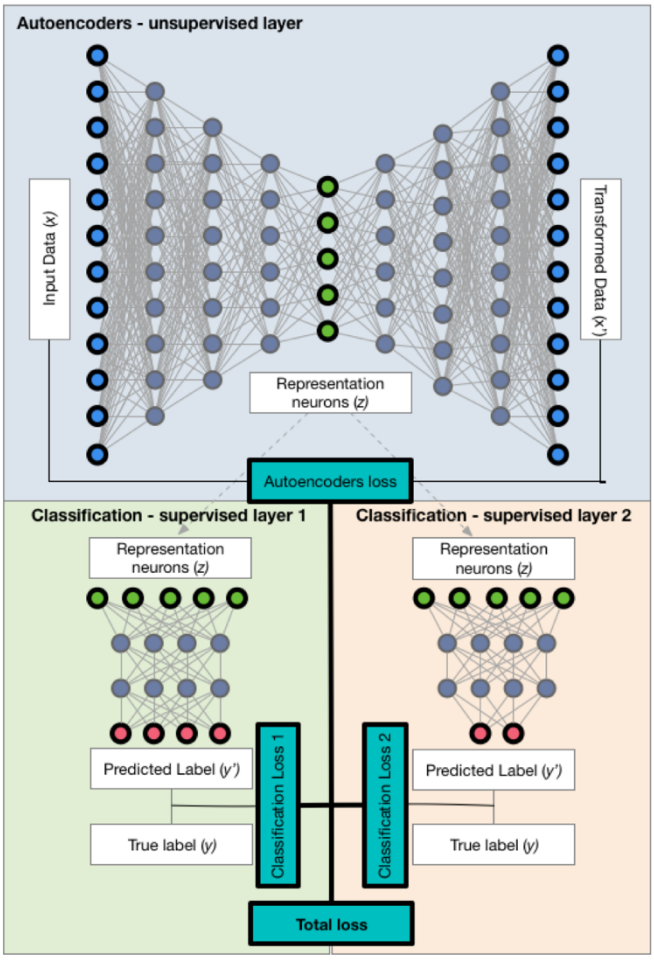


Figure 1. Overview of Moanna's neural network architecture for predicting breast cancer subtypes using multi-omics data. Moanna consists of a semi-supervised autoencoder for dimensionality reduction and a multi-task learning classification layers for predicting different breast cancer biomarkers. For the results described in this paper, Moanna used input data that comprised of equal number of gene expression, copy number variation, and somatic mutation features from over 15,000 genes. Moanna is then trained by minimising the sum of autoencoder and classification losses.

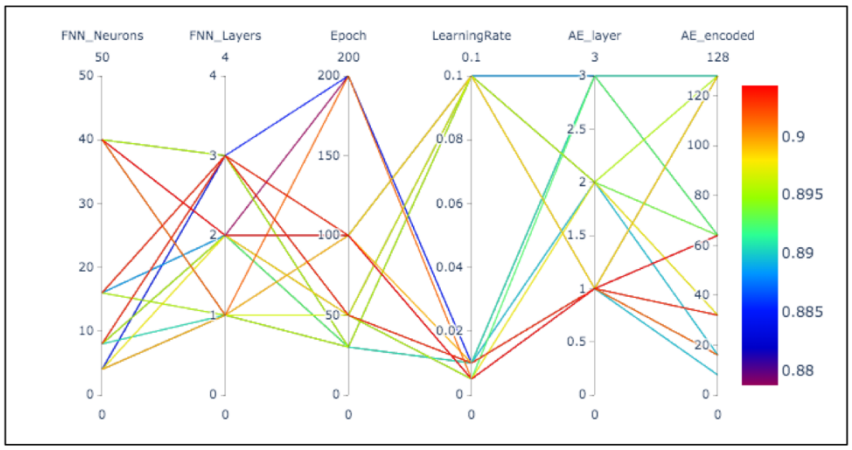


Figure 2. Parallel coordinates plot of different Moanna's parameters combination. Red arrows represent the combinations that Moanna employs on its final model. These are the parameters with the highest PAM50 subtype prediction accuracy from doing a grid search on our validation data.

Breast cancer datasets

Moanna was trained on Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) [21, 22] datasets downloaded from cbioportal [40, 41]. METABRIC is a comprehensive breast cancer study from over 2000 primary tumours, including gene expression and copy number profiles of 25160 genes alongside somatic mutations of 173 frequently mutated breast cancer genes. This dataset also comprises of clinical data and long-term follow-up information, including the PAM50 subtypes, estrogen receptor (ER) and HER2 status that Moanna uses as its training label. We excluded samples that are not one of the four intrinsic subtypes (Basal-like, HER2-enriched, Luminal A and Luminal B) and samples that do not have all three genomics profiles (gene expression, copy number and somatic mutation). This left us with a total of 1689 samples which are then randomly split into 70% training and 30% hold-out validation data. The distributions of subtypes from this METABRIC dataset is shown in Table 1.

To evaluate the robustness of Moanna, we use METABRIC-trained Moanna model for predicting subtypes of independent breast cancer datasets from The Cancer Genome Atlas (TCGA) [19, 20]. This TCGA dataset was also retrieved from cbioportal [40, 41], where a total of 954 samples were selected using the same criteria that we applied for METABRIC. Majority of these samples come with PAM50 subtype, ER, HER2 status and long-term follow-up information. The distributions of subtypes from this TCGA datasets is shown in Table 1.

Table 1. PAM50 subtype samples distribution from our training, validation and testing datasets.

PAM50 Subtype	Training (70% METABRIC) n= 1182	Validation (30% METABRIC) n = 507	Independent Test (TCGA) n = 631
Basal-like	18% (n= 213)	17.6% (n= 89)	17.7% (n=112)
HER2-enriched	12.8% (n= 151)	16.2% (n=82)	9.4% (n=59)
Luminal A	41% (n= 485)	39.8% (n=202)	52.8% (n=333)
Luminal B	28.2% (n= 333)	26.4% (n=134)	20.1% (n=127)

Data Pre-processing

One of the major issues when dealing with gene expression profiles is the different platform used to generate these data and possible batch effects associated with the experiments. Gene expression data from METABRIC were obtained through microarray data on Illumina HT-12 v3 platform while TCGA transcriptomic profiles were from RNA-sequencing performed on Illumina HiSeq. Hence, we used the relative expression (z-score transformed) calculated by cbioportal where expression values have been further normalised based on the distribution of the diploid samples in the datasets.

For copy number variation (CNV) and somatic single nucleotide polymorphism (SNP) data, information is summarised into a matrix form of gene and sample combination. CNV data has a range of [-2, 2], where 0 is copy number neutral; -1 represents heterozygous deletion; -2 indicates homozygous loss; 1 and 2 are low-level gain and high-level amplification respectively. SNP data is constructed in a binary format where 0 indicates no detected somatic mutation in that gene, and 1 represents a mutated gene. For METABRIC, any genes that are not sequenced by the targeted panel will be assigned 0 for its somatic mutation status.

The combinations of these pre-processed data were used as the input features to Moanna. Equal number of features from each *'omics type* (gene expression, CNV, SNP) were included in the overall neural network design. For the results presented in this paper, we only include genes that have expression values in both METABRIC and TCGA datasets. After filtering, our input features consisted of approximately 47,000 input features from over 15,000 genes.

Results

Autoencoders as the best dimensionality reduction method through biomarker cluster analysis

To address large p small n problems [32] on our datasets, we evaluated multiple dimensionality reduction techniques to prevent overfitting or poor generalisation to new data. Moanna's architecture uses autoencoders for feature extraction, where input features are transformed into a set of reduced input vectors through linear or non-linear mapping. This strategy can be compared to principal component analysis (PCA) where high dimensional data is transformed to a series eigenvectors and eigenvalues, and the

top *N* principal components represented the majority of the variance of the original data [18, 32]. On the other hand, alternative strategies through feature selection based on prior knowledge or level of activities have also been widely applied [32]. For this comparison, we will contrast Moanna's extracted features against randomly selected genes, PAM50 genes, top differentially expressed genes and features extracted from the top 64 principal components.

We first projected the input data into two-dimensional space with t-SNE [42] and compared the sample distribution with the t-SNE plot of the extracted features from Moanna's autoencoder. Figure 3 reports multiple clusters from Moanna's extracted features annotated by PAM50 subtypes and ER status. This indicates that the 64 neurons from the neural network model's representation layer have extracted important biological characteristics of the 47,000 input features for the purpose of subtyping, even before going through the final classification layers. We observed the same result when we repeat the exercise on TCGA breast cancer datasets, showing a vast improvement when compared to the clusters from the original input features.

To further evaluate the performance of Moanna' autoencoder, we performed clustering analysis on different selected and extracted features. The comparison includes: 1) gene expression of 50 genes from PAM50, 2) top 200 differentially expressed genes (DEG), 3) first 50 principal components (from PCA) of all input features, 4) first 50 principal components (from PCA) of all gene expression input features, and 6) randomly selected 64 genes. Following the clustering evaluation strategy from Geddes et al. [43], we calculate three metrics for assessing the performance of these dimensionality reduction strategies in retaining relevant features required for clustering breast cancer samples to

their subtypes. These metrics are Fowlkes-Mallows index (FM), Adjusted Rand Index (ARI) and normalised mutual information (NMI) score, which was calculated for each method after running k-means clustering on its selected / extracted features. In addition, we apply these features into Moanna's feed forward neural network, by replacing the autoencoder layer, to measure their usefulness when employed to solve classification problem.

The result of this evaluation (see Table 2) indicates that Moanna's autoencoder performed the best in clustering samples to their subtypes. It also achieved an overall better accuracy when deployed alongside a neural network classifier, in comparison to other dimensionality reduction techniques tested. Moanna's extracted features are better at clustering samples to subgroups, and significantly improved clusters that are only based on 50 genes from PAM50. We used feature selection on PAM50 genes as our benchmark for this evaluation, given that our subtype training labels were originated from this 50-gene signature, and that they were expected to perform the closest to the label. On the other hand, although it struggled to separate the clusters of luminal samples, unsupervised feature extraction using PCA achieved reasonable high classification accuracy when paired with Moanna's multi-task learning (see Figure 3).

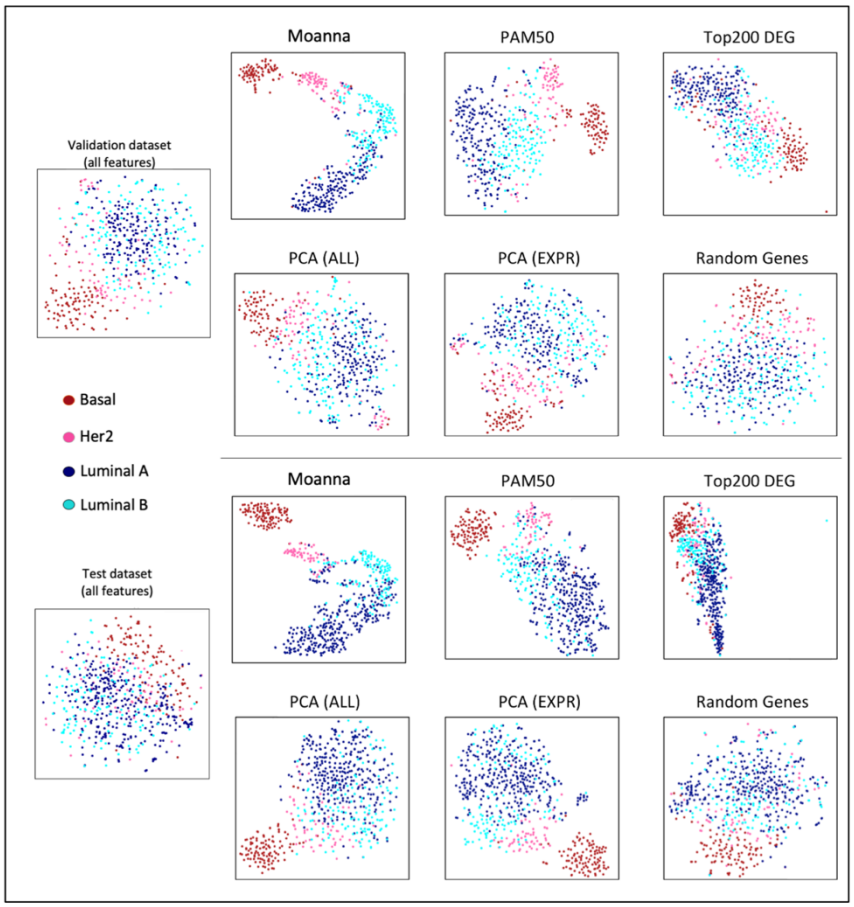


Figure 3. T-SNE plots of all the extracted/selected input-features through various dimensionality techniques described in Table 2. Top plots are from validation dataset, while bottom half plots are from testing dataset. Feature selection based on PAM50 genes is our benchmark for this evaluation, given that our subtype training labels were originated from this 50-gene signature. The result indicates that Moanna's autoencoder performed best in clustering samples to their subgroups even before entering the multi-task learning layer.

Table 2. Results of dimensionality reduction evaluation based on clustering metrics and classification accuracy when applied to Moanna's multi-task learning classifier. To measure how well the extracted features are clustering samples to its subtype, we did k-means clustering on the encoded space and measured the adjusted rand index (ARI), normalized mutual information score (NMI) and Fowlkes-Mallows (FM) index of the clusters. We compared this with other strategies such as feature extraction with principal component analysis (PCA) and feature selection of 1) 50 genes from PAM50 (PAM50), 2) top 200 differentially expressed genes (DEG) and 3) randomly selected 64 genes. We completed this evaluation on both validation (V) and testing (T) datasets. The various clustering indices show that semi-supervised autoencoder used in Moanna performed better and achieved a better overall accuracy on the independent testing dataset than other techniques compared.

Metrics	Dataset	Moanna	PAM50	Top DEG (200)	PCA (All)	PCA (EXPR)	Random
+ k-means clustering							
ARI	V	**0.628**	0.470	0.262	0.323	0.242	0.206
	T	**0.621**	0.536	0.311	0.306	0.264	0.250
NMI	V	**0.629**	0.487	0.303	0.347	0.300	0.259
	T	**0.630**	0.476	0.293	0.376	0.389	0.310
FM	V	**0.733**	0.628	0.467	0.536	0.474	0.422
	T	**0.752**	0.718	0.543	0.527	0.502	0.559
Classification Accuracy							
ER Status	V	0.964	0.966	0.921	0.961	**0.968**	0.935
	T	**0.946**	0.935	0.905	0.941	0.937	0.926
HER2 Status	V	0.959	**0.974**	0.880	0.937	0.945	0.878
	T	**0.864**	0.853	0.773	0.854	0.859	0.773
PAM50 Subtype	V	0.850	0.838	0.755	0.805	**0.854**	0.694
	T	0.848	**0.851**	0.791	0.810	0.843	0.754

Note: Bold denotes the best in its category

Moanna achieves high accuracy in predicting ER-status, HER2-status and PAM50 subtypes

We applied the proposed method on our training datasets (70% METABRIC, n= 1182) and evaluated the classification accuracy, precision and recall on our validation samples (30% METABRIC, n= 507). Table 3 summarises Moanna classification performance where it accurately differentiates well-characterised markers, for instance, differentiating ER-positive (ER+) and ER-negative (ER-) samples (96.5% accuracy),

as well as the difference between basal and non-basal-like samples (98.4% accuracy). In addition, the majority of the subtypes predicted by Moanna (85.6%) agree with the original subtypes identified by PAM50. From a total of 507 validation samples, Moanna classifies 16.6% (n= 84) basal-like, 13.6% (n= 69) HER2-like, 32.7% (n=166) LumA-like and 22.1% (n=112) LumB-like subtype.

We then further evaluated the 76 samples that were classified differently by Moanna in comparison to PAM50 (see Figure 4a). We found that 28.9% (n= 22) of the dissimilarities are on ER+/HER2- High Proliferation samples that were classified as Luminal B-like by Moanna, but predicted as Luminal A in PAM50. There were also 15.8% (n=12) samples that are ERBB2 amplified and classified as HER2-enriched by Moanna but called differently in PAM50. This discordance suggests that this Moanna's subtype prediction model did not only fit to the training subtypes label, but also integrated information learnt from ER and HER2-status predictions.

Table 3. Classification metrics on multiple tasks predicted by Moanna on validation (V) and testing (T) datasets.

Classification Task	Dataset	Accuracy	Precision*	Recall*	F1-score*
ER Status	V	0.964	0.965	0.964	0.965
	T	0.946	0.947	0.946	0.947
HER2 Status	V	0.959	0.960	0.959	0.959
	T	0.864	0.872	0.863	0.844
PAM50 Subtype	V	0.850	0.857	0.850	0.852
	T	0.848	0.864	0.848	0.852
Basal vs other subtypes	V	0.984	0.984	0.984	0.984
	T	0.989	0.989	0.989	0.989

* weighted-average, calculated by scikit-learn package [44]

Application of Moanna on independent datasets show the model does not overfit

We applied METABRIC-trained Moanna on an independent dataset (TCGA breast cancer samples) to evaluate the robustness of the architecture when dealing with new data from different experiments. Table 3 shows the precision and recall from this classification are consistent with the previous result Moanna achieved on the METABRIC validation dataset. The model predicted the ER status at 94.7% accuracy when compared to the label acquired from cbioportal. It also managed to differentiate basal-like samples from the other subtypes at 98.9% accuracy while 86.4% of the subtypes predicted are concordant with the PAM50 subtype from TCGA. From a total of 631 test samples, Moanna classifies 17.6% (n= 111) as basal-like, 7.6% (n= 48) as HER2-like, 52.8% (n=333) LumA-like and 20.1% (n=127) LumB-like.

Figure 4b shows the confusion matrix of Moanna's classification from both datasets, where it is obvious that the proportion of samples' subtypes are not balanced. HER2-enriched subtype has the least number of samples while luminal A samples represent almost half of the cases on both datasets. Imbalance class training has been studied to affect classifiers' performance [45], and we hypothesized that this would be one of the reasons for the lower concordance between predicted HER2-like subtype and the training label. The other major dissimilarities are concentrated between the classification of the two luminal subtypes. A few studies on the same datasets have identified potential admixed cases in luminal A and luminal B samples, as well as further subclasses due to the heterogeneity of luminal breast cancer [15-17].

Figure 4. a) The stacked bar plots show the differences between PAM50 and Moanna's predicted subtypes from the 76 misclassified samples, grouped by the 3-gene classifier and the HER2 copy number gain status from SNP6 data. Almost half of the disagreements happened around 1) ER+/HER2- High Proliferation samples that were classified as Luminal B-like in Moanna, but predicted as Luminal A in PAM50 2) Samples with HER2 gain were predicted as HER2-like in Moanna, but not in PAM50. b) Heatmap visualisation of Moanna's confusion matrix (left: validation data; right: testing data). Results from both classifications are consistent, where basal-like subtypes prediction matched the PAM50 subtypes and dissimilarities are concentrated around Luminal A and Luminal B predictions.

Moanna's predicted subtypes show better correlation to patients' survival

To validate the clinical significance of Moanna's classification, we perform disease-free-survival analysis using these predicted subtypes. Kaplan-Meier plots (Figure 5) show that Moanna's predicted subtypes display a more distinct separation of survival patterns compared to the original subtypes. To assess this further, we compare the prognosis between the two luminal subtypes (LumA-like vs LumB-like), which is one of the main dissimilarities between Moanna's and the original PAM50 classes. Cox proportional hazard ratio from our analysis shows a stronger correlation to patient survival between luminal A and luminal B samples (HR=2.95, CI=1.45–6.00, $p<0.005$) when compared to the original subtypes (HR=1.98, CI=1.03–3.82, $p<0.005$). This is consistent with literature where luminal A has a better prognosis than luminal B patients [9]. This result also implies subtypes that were predicted differently by Moanna were not necessarily misclassified, but rather a potential improvement to the original subtyping.

Moanna performs more consistently than other machine learning classifiers

To further benchmark Moanna's performance, we constructed four others widely used machine learning algorithms for classification tasks based on random forest (RF), support vector machine (SVM), multinomial logistics regression and stochastic gradient descent (SGD) based classifier. These algorithms were trained with an identical setup, including datasets split, number of samples and input features. Figure 6 summarises the performance of all these machine algorithms when compared to the original hormone status and PAM50 subtypes. The precision and recall values indicate

similar performance across all of these machine learning implementations with Moanna and SVM being the top performers. The average f1- score (harmonic mean of precision of recall), calculated as the average of f1-score across all three classifications on independent testing datasets, shows that Moanna outperforms SVM and other methods (see Table 4).

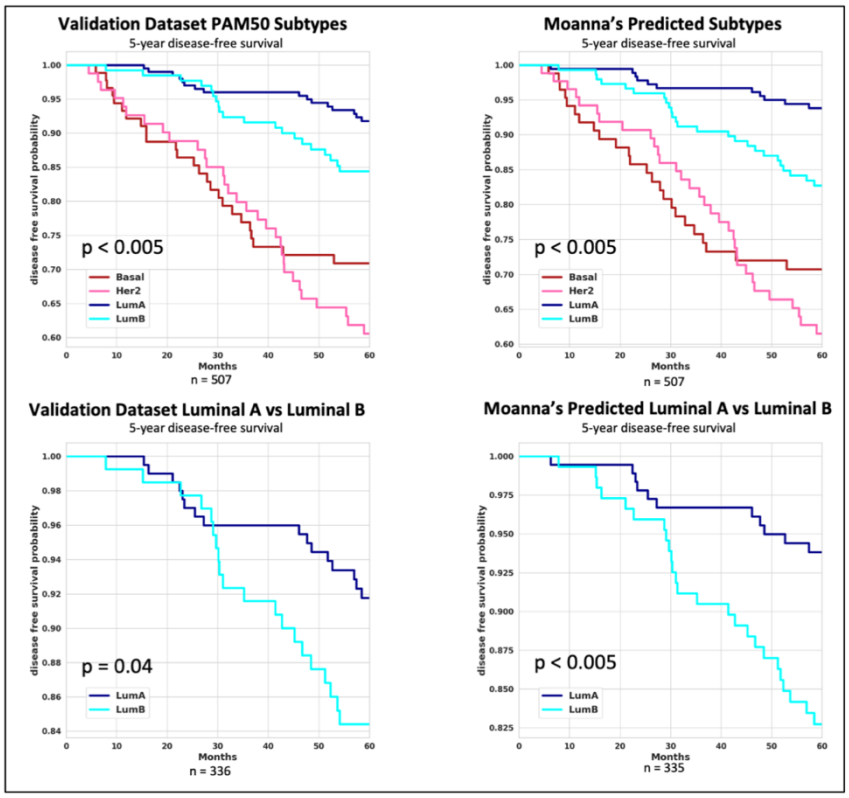


Figure 5. Kaplan-Meier survival-plot of 5-year disease-free-survival (DFS) of the predicted subtypes from all patients in our validation datasets. The top plots show the differences between all four PAM50 subtypes predicted by Moanna (right) and the original label (left). The bottom plots compare the survival analysis between Luminal A and Luminal B subtypes. Cox proportional hazard ratio analysis was performed using the Python package lifelines (https://doi.org/10.5281/zenodo.3267531).

Table 4. Comparisons of f1-score across all five machine learning algorithms on independent datasets classifications.

F1-score	Moanna	SGD Classifier	RF	Logistic Regression	SVM
ER Status	**0.947**	0.926	0.917	0.936	0.941
HER2 Status	**0.844**	0.816	0.827	0.838	0.841
PAM50 Subtype	0.852	0.851	**0.858**	0.850	**0.858**
Average	***0.881***	*0.864*	*0.868*	*0.875*	*0.880*

Note: Bold denotes the best in its category;

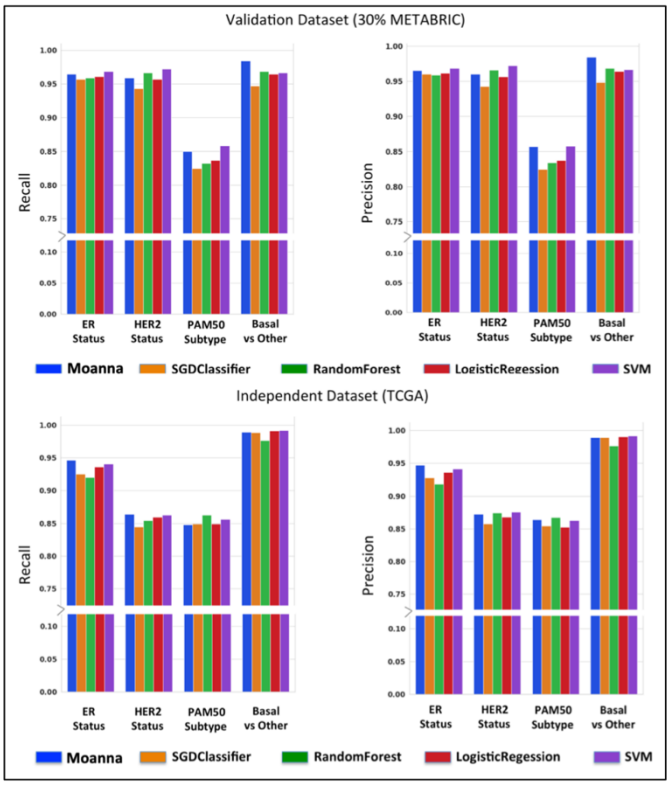


Figure 6. Precision and recall summary of Moanna's evaluation against other machine learning algorithms (top: validation dataset; bottom: testing dataset). Moanna's classification accuracy is comparable to other widely used machine learning algorithms including SGD classifier, random forest, logistic regression classifier and support vector machine.

Moanna's main driver is correlated with the genomic data type that drives PAM50 subtype classification

To assess the benefit of using multi-omics data over a single type of genomics data, we re-evaluated the classification accuracy of Moanna when trained with the individual omics data type. We set up multiple models trained on input features consisting of gene expression profiles (EXPR), copy number variation (CNV), and somatic mutation (SNP) data, and multiple combinations between them. The final evaluated Moanna referred throughout this manuscript was trained and evaluated using a combination of all three data types.

The contribution of each data types and their combinations towards the classifying breast cancer subtypes on our datasets is summarised in Table 5. Looking at individual data, it is clear that gene expression profile is a better classifier in comparison to CNV and SNP data. This is not surprising given that many studies have demonstrated the utility of gene expression assays in capturing different breast cancer subtypes, including PAM50 label that is being used for this study [3, 12, 13]. In addition, while CNV data alone do not have the same predictive power, the combined data classification result suggests that CNVs are complementing the gene expression data in improving the classification accuracy. This is consistent with literature that studies how CNVs on certain genes cause them to be up-or-down regulated [46, 47]. On the other hand, we observe that the presence of SNP data as part of our input features contributes towards differentiating basal-like subtypes from the other subtypes. This is aligned with SNP analysis of these datasets where different breast cancer subtypes were described with different frequently mutated genes. For example, basal-like datasets have a higher

frequency of *TP53* mutations, while luminal subtypes samples tend to see more *PIK3CA* mutations [19]. This analysis indicates that Moanna's neural network architecture setup provides a mechanism for combining the knowledge from different resolution of omics data to achieve good classification accuracy.

Table 5. Classification accuracy of Moanna trained with various combinations of genomics data (EXPR = gene expression profile; CNV = copy number variation; SNP = somatic mutation data). We completed this evaluation on both validation (V) and testing (T) datasets. Result suggests that EXPR drives the majority of the prediction of our model, followed by CNV, while SNP contributes towards predicting basal-like samples. Overall, the combination of all three data performed better than any individual data.

Task	Dataset	EXPR	CNV	SNV	EXPR-CNV	EXPR-SNV	CNV-SNV	EXPR-CNV-SNV
ER Status	V	**0.951**	0.901	0.807	**0.970**	0.968	0.919	0.964
	T	**0.946**	0.908	0.810	**0.948**	0.941	0.903	0.946
HER2 Status	V	**0.957**	0.955	0.866	**0.968**	0.955	0.947	0.959
	T	**0.872**	0.853	0.773	**0.862**	0.859	0.848	0.864^
PAM50 Subtype	V	**0.815**	0.649	0.513	0.826	**0.842**	0.645	0.850
	T	**0.851**	0.686	0.517	**0.857**	0.853	0.669	0.848^
Basal vs other subtypes	V	**0.961**	0.931	0.834	**0.976**	0.972	0.929	0.984
	T	**0.987**	0.959	0.838	**0.987**	**0.987**	0.954	0.989

Note: Bold denotes the best in its category; ^ denotes worse evaluation than any individual data

Discussions & Conclusion

Breast cancer is a heterogeneous disease with various subtypes that exhibits different characteristics. The four main molecular subtypes are Basal-like, HER2-enriched, Luminal A and Luminal B. These subtypes have been studied extensively to show differences in prognosis, incidence rate, and response to treatments and therapies [3, 4, 9]. Gene expression-based assays, such as the 50-gene panel called PAM50, are one of the well-established methods to infer molecular breast cancer subtypes [10]. However, there have been many studies analysing the discordance between gene expression and

IHC-based subtypes. Various explanations have been proposed, such as the limitations of these assays and the presence of intra-tumour heterogeneity [11, 14, 15, 17, 18]. To evaluate this further, we developed a novel deep-learning-based framework, Moanna, to predict breast cancer subtypes by integrating gene expression, SNP and CNV data.

In this manuscript, we demonstrated that a trained Moanna model is capable of extracting biological patterns from its training datasets and predicting the biomarkers of breast cancer samples with high accuracy. Although not all of the predicted breast cancer subtypes agree with the provided labels on the validation and testing datasets, Moanna's predicted subtypes show a more significant correlation with patient survival when compared to the original label subtypes. This suggests that the mis-predictions might not be necessarily incorrect, but rather a potential further investigation into the accuracy of the original subtype' labels.

The neural network architecture of Moanna is designed to handle the high-dimensionality of integrated 'omics data. It is a joint semi-supervised learning algorithm, based on the concept of ladder network, combining the training of unsupervised autoencoders and multi-task learning feed-forward neural networks. The ladder network design allows the autoencoder to find relevant latent variables faster by discarding irrelevant features to the classification, while maintaining a decoder that can reconstruct a representation of the input features. In addition, multitask learning setup improves the model generalisation, essentially equivalent to adding regularisation to the overall training by learning independent patterns using shared hidden layers. In combination, this implementation enables Moanna to be extended for other classifications beyond cancer subtyping.

There are, however, some limitations to this approach. First, the implementation of Moanna for breast cancer subtypes prediction currently does not work with a single sample as Moanna expected the gene expression data to be normalised against a control. Although this limitation can be addressed on future implementation by adding a baseline reference, it will still be largely restricted in the absence of normal samples in the cohort. Second, Moanna currently combines multi-omics data with early integration approach, despite dealing with combination of discrete and continuous variables. While the chosen activation functions and linear combination on hidden layers could potentially deal with this limitation, various studies have proposed better approaches in dealing with data from multiple modalities [48, 49]. One possible solution is to implement separate encoding subnetworks for extracting information from each *'omics* type, before integrating the features into the current architecture. In future work, we would explore options to extend Moanna for addressing these limitations.

In summary, we presented Moanna, a multi-omics neural network algorithm for predicting breast cancer subtypes. Through training and evaluation on public breast cancer datasets, we have demonstrated Moanna's performance in generalising knowledge extracted from gene expression, CNV and SNP data. Despite the heavy focus on breast cancer subtypes in this manuscript, Moanna's proof-of-concept implementation can be extended for predicting other biomarkers, such as the TILs or even for building a prognosis model. The generalised neural network architecture can also be deployed on other cancer types, extracting valuable information from vast amount of public cancer datasets.

Code availability

Software package

Moanna was implemented as a collection of python scripts packaged in a docker image with all its python libraries dependencies. Moanna is developed with PyTorch [50] deep learning framework.

Source code is available at https://github.com/rlupat/moanna. Pre-processed data used for the study and our trained model are also available for download.

References

1. Alberts, B., *Molecular biology of the cell*. 2015.
2. Perou, C.M., et al., *Molecular portraits of human breast tumours.* Nature, 2000. **406**(6797): p. 747-752.
3. Sorlie, T., et al., *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.* Proc Natl Acad Sci U S A, 2001. **98**(19): p. 10869-74.
4. Sorlie, T., et al., *Repeated observation of breast tumor subtypes in independent gene expression data sets.* Proc Natl Acad Sci U S A, 2003. **100**(14): p. 8418-23.
5. Herschkowitz, J.I., et al., *Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors.* Genome Biol, 2007. **8**(5): p. R76.
6. Prat, A., et al., *Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer.* Breast Cancer Res, 2010. **12**(5): p. R68.
7. Malhotra, G.K., et al., *Histological, molecular and functional subtypes of breast cancers.* Cancer Biol Ther, 2010. **10**(10): p. 955-60.

8. Weigelt, B., F.C. Geyer, and J.S. Reis-Filho, *Histological types of breast cancer: How special are they?* Molecular Oncology, 2010. **4**(3): p. 192-208.

9. Feng, Y., et al., *Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis.* Genes & diseases, 2018. **5**(2): p. 77-106.

10. Bastien, R.R., et al., *PAM50 breast cancer subtyping by RT-qPCR and concordance with standard clinical molecular markers.* BMC Med Genomics, 2012. **5**: p. 44.

11. Gao, J.J. and S.M. Swain, *Luminal A Breast Cancer and Molecular Assays: A Review.* The oncologist, 2018. **23**(5): p. 556-565.

12. Whitworth, P., et al., *Chemosensitivity predicted by BluePrint 80-gene functional subtype and MammaPrint in the Prospective Neoadjuvant Breast Registry Symphony Trial (NBRST).* Annals of surgical oncology, 2014. **21**(10): p. 3261-3267.

13. Parker, J.S., et al., *Supervised risk predictor of breast cancer based on intrinsic subtypes.* J Clin Oncol, 2009. **27**(8): p. 1160-7.

14. Kim, H.K., et al., *Discordance of the PAM50 Intrinsic Subtypes Compared with Immunohistochemistry-Based Surrogate in Breast Cancer Patients: Potential Implication of Genomic Alterations of Discordance.* Cancer Research and Treatment, 2019. **51**(2): p. 737-747.

15. Kumar, N., et al., *Quantification of intrinsic subtype ambiguity in Luminal A breast cancer and its relationship to clinical outcomes.* BMC Cancer, 2019. **19**(1).

16. Allott, E.H., et al., *Intratumoral heterogeneity as a source of discordance in breast cancer biomarker classification.* Breast Cancer Research, 2016. **18**(1).

17. Martelotto, L.G., et al., *Breast cancer intra-tumor heterogeneity.* Breast Cancer Research, 2014. **16**(3): p. R48.

18. Raj-Kumar, P.-K., et al., *PCA-PAM50 improves consistency between breast cancer intrinsic and clinical subtyping reclassifying a subset of luminal A tumors as luminal B.* Scientific Reports, 2019. **9**(1).

19. The Cancer Genome Atlas Research Network, *Comprehensive molecular portraits of human breast tumours.* Nature, 2012. **490**(7418): p. 61-70.

20. Ciriello, G., et al., *Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer.* Cell, 2015. **163**(2): p. 506-19.

21. Curtis, C., et al., *The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.* Nature, 2012. **486**(7403): p. 346-52.

22. Pereira, B., et al., *The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes.* Nature Communications, 2016. **7**(1): p. 11479.

23. Arvaniti, E., et al., *Automated Gleason grading of prostate cancer tissue microarrays via deep learning.* Scientific Reports, 2018. **8**(1): p. 12054.

24. Ehteshami Bejnordi, B., et al., *Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer.* Jama, 2017. **318**(22): p. 2199-2210.

25. Coudray, N., et al., *Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning.* Nature Medicine, 2018. **24**(10): p. 1559-1567.

26. Saltz, J., et al., *Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images.* Cell reports, 2018. **23**(1): p. 181-193.e7.

27. Katzman, J.L., et al., *DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network.* BMC Medical Research Methodology, 2018. **18**(1): p. 24.

28. Ching, T., X. Zhu, and L.X. Garmire, *Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data.* PLOS Computational Biology, 2018. **14**(4): p. e1006076.

29. Way, G.P. and C.S. Greene, *Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders.* Pac Symp Biocomput, 2018. **23**: p. 80-91.

30. Gao, F., et al., *DeepCC: a novel deep learning-based framework for cancer molecular subtype classification.* Oncogenesis, 2019. **8**(9).

31. Beykikhoshk, A., et al., *DeepTRIAGE: interpretable and individualised biomarker scores using attention mechanism for the classification of breast cancer sub-types.* BMC Medical Genomics, 2020. **13**(S3).

32. Hira, Z.M. and D.F. Gillies, *A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data.* Advances in Bioinformatics, 2015. **2015**: p. 1-13.

33. Glaab, E., *Using prior knowledge from cellular pathways and molecular networks for diagnostic specimen classification.* Briefings in Bioinformatics, 2016. **17**(3): p. 440-452.

34. Li, J., et al., *Feature Selection.* ACM Computing Surveys, 2018. **50**(6): p. 1-45.

35. Valpola, H., *From neural PCA to deep unsupervised learning.* arXiv:1411.7783, 2015(Adv. in Independent Com-ponent Analysis and Learning Machines): p. 143–171.

36. Rasmus, A.V., Harri; Honkala, Mikko; Berglund, Mathias; Raiko, Tapani, *Semi-Supervised Learning with Ladder Networks.* arXiv:1507.02672, 2015.

37. Mohammad, P., et al., *Deconstructing the Ladder Network Architecture*. 2016, PMLR. p. 2368-2376.

38. Goodfellow, I., Y. Bengio, and A. Courville, *Deep learning*. Adaptive computation and machine learning. 2016, Cambridge, Massachusetts: The MIT Press. xxii, 775 pages.

39. Ruder, S., *An overview of multi-task learning in deep neural networks.* arXiv preprint arXiv:1706.05098, 2017.

40. Cerami, E., et al., *The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data.* Cancer Discov, 2012. **2**(5): p. 401-4.

41. Gao, J., et al., *Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal.* Sci Signal, 2013. **6**(269): p. pl1.

42. Maaten, L.v.d. and G. Hinton, *Visualizing data using t-SNE.* Journal of machine learning research, 2008. **9**(Nov): p. 2579-2605.

43. Geddes, T.A., et al., *Autoencoder-based cluster ensembles for single-cell RNA-seq data analysis.* BMC bioinformatics, 2019. **20**(19): p. 660.

44. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* the Journal of machine Learning research, 2011. **12**: p. 2825-2830.

45. Buda, M., A. Maki, and M.A. Mazurowski, *A systematic study of the class imbalance problem in convolutional neural networks.* Neural Networks, 2018. **106**: p. 249-259.

46. Shao, X., et al., *Copy number variation is highly correlated with differential gene expression: a pan-cancer study.* BMC Medical Genetics, 2019. **20**(1).

47. Zhao, M. and Z. Zhao, *Concordance of copy number loss and down-regulation of tumor suppressor genes: a pan-cancer study.* 2016. **17**(S7).

48. Sharifi-Noghabi, H., et al., *MOLI: multi-omics late integration with deep neural networks for drug response prediction.* Bioinformatics, 2019. **35**(14): p. i501-i509.

49. Rappoport, N. and R. Shamir, *Multi-omic and multi-view clustering algorithms: review and cancer benchmark.* Nucleic acids research, 2018. **46**(20): p. 10546-10562.

50. Paszke, A., et al. *Pytorch: An imperative style, high-performance deep learning library*. in *Advances in neural information processing systems*. 2019.

Chapter 5 - Summary and Future Directions

The availability of large-scale public cancer datasets from projects like TCGA has provided researchers with a vast range of new insights to be uncovered. This thesis explores the application of machine learning techniques in predicting clinically-relevant molecular features from these large-scale multi-omics data. It evaluated computational algorithms that predict breast cancer driver genes (Chapter 3) and presented a novel deep learning method for classifying breast cancer to its molecular subtypes (Chapter 4). This concluding chapter summarises the finding from this research and highlights its relevance to the fields. Both topics, genomics and machine learning, are actively growing research focus, and there will be many opportunities to explore this study further, and this could lead to new hypothesis and improved algorithms. As such, it is also imperative to discuss the limitations of the approaches presented in this thesis, as well as possible future development and potential applications in classifying other cancer-related features.

5.1. Summary and Limitations

5.1.1. Evaluation of driver gene prediction algorithms

Multiple computational methods for predicting driver genes have been proposed in the literature, but the evaluation of these algorithms remains challenging in the absence of *complete* sets of comparative drivers. This thesis presented a comprehensive evaluation strategy and its application on breast cancer datasets for assessing five driver gene prediction algorithms. The assessment comprised of: comparisons of predicted driver genes against established cancer genes list, agreements between methods, prediction consistency, support for detecting copy number and subtype-specific drivers. Methods for this evaluation were selected to cover various published strategies for predicting driver genes, including network-based approaches (DriverNet and DawnRank), analysis of functional genomic positions (OncodriveFML and OncodriveCLUSTL), and a random-forest-based method (20/20+).

Overall evaluation results suggested that the selection of the most appropriate algorithms to be applied to a study is likely to depend on the objectives and datasets characteristics. Genomics datasets that are complemented with gene expression data are expected to benefit more from DriverNet and DawnRank. These network-based algorithms are also able to predict driver genes that are not frequently point-mutated, such as genes that are mostly affected by epigenetic changes, large deletion or amplification. In addition, DriverNet and Dawnrank support smaller datasets and had consistent predictions across different datasets. On the other hand, for larger datasets that contain only somatic mutation data, 20/20+ was the best-evaluated method for predicting established genes. Its predictions had the highest precision and f1-score when compared to the list of driver genes from CGC, as well as, the consensus drivers identified by the other algorithms. As a supervised machine learning model, 20/20+ excellent performance on well-studied genes is not unexpected. This prediction algorithm combines many predictive features that are likely to overlap with other algorithms' principles and was trained with some known oncogenes and tumour suppressor genes. Nevertheless, it is also important to note that genes that are identified by more than two methods are all well-

studied genes, suggesting that a consensus-based approach could be a possible alternative for high sensitivity driver gene predictions.

While all the evaluated algorithms identified some critical cancer genes, these methods also come with their drawbacks. All five algorithms depend heavily on other sources of knowledge that are still ongoing research, such as gene interaction network, functional impact score databases and predefined training driver genes. Their predictions also rely on the sensitivity of the upstream variant-calling steps. Although these genomics analysis tools have matured over the last decade, much of the analysis of true somatic variants still require additional filtering and thresholding. Furthermore, the evaluation result presented in Chapter 3 revealed that the predicted driver genes from OncodriveFML, OncodriveCLUSTL, and 20/20+ were inconsistent between different datasets. The main contributing factor is likely to be the size of the datasets, as the number of predicted genes is observed to increase as the number of samples grows.

The evaluation strategy presented in this thesis still has some limitations. As driver genes identification is an ongoing research topic, the assessment of algorithms' predictions accuracy is likely to stay incomplete in the absence of orthogonal validation of the novel predicted genes. Moreover, this evaluation could not assess the utility of these algorithms for diagnostics purposes. Most algorithms require many other samples to provide significant drivers prediction, and not all patients would have these mutations. This is compounded by the fact that tumour heterogeneity also affects the main drivers to oncogenesis. As presented in the result from Chapter 3, the analysis of subtype-specific subsets of a dataset uncovered additional potential breast cancer critical genes. From this finding, it can be hypothesised that a thorough detection of driver genes should not only analyse all mutations presented in a dataset but should also detect and take into account the molecular differences between groups of samples.

5.1.2. Deep learning model for predicting breast cancer subtypes

To interrogate breast cancer subtypes further, this thesis also presents a novel deep learning approach that integrates multi-omics data for predicting various samples' biomarkers and molecular subtypes. This method incorporates the architecture of a semi-supervised autoencoder for dimensionality reduction and a supervised multitask learning setup for the classifications. Taking an input of gene expression, somatic point mutation and copy number data, the algorithm outputs the probability of a sample belonging to classes it learnt from the training labels. In this study, this deep-learning-based model was trained to project breast cancer samples' ER-status, HER2-status and PAM50 subtypes.

The proposed neural network algorithm was trained with METABRIC data and then evaluated on TCGA breast cancer samples. The overall result demonstrated that a semi-supervised autoencoder could effectively reduce the dimensions of multi-omics data while retaining most of the essential molecular characteristics in its extracted features. Moreover, when it is combined with multi-task learning, it helps reduce overfitting when dealing with smaller training datasets. This combination also contributes to improve the accuracy of the supervised classifications, where predictions are highly concordant with the provided labels.

It differentiates ER+ from ER- samples with 96% accuracy, and 85% of the samples' predicted molecular subtypes agree with PAM50. Further analyses of these minor predictions' differences suggested that the subtypes identified by this deep learning model show a stronger correlation with patient prognosis.

While the indicated accuracy of the deep learning method presented in this thesis has been promising, this initial implementation still has room for improvements. The current approach requires gene expression data to be normalised, or standardised, before being integrated into the overall model. This normalisation step limits the application of this algorithm to only those with normal samples and larger scale datasets. This limitation can potentially be addressed by adopting baseline expression data from multiple studies and sequencing machines, as an additional pre-processing step prior to fitting the data to the trained model.

Another limitation to this proposed method was the use of early integration for combining mixed input variables. As described in Chapter 2 and Chapter 4, this thesis used gene expression data that was defined as continuous real numbers, while somatic point mutation and copy number information were provided as discrete events. Furthermore, these data were generated from different sequencing machines and experiments. Although the selected activation functions and linear combination on hidden layers in the deep learning method were likely to deal with this mixture, there are also other approaches that can be applied to explicitly handle data from multiple modalities. For example, Sharifi-Noghabi [169] proposed the adoption of late integration neural network algorithm, where each omics data type have separate encoding subnetworks to extract features, before concatenating them into a multi-omics representation. Early and late integration on multimodal deep learning have different advantages [170, 171], and this should be explored further in future iterations of Moanna.

5.2. Future direction

5.2.1. Other related works

Deep learning architecture can be designed to handle many different combinations of unsupervised and supervised learning, and this extensibility is particularly interesting for exploring patterns from heterogeneous data like cancer genomics. Beyond the details presented in the result chapters of this thesis, the proposed neural network algorithm (Moanna) was also extended to examine its application in building a prognosis model. This was done by connecting another sub-network to the encoded features of Moanna's autoencoder, and this extension was then added to the existing multitask learning setup (Figure 5.1). The objective function of this prediction adopts the proposed deep Cox proportional hazards neural network by DeepSurv [16], which predicts the risk score of a particular sample. The preliminary training shows promising results, where higher risk scores were predicted for samples that have an event within the first five years (Figure 5.1). However, the overfitting issue has not been fully addressed, and this application still requires further research and validation.

While recent development of deep learning methods has delivered promising results, its broader adoption in healthcare is still restricted by the challenges in explaining and interpreting the algorithms' decisions [172]. High-performing methods are often difficult to

explain, while more transparent approaches are usually not as accurate [173]. For the method developed in this thesis, it would also be interesting to understand how different components of Moanna come to their final predictions. As a proof-of-concept to explain this algorithm, a heatmap was generated from all the extracted features post-autoencoder (Figure 5.2). The initial analysis of these patterns suggested that some *hidden* neurons were responsible for predicting specific tasks, such as *neuron 6* (Luminal A vs Luminal B), *neuron 10* (ER+ vs ER-), and *neuron 27* (HER2+ vs HER2-) (Figure 5.2). Furthermore, by interrogating the weightings of the input features in this network, it can be observed that gene expression and copy number data are the most influential to the overall prediction. This observation is consistent with the result of the evaluation in Chapter 4. Well-studied aberrations in genes such as *ERBB2*, *ESR1*, and *GRB7*, dominated the top 10 features with the highest combined weightings across the model (Table 5.1). As part of this ongoing work, other popular techniques for addressing the black-box nature of deep learning would be further examined. This includes sensitivity analysis, layer-wise relevance propagation, and inputs perturbations [174-176] that are unfortunately beyond the scope of this thesis.

Table 5.1 - Top 10 influential genes in Moanna's subtyping model, based on the combined input features' absolute weights in the neural network. The bolded numbers represent the top individual input neurons.

	Gene Expression	Copy-Number	Point Mutation
PGAP3	**0.98**	**0.86**	0.61
STARD3	**0.91**	0.73	0.61
GRB7	**0.90**	0.75	0.60
KRT17	**0.83**	0.62	0.60
ERBB2	**0.82**	**0.82**	0.58
PNMT	0.75	**0.81**	0.57
KRT14	**0.78**	0.61	0.60
MED1	**0.77**	0.62	0.59
ESR1	0.75	0.61	0.60
CDK12	0.75	0.63	0.60

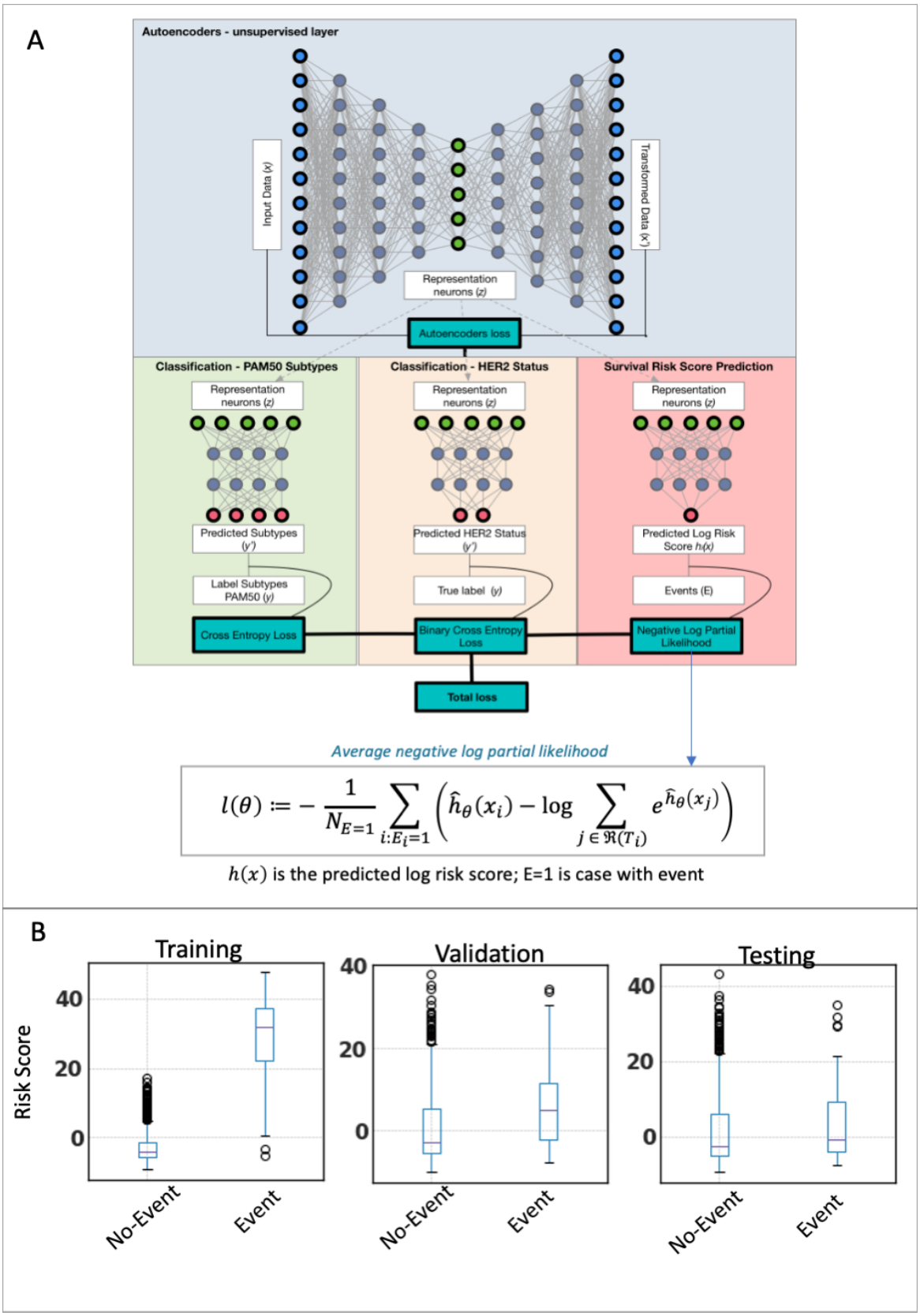


Figure 5.1 - A) Extended Moanna architecture to include survival risk score prediction (pink box) with average negative log partial likelihood as the objective function. This equation is adopted from DeepSurv [16]. B) Boxplots representing predicted risk score among groups with or without observed events, across three different datasets

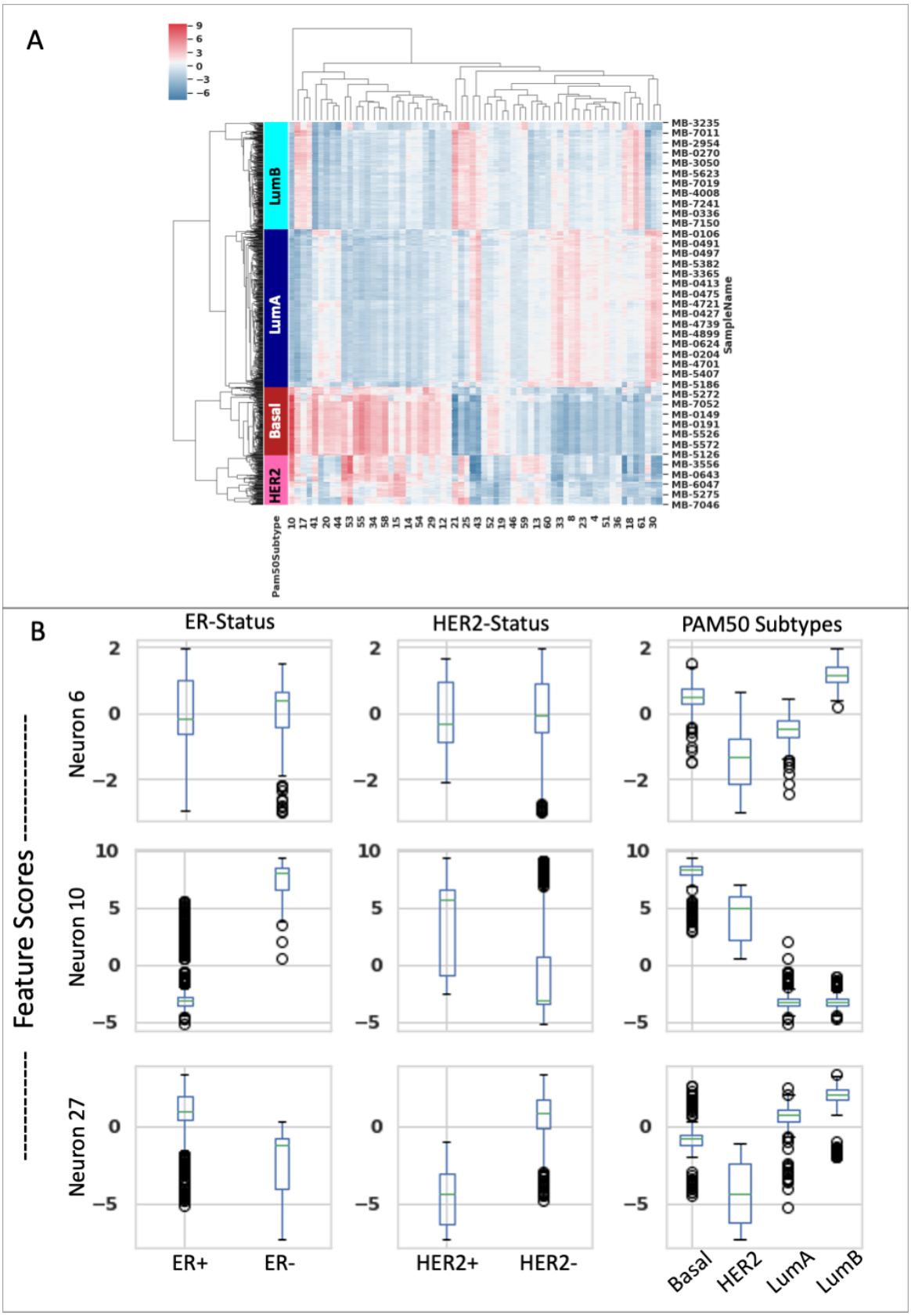


Figure 5.2 - A) Heatmap describing the values of autoencoder extracted feature across PAM50 subtypes. B) Boxplots representing the latent features (neurons) scores within different groups of comparisons.

5.2.2 Future application and data integration

Although all the methods proposed in this thesis have been described in the context of breast cancer, they were designed to be independent of the cancer site of origin. It would be interesting to evaluate Moanna's applications in classifying other cancer data, including:

1. Predicting molecular subtypes from other cancer data available in TCGA, such as ovarian cancer, acute myeloid leukemia, and lung cancer. Similar to breast cancer, molecular profiling of these cancer types has also stratified them into multiple subtypes with prognostic implications [177-180].
2. Pan-cancer molecular classifications. While cancer is usually treated based on the anatomical site, the increasing availability of pathway-specific targeted treatments means the traditional classification can be further supplemented by molecular information. Studies have identified potential classifications beyond the origin of the tissues, and these pan-cancer subtypes provide supporting information for predicting prognosis [181, 182].
3. Identifying primary sites of metastatic cancer. Cancer of unknown primary (CUP) is a common occurrence with 3-5% incidence rate, and its disease management is complicated due to the absence of conclusive site of origin [183]. However, studies have shown that comprehensive somatic mutation analysis can assist in the identification of the primary site and therefore, provide potential options for treatments [184].

As the relationship between cancer genomics and its underlying biology continue to be unravelled, integrating other information to the machine learning methods would undoubtedly be beneficial. Data from other *'omics* platform, such as DNA methylation and protein phosphorylation data, could be included in the deep learning model to improve its classifications. Furthermore, outside of sequencing, cancer management in diagnostics settings have also generated a massive amount of medical imaging data as part of its routine care and monitoring. Deep learning techniques have been widely deployed to analyse these images [185-187] and further integration to the algorithm that analysed multi-omics data, such as the one presented in this thesis, has the potential to explain some of the complexities of cancer biology.

5.3. Conclusion

In summary, this thesis demonstrates the application of machine learning in predicting clinically-relevant molecular features from large-scale genomic breast cancer datasets. It evaluated multiple driver gene predictions algorithms and concluded that, while all computational methods could identify well-catalogued genes, the choice of algorithm is driven by study objectives and datasets characteristics. Moreover, as its research output, this thesis also proposed an extensible novel deep learning algorithm for classifying breast cancer subtypes and hormone-receptor status. The breast cancer subtypes predicted by this deep learning model showed a significant correlation with patient survival.

Although the clinical utility of machine learning algorithms still requires further validation, it is the view of the author that the adoption of machine learning approaches in cancer

diagnostics will become increasingly common. These applications in cancer research will continue to assist the interpretation of large-scale cancer genomics study and improve researchers' understanding of tumour heterogeneity. Ultimately, it is hopeful that these state-of-the-art algorithms will eventually lead to the discovery of better cancer treatments that improve patients' quality of life.

References

1. International Human Genome Sequencing Consortium, *Finishing the euchromatic sequence of the human genome.* Nature, 2004. **431**(7011): p. 931-945.
2. Frio, T.R., *High-Throughput Technologies: DNA and RNA Sequencing Strategies and Potential*, in *Pan-cancer Integrative Molecular Portrait Towards a New Paradigm in Precision Medicine*, C. Le Tourneau and M. Kamal, Editors. 2015, Springer International Publishing: Cham. p. 47-68.
3. Auton, A., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
4. The Cancer Genome Atlas Research Network.
5. International Cancer Genome Consortium, et al., *International network of cancer genome projects.* Nature, 2010. **464**(7291): p. 993-998.
6. Alberts, B., *Molecular biology of the cell*. 2015.
7. Ciriello, G., et al., *Emerging landscape of oncogenic signatures across human cancers.* Nature Genetics, 2013. **45**(10): p. 1127-1133.
8. ICGC TCGA Pan-Cancer Analysis of Whole Genomes Consortium, *Pan-cancer analysis of whole genomes.* Nature, 2020. **578**(7793): p. 82-93.
9. Pereira, B., et al., *The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes.* Nature Communications, 2016. **7**(1): p. 11479.
10. Turashvili, G. and E. Brogi, *Tumor Heterogeneity in Breast Cancer.* Frontiers in Medicine, 2017. **4**: p. 227-227.
11. Gao, J., et al., *Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal.* Sci Signal, 2013. **6**(269): p. pl1.
12. Kourou, K., et al., *Machine learning applications in cancer prognosis and prediction.* Computational and Structural Biotechnology Journal, 2015. **13**: p. 8-17.
13. Quang, D., Y. Chen, and X. Xie, *DANN: a deep learning approach for annotating the pathogenicity of genetic variants.* Bioinformatics, 2015. **31**(5): p. 761-3.
14. Sundaram, L., et al., *Predicting the clinical impact of human mutation with deep neural networks.* Nature Genetics, 2018. **50**(8): p. 1161-1170.
15. Way, G.P. and C.S. Greene, *Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders.* Pac Symp Biocomput, 2018. **23**: p. 80-91.
16. Katzman, J.L., et al., *DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network.* BMC Medical Research Methodology, 2018. **18**(1): p. 24.
17. Poplin, R., et al., *A universal SNP and small-indel variant caller using deep neural networks.* Nature Biotechnology, 2018. **36**(10): p. 983-987.
18. Goodfellow, I., Y. Bengio, and A. Courville, *Deep learning*. Adaptive computation and machine learning. 2016, Cambridge, Massachusetts: The MIT Press. xxii, 775 pages.
19. Mao, H., et al., *Resource management with deep reinforcement learning*, in *Proceedings of the 15th ACM Workshop on Hot Topics in Networks*. 2016. p. 50-56.
20. Blum, A.L. and P. Langley, *Selection of relevant features and examples in machine learning.* Artificial intelligence, 1997. **97**(1-2): p. 245-271.

21. Bengio, Y., A. Courville, and P. Vincent, *Representation learning: A review and new perspectives.* IEEE transactions on pattern analysis and machine intelligence, 2013. **35**(8): p. 1798-1828.
22. Hira, Z.M. and D.F. Gillies, *A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data.* Advances in Bioinformatics, 2015. **2015**: p. 1-13.
23. Bengio, Y. and Y. LeCun, *Scaling learning algorithms towards AI.* Large-scale kernel machines, 2007. **34**(5): p. 1-41.
24. LeCun, Y., et al., *Backpropagation applied to handwritten zip code recognition.* Neural computation, 1989. **1**(4): p. 541-551.
25. Abadi, M., et al., *Tensorflow: Large-scale machine learning on heterogeneous distributed systems.* arXiv preprint arXiv:1603.04467, 2016.
26. Paszke, A., et al., *Pytorch: An imperative style, high-performance deep learning library*, in *Advances in Neural Information Processing Systems 32*. 2019, Curran Associates, Inc. p. 8024-8035.
27. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* Nature, 2015. **521**(7553): p. 436-444.
28. Kingma, D.P. and J. Ba, *Adam: A method for stochastic optimization.* arXiv preprint arXiv:1412.6980, 2014.
29. Tieleman, T. and G. Hinton, *Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude.* COURSERA: Neural networks for machine learning, 2012. **4**(2): p. 26-31.
30. Duchi, J., E. Hazan, and Y. Singer, *Adaptive subgradient methods for online learning and stochastic optimization.* Journal of machine learning research, 2011. **12**(7): p. 2121-2159.
31. Nwankpa, C., et al., *Activation functions: Comparison of trends in practice and research for deep learning.* arXiv preprint arXiv:1811.03378, 2018.
32. Srivastava, N., et al., *Dropout: a simple way to prevent neural networks from overfitting.* The journal of machine learning research, 2014. **15**(1): p. 1929-1958.
33. Ruder, S., *An overview of multi-task learning in deep neural networks.* arXiv preprint arXiv:1706.05098, 2017.
34. Pascanu, R., T. Mikolov, and Y. Bengio, *On the difficulty of training recurrent neural networks*, in *International conference on machine learning*. 2013. p. 1310-1318.
35. Hinton, G.E. and R.R. Salakhutdinov, *Reducing the dimensionality of data with neural networks.* science, 2006. **313**(5786): p. 504-507.
36. LeCun, Y. and Y. Bengio, *Convolutional networks for images, speech, and time series.* The handbook of brain theory and neural networks, 1995. **3361**(10): p. 1995.
37. Erhan, D., et al., *Why does unsupervised pre-training help deep learning?*, in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. 2010, Journal of Machine Learning Research. p. 625-660.
38. Krizhevsky, A., I. Sutskever, and G.E. Hinton, *Imagenet classification with deep convolutional neural networks.* Communications of the ACM, 2017. **60**(6): p. 84-90.
39. Valpola, H., *From neural PCA to deep unsupervised learning.* arXiv:1411.7783, 2015(Adv. in Independent Com-ponent Analysis and Learning Machines): p. 143–171.
40. Rasmus, A.V., Harri; Honkala, Mikko; Berglund, Mathias; Raiko, Tapani, *Semi-Supervised Learning with Ladder Networks.* arXiv:1507.02672, 2015.
41. Caruana, R., *Multitask learning.* Machine learning, 1997. **28**(1): p. 41-75.

42. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.
43. Denisko, D. and M.M. Hoffman, *Classification and interaction in random forests.* Proceedings of the National Academy of Sciences, 2018. **115**(8): p. 1690-1692.
44. Fawagreh, K., M.M. Gaber, and E. Elyan, *Random forests: from early developments to recent advancements.* Systems Science & Control Engineering: An Open Access Journal, 2014. **2**(1): p. 602-609.
45. Ahmad, M.W., M. Mourshed, and Y. Rezgui, *Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption.* Energy and Buildings, 2017. **147**: p. 77-89.
46. Handelman, G.S., et al., *Peering Into the Black Box of Artificial Intelligence: Evaluation Metrics of Machine Learning Methods.* American Journal of Roentgenology, 2018. **212**(1): p. 38-43.
47. Japkowicz, N., *Why question machine learning evaluation methods*, in *AAAI workshop on evaluation methods for machine learning*. 2006. p. 6-11.
48. Ting, K.M., *Confusion Matrix*, in *Encyclopedia of Machine Learning*, C. Sammut and G.I. Webb, Editors. 2010, Springer US: Boston, MA. p. 209-209.
49. Palacio-Niño, J.-O. and F. Berzal, *Evaluation metrics for unsupervised learning algorithms.* arXiv preprint arXiv:1905.05667, 2019.
50. Fanaee-T, H. and M. Thoresen, *Performance evaluation of methods for integrative dimension reduction.* Information Sciences, 2019. **493**: p. 105-119.
51. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* Journal of Machine Learning Research, 2011. **12**: p. 2825-2830.
52. Martin, T.A., et al., *Cancer invasion and metastasis: molecular and cellular perspective*, in *Madame Curie Bioscience Database [Internet]*. 2013, Landes Bioscience.
53. Gupta, G.P. and J. Massagué, *Cancer metastasis: building a framework.* Cell, 2006. **127**(4): p. 679-695.
54. Hanahan, D. and R.A. Weinberg, *The hallmarks of cancer.* Cell, 2000. **100**(1): p. 57-70.
55. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-674.
56. Crick, F., *Central dogma of molecular biology.* Nature, 1970. **227**(5258): p. 561-563.
57. Talseth-Palmer, B.A. and R.J. Scott, *Genetic variation and its role in malignancy.* International journal of biomedical science : IJBS, 2011. **7**(3): p. 158-171.
58. Loewe, L. and W.G. Hill, *The population genetics of mutations: good, bad and indifferent.* Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 2010. **365**(1544): p. 1153-1167.
59. Stratton, M.R., P.J. Campbell, and P.A. Futreal, *The cancer genome.* Nature, 2009. **458**(7239): p. 719-724.
60. Jackson, S.P. and J. Bartek, *The DNA-damage response in human biology and disease.* Nature, 2009. **461**(7267): p. 1071-1078.
61. Milholland, B., et al., *Differences between germline and somatic mutation rates in humans and mice.* Nature Communications, 2017. **8**(1): p. 15183.
62. Rahman, N., *Realizing the promise of cancer predisposition genes.* Nature, 2014. **505**(7483): p. 302-308.
63. Knudson, A.G., *Mutation and cancer: statistical study of retinoblastoma.* Proceedings of the National Academy of Sciences, 1971. **68**(4): p. 820-823.

64. Wang, Q., *Cancer predisposition genes: molecular mechanisms and clinical impact on personalized cancer care: examples of Lynch and HBOC syndromes.* Acta pharmacologica Sinica, 2016. **37**(2): p. 143-149.
65. Ford, D., et al., *Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium.* American Journal of Human Genetics, 1998. **62**(3): p. 676-689.
66. Tomasetti, C., B. Vogelstein, and G. Parmigiani, *Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation.* Proceedings of the National Academy of Sciences, 2013. **110**(6): p. 1999.
67. Vogelstein, B., et al., *Cancer Genome Landscapes.* Science, 2013. **339**(6127): p. 1546-1558.
68. McLaren, W., et al., *Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor.* Bioinformatics, 2010. **26**(16): p. 2069-2070.
69. Zhao, M. and Z. Zhao, *Concordance of copy number loss and down-regulation of tumor suppressor genes: a pan-cancer study.* 2016. **17**(S7).
70. Sircoulomb, F., et al., *Genome profiling of ERBB2-amplified breast cancers.* BMC cancer, 2010. **10**: p. 539-539.
71. Adams, J., *Transcriptome: connecting the genome to gene function.* Nat Educ, 2008. **1**(1): p. 195.
72. Sager, R., *Expression genetics in cancer: Shifting the focus from DNA to RNA.* Proceedings of the National Academy of Sciences, 1997. **94**(3): p. 952.
73. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-140.
74. Anders, S. and W. Huber, *Differential expression analysis for sequence count data.* Genome Biol, 2010. **11**: p. R106.
75. Sorlie, T., et al., *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.* Proc Natl Acad Sci U S A, 2001. **98**(19): p. 10869-74.
76. Perou, C.M., et al., *Molecular portraits of human breast tumours.* Nature, 2000. **406**(6797): p. 747-752.
77. Sorlie, T., et al., *Repeated observation of breast tumor subtypes in independent gene expression data sets.* Proc Natl Acad Sci U S A, 2003. **100**(14): p. 8418-23.
78. Herschkowitz, J.I., et al., *Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors.* Genome Biol, 2007. **8**(5): p. R76.
79. Prat, A., et al., *Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer.* Breast Cancer Res, 2010. **12**(5): p. R68.
80. Forbes, S.A., et al., *COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer.* Nucleic acids research, 2011. **39**(Database issue): p. D945-D950.
81. Tate, J.G., et al., *COSMIC: the Catalogue Of Somatic Mutations In Cancer.* Nucleic Acids Research, 2019. **47**(D1): p. D941-D947.
82. OIKONOMOU, E. and A. PINTZAS, *Cancer genetics of sporadic colorectal cancer: BRAF and PI3KCA mutations, their impact on signaling and novel targeted therapies.* Anticancer research, 2006. **26**(2A): p. 1077-1084.

83. Samuels, Y. and V.E. Velculescu, *Oncogenic mutations of PIK3CA in human cancers.* Cell Cycle, 2004. **3**(10): p. 1221-4.
84. Xu, J., Y. Chen, and O.I. Olopade, *MYC and Breast Cancer.* Genes & Cancer, 2010. **1**(6): p. 629-640.
85. Sondka, Z., et al., *The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers.* Nature Reviews Cancer, 2018. **18**(11): p. 696-705.
86. Olivier, M., M. Hollstein, and P. Hainaut, *TP53 mutations in human cancers: origins, consequences, and clinical use.* Cold Spring Harbor perspectives in biology, 2010. **2**(1): p. a001008-a001008.
87. Heather, J.M. and B. Chain, *The sequence of sequencers: The history of sequencing DNA.* Genomics, 2016. **107**(1): p. 1-8.
88. Metzker, M.L., *Sequencing technologies — the next generation.* Nature Reviews Genetics, 2010. **11**(1): p. 31-46.
89. Goodwin, S., J.D. McPherson, and W.R. McCombie, *Coming of age: ten years of next-generation sequencing technologies.* Nature Reviews Genetics, 2016. **17**(6): p. 333-351.
90. Tan, G., et al., *Long fragments achieve lower base quality in Illumina paired-end sequencing.* Scientific Reports, 2019. **9**(1): p. 2856.
91. Belkadi, A., et al., *Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants.* Proceedings of the National Academy of Sciences, 2015. **112**(17): p. 5473.
92. Bewicke-Copley, F., et al., *Applications and analysis of targeted genomic sequencing in cancer studies.* Computational and structural biotechnology journal, 2019. **17**: p. 1348-1359.
93. Zhao, S., et al., *Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells.* PloS one, 2014. **9**(1): p. e78644.
94. Nookaew, I., et al., *A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae.* Nucleic Acids Research, 2012. **40**(20): p. 10084-10097.
95. Wang, B., et al., *Reviving the transcriptome studies: an insight into the emergence of single-molecule transcriptome sequencing.* Frontiers in Genetics, 2019. **10**: p. 384.
96. Branton, D., et al., *The potential and challenges of nanopore sequencing*, in *Nanoscience and technology: A collection of reviews from Nature Journals*. 2010, World Scientific. p. 261-268.
97. Rhoads, A. and K.F. Au, *PacBio sequencing and its applications.* Genomics, Proteomics & Bioinformatics, 2015. **13**(5): p. 278-289.
98. Zheng, G.X., et al., *Massively parallel digital transcriptional profiling of single cells.* Nature Communications, 2017. **8**(1): p. 1-12.
99. Bray, F., et al., *Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* CA Cancer J Clin, 2018. **68**(6): p. 394-424.
100. Australian Institute for Health Welfare, *Cancer in Australia 2019 (Cat No. CAN 119; Cancer Series No. 123).* 2019.
101. Sun, Y., et al., *Identification of 12 cancer types through genome deep learning.* Scientific Reports, 2019. **9**(1): p. 17256.

102. Ly, D., et al., *An international comparison of male and female breast cancer incidence rates.* International Journal of Cancer, 2013. **132**(8): p. 1918-1926.
103. Bellanger, M., et al., *Are Global Breast Cancer Incidence and Mortality Patterns Related to Country-Specific Economic Development and Prevention Strategies?* Journal of Global Oncology, 2018(4): p. 1-16.
104. Malhotra, G.K., et al., *Histological, molecular and functional subtypes of breast cancers.* Cancer Biol Ther, 2010. **10**(10): p. 955-60.
105. Weigelt, B., F.C. Geyer, and J.S. Reis-Filho, *Histological types of breast cancer: How special are they?* Molecular Oncology, 2010. **4**(3): p. 192-208.
106. Gao, J.J. and S.M. Swain, *Luminal A Breast Cancer and Molecular Assays: A Review.* The Oncologist, 2018. **23**(5): p. 556-565.
107. Makki, J., *Diversity of Breast Carcinoma: Histological Subtypes and Clinical Relevance.* Clinical Medicine Insights: Pathology, 2015. **8**: p. 23-31.
108. Feng, Y., et al., *Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis.* Genes & diseases, 2018. **5**(2): p. 77-106.
109. The Cancer Genome Atlas Research Network, *Comprehensive molecular portraits of human breast tumours.* Nature, 2012. **490**(7418): p. 61-70.
110. Testa, U., G. Castelli, and E. Pelosi, *Breast Cancer: A Molecularly Heterogenous Disease Needing Subtype-Specific Treatments.* Medical sciences (Basel, Switzerland), 2020. **8**(1): p. 18.
111. Early Breast Cancer Trialists' Collaborative Group, *Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials.* The lancet, 2011. **378**(9793): p. 771-784.
112. Tran, B. and P.L. Bedard, *Luminal-B breast cancer and novel therapeutic targets.* Breast Cancer Research, 2011. **13**(6): p. 221.
113. Yang, Y., et al., *HER2-Driven Breast Tumorigenesis Relies upon Interactions of the Estrogen Receptor with Coactivator MED1.* Cancer Research, 2018. **78**(2): p. 422-435.
114. Parker, J.S., et al., *Supervised risk predictor of breast cancer based on intrinsic subtypes.* J Clin Oncol, 2009. **27**(8): p. 1160-7.
115. Bertucci, F., et al., *How basal are triple-negative breast cancers?* International Journal of Cancer, 2008. **123**(1): p. 236-240.
116. Chen, Y. and O.I. Olopade, *MYC in breast tumor progression.* Expert Review of Anticancer Therapy, 2008. **8**(10): p. 1689-1698.
117. Masuda, H., et al., *Role of epidermal growth factor receptor in breast cancer.* Breast Cancer Research and Treatment, 2012. **136**(2): p. 331-345.
118. Kennecke, H., et al., *Metastatic Behavior of Breast Cancer Subtypes.* Journal of Clinical Oncology, 2010. **28**(20): p. 3271-3277.
119. Garrido-Castro, A.C., N.U. Lin, and K. Polyak, *Insights into Molecular Classifications of Triple-Negative Breast Cancer: Improving Patient Selection for Treatment.* Cancer discovery, 2019. **9**(2): p. 176-198.
120. Loi, S., et al., *Tumor-Infiltrating Lymphocytes and Prognosis: A Pooled Individual Patient Analysis of Early-Stage Triple-Negative Breast Cancers.* Journal of clinical oncology : official journal of the American Society of Clinical Oncology, 2019. **37**(7): p. 559-569.
121. Zaha, D.C., *Significance of immunohistochemistry in breast cancer.* World journal of clinical oncology, 2014. **5**(3): p. 382-392.

122. Goldhirsch, A., et al., *Strategies for subtypes--dealing with the diversity of breast cancer: highlights of the St. Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011.* Annals of oncology : official journal of the European Society for Medical Oncology, 2011. **22**(8): p. 1736-1747.
123. Bastien, R.R., et al., *PAM50 breast cancer subtyping by RT-qPCR and concordance with standard clinical molecular markers.* BMC Med Genomics, 2012. **5**: p. 44.
124. Whitworth, P., et al., *Chemosensitivity predicted by BluePrint 80-gene functional subtype and MammaPrint in the Prospective Neoadjuvant Breast Registry Symphony Trial (NBRST).* Annals of surgical oncology, 2014. **21**(10): p. 3261-3267.
125. Kim, H.K., et al., *Discordance of the PAM50 Intrinsic Subtypes Compared with Immunohistochemistry-Based Surrogate in Breast Cancer Patients: Potential Implication of Genomic Alterations of Discordance.* Cancer Research and Treatment, 2019. **51**(2): p. 737-747.
126. Kumar, N., et al., *Quantification of intrinsic subtype ambiguity in Luminal A breast cancer and its relationship to clinical outcomes.* BMC Cancer, 2019. **19**(1).
127. Allott, E.H., et al., *Intratumoral heterogeneity as a source of discordance in breast cancer biomarker classification.* Breast Cancer Research, 2016. **18**(1).
128. Martelotto, L.G., et al., *Breast cancer intra-tumor heterogeneity.* Breast Cancer Research, 2014. **16**(3): p. R48.
129. Raj-Kumar, P.-K., et al., *PCA-PAM50 improves consistency between breast cancer intrinsic and clinical subtyping reclassifying a subset of luminal A tumors as luminal B.* Scientific Reports, 2019. **9**(1).
130. Ciriello, G., et al., *Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer.* Cell, 2015. **163**(2): p. 506-19.
131. Cerami, E., et al., *The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data.* Cancer discovery, 2012. **2**(5): p. 401-404.
132. Wang, K., et al., *MapSplice: accurate mapping of RNA-seq reads for splice junction discovery.* Nucleic Acids Res, 2010. **38**(18): p. e178.
133. Li, B. and C.N. Dewey, *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.* BMC Bioinformatics, 2011. **12**(1): p. 323.
134. Curtis, C., et al., *The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.* Nature, 2012. **486**(7403): p. 346-52.
135. Bashashati, A., et al., *DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer.* Genome Biology, 2012. **13**(12): p. R124.
136. Thingholm, L.B., et al., *Strategies for Integrated Analysis of Genetic, Epigenetic, and Gene Expression Variation in Cancer: Addressing the Challenges.* Frontiers in Genetics, 2016. **7**.
137. Chatterjee, A., E.J. Rodger, and M.R. Eccles, *Epigenetic drivers of tumourigenesis and cancer metastasis.* Seminars in Cancer Biology, 2018. **51**: p. 149-159.
138. Dietlein, F., et al., *Identification of cancer driver genes based on nucleotide context.* Nature Genetics, 2020. **52**(2): p. 208-218.
139. Cheng, F., J. Zhao, and Z. Zhao, *Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes.* Briefings in Bioinformatics, 2016. **17**(4): p. 642-656.
140. Dees, N.D., et al., *MuSiC: Identifying mutational significance in cancer genomes.* Genome Research, 2012. **22**(8): p. 1589-1598.

141. Lawrence, M.S., et al., *Mutational heterogeneity in cancer and the search for new cancer-associated genes.* Nature, 2013. **499**(7457): p. 214-218.
142. Reimand, J. and G.D. Bader, *Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers.* Molecular Systems Biology, 2013. **9**(1): p. 637.
143. Gonzalez-Perez, A. and N. Lopez-Bigas, *Functional impact bias reveals cancer drivers.* Nucleic Acids Research, 2012. **40**(21): p. e169-e169.
144. Arnedo-Pac, C., et al., *OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers.* Bioinformatics, 2019.
145. Hou, J.P. and J. Ma, *DawnRank: discovering personalized driver genes in cancer.* Genome Medicine, 2014. **6**(7).
146. Bertrand, D., et al., *Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles.* Nucleic Acids Research, 2015. **43**(7): p. e44-e44.
147. Tokheim, C.J., et al., *Evaluating the evaluation of cancer driver genes.* Proceedings of the National Academy of Sciences, 2016. **113**(50): p. 14330-14335.
148. Luo, P., et al., *deepDriver: Predicting Cancer Driver Genes Based on Somatic Mutations Using Deep Convolutional Neural Networks.* Frontiers in Genetics, 2019. **10**.
149. Tamborero, D., A. Gonzalez-Perez, and N. Lopez-Bigas, *OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes.* Bioinformatics, 2013. **29**(18): p. 2238-2244.
150. Mularoni, L., et al., *OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations.* Genome Biology, 2016. **17**(1).
151. Ng, S., et al., *PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis.* Bioinformatics, 2012. **28**(18): p. i640-i646.
152. Wong, W.C., et al., *CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer.* Bioinformatics, 2011. **27**(15): p. 2147-2148.
153. Mao, Y., et al., *CanDrA: Cancer-Specific Driver Missense Mutation Annotation with Optimized Features.* PLoS ONE, 2013. **8**(10): p. e77945.
154. Vastrik, I., et al., *Reactome: a knowledge base of biologic pathways and processes.* Genome Biology, 2007. **8**(3): p. R39.
155. Kanehisa, M., et al., *KEGG for integration and interpretation of large-scale molecular data sets.* Nucleic acids research, 2012. **40**(Database issue): p. D109-D114.
156. Rentzsch, P., et al., *CADD: predicting the deleteriousness of variants throughout the human genome.* Nucleic Acids Research, 2019. **47**(D1): p. D886-D894.
157. Banerji, S., et al., *Sequence analysis of mutations and translocations across breast cancer subtypes.* Nature, 2012. **486**(7403): p. 405-409.
158. Berx, G., et al., *Mutations of the human E-cadherin (CDH1) gene.* Human Mutation, 1998. **12**(4): p. 226-237.
159. Tamborero, D., et al., *Comprehensive identification of mutational cancer driver genes across 12 tumor types.* Scientific Reports, 2013. **3**(1): p. 2650.
160. Balko, J.M., et al., *Molecular Profiling of the Residual Disease of Triple-Negative Breast Cancers after Neoadjuvant Chemotherapy Identifies Actionable Therapeutic Targets.* Cancer Discovery, 2014. **4**(2): p. 232-245.
161. Shao, X., et al., *Copy number variation is highly correlated with differential gene expression: a pan-cancer study.* BMC Medical Genetics, 2019. **20**(1).

162. Zhang, Y., et al., *Elevated Aurora B expression contributes to chemoresistance and poor prognosis in breast cancer.* International journal of clinical and experimental pathology, 2015. **8**(1): p. 751-757.
163. Campbell, P.J., et al., *Pan-cancer analysis of whole genomes.* Nature, 2020. **578**(7793): p. 82-93.
164. Yoon, N.K., et al., *Higher levels of GATA3 predict better survival in women with breast cancer.* Human pathology, 2010. **41**(12): p. 1794-1801.
165. Engel, C., et al., *Prevalence of pathogenic BRCA1/2 germline mutations among 802 women with unilateral triple-negative breast cancer without family cancer history.* BMC Cancer, 2018. **18**(1): p. 265.
166. Shrestha, Y., et al., *PAK1 is a breast cancer oncogene that coordinately activates MAPK and MET signaling.* Oncogene, 2012. **31**(29): p. 3397-3408.
167. Lundberg, A., et al., *The long-term prognostic and predictive capacity of cyclin D1 gene amplification in 2305 breast tumours.* Breast Cancer Research, 2019. **21**(1): p. 34.
168. Balko, J.M., et al., *Triple-negative breast cancers with amplification of JAK2 at the 9p24 locus demonstrate JAK2-specific dependence.* 2016. **8**(334): p. 334ra53-334ra53.
169. Sharifi-Noghabi, H., et al., *MOLI: multi-omics late integration with deep neural networks for drug response prediction.* Bioinformatics, 2019. **35**(14): p. i501-i509.
170. Rappoport, N. and R. Shamir, *Multi-omic and multi-view clustering algorithms: review and cancer benchmark.* Nucleic acids research, 2018. **46**(20): p. 10546-10562.
171. Zitnik, M., et al., *Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities.* An international journal on information fusion, 2019. **50**: p. 71-91.
172. Kelly, C.J., et al., *Key challenges for delivering clinical impact with artificial intelligence.* BMC Medicine, 2019. **17**(1): p. 195.
173. Holzinger, A., et al., *What do we need to build explainable AI systems for the medical domain?* arXiv preprint arXiv:1712.09923, 2017.
174. Lundberg, S.M. and S.-I. Lee. *A unified approach to interpreting model predictions*. in *Advances in neural information processing systems*. 2017.
175. Ribeiro, M.T., S. Singh, and C. Guestrin. *" Why should I trust you?" Explaining the predictions of any classifier*. in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016.
176. Samek, W., T. Wiegand, and K.-R. Müller, *Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models.* arXiv preprint arXiv:1708.08296, 2017.
177. Leong, H.S., et al., *Efficient molecular subtype classification of high-grade serous ovarian cancer.* The Journal of pathology, 2015. **236**(3): p. 272-277.
178. Verhaak, R.G., et al., *Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling.* haematologica, 2009. **94**(1): p. 131-134.
179. West, L., et al., *A novel classification of lung cancer into molecular subtypes.* PloS one, 2012. **7**(2): p. e31906.
180. Mrozek, K., et al., *Clinical relevance of mutations and gene-expression changes in adult acute myeloid leukemia with normal cytogenetics: are we ready for a prognostically prioritized molecular classification?* Blood, 2007. **109**(2): p. 431-448.
181. Hoadley, K.A., et al., *Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin.* Cell, 2014. **158**(4): p. 929-944.

182. Hoadley, K.A., et al., *Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer.* Cell, 2018. **173**(2): p. 291-304.e6.
183. Pavlidis, N., H. Khaled, and R. Gaafar, *A mini review on cancer of unknown primary site: A clinical puzzle for the oncologists.* Journal of advanced research, 2015. **6**(3): p. 375-382.
184. Tothill, R.W., et al., *Massively-parallel sequencing assists the diagnosis and guided treatment of cancers of unknown primary.* J Pathol, 2013. **231**(4): p. 413-23.
185. Saltz, J., et al., *Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images.* Cell reports, 2018. **23**(1): p. 181-193.e7.
186. Coudray, N., et al., *Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning.* Nature Medicine, 2018. **24**(10): p. 1559-1567.
187. Ehteshami Bejnordi, B., et al., *Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer.* Jama, 2017. **318**(22): p. 2199-2210.

Appendix

Complete list of predicted driver genes by algorithms described in Chapter 3

1. DriverNet

Table A.1 – DriverNet's predicted driver genes (datasets consist of all 'omics data, including point mutation, copy number, and gene expression profiles)

TCGA (SNP + CNV + EXPR)					METABRIC (SNP + CNV + EXPR)				
All subtypes	Basal	Her2	LumA	LumB	All subtypes	Basal	Her2	LumA	LumB
PRDM7	ZNF124	ARRB2	PRDM7	ASH2L	ARRB2	ARRB2	ARRB2	ARRB2	ARRB2
ARRB2	CTNNB1	ZFP3	ARRB2	ARRB2	ZFP90	CTNNB1	ASH2L	ZFP90	ASH2L
CREBBP	ARRB2	GRB2	CREBBP	CTNNB1	TP53	ASH2L	TP53	CREBBP	CBL
AURKB	AKT1	CTNNB1	CDH1	GRB2	EP300	TP53	AURKB	CDH1	MYC
GRB2	ACTB	TP53	AURKB	EP300	ACTA1	AKT1	CRK	ACTN2	EP300
ACTB	TP53	ACTG1	TP53	ACTG1	AURKB	ACTB	EP300	TP53	ACTA1
TP53	GRB2	CREBBP	ACTN2	TP53	MYC	MYC	ACTA1	PIK3CA	AURKB
AKT1	MYC	AURKB	GRB2	AURKB	AKT1	AURKB	MYC	AURKB	TP53
MYC	AURKB	MYC	DHX38	AKT1	CDH1	GRB2	GRB2	DHX38	CDH1
CPSF1	CDK7	CBL	RPL13	MYC	CBL	CDK7	CDH1	GNAO1	GRB2
CDH1	EP300	JUN	MYC	CPSF1	ACTN2	EP300	AR	RPL13	GNB1
RPL13	EEF2	EFTUD2	EP300	CBL	CREBBP	EIF1AX	RPL13	ACTA1	ACTN2
CDC42	CPSF1	CDC42	CBL	CDC42	CPSF1	PIK3R1	CPSF1	EP300	CPSF1
CDK7	CDC42	LHX1	GNAO1	EEF1A1	GRB2	CPSF1	ACTN2	CBL	AKT1
EP300	CD247	AKT1	ACTA1	JUN	ASH2L	ACTA1	AKT1	MAPK3	ZFP90
ASH2L	ACTN2	BRCA1	AKT1	GNB1	RPL13	GNAI2	ZFP90	MYC	CREBBP
CBL	HSP90AA1	RPL12	MAPK3	CREBBP	GNAO1	ZNF124	DLG4	PLK1	EEF1A1
ACTN2	HDAC1	ACTN2	CDC42	ACTN2	PIK3CA	FOS	CBL	ARNT	FYN
HDAC1	GNAI2	HDAC1	ARF1	MAPK1	CDK7	AR	PIK3CA	AKT1	JUN
ARF1	JUN	GNAI3	ZFP3	HDAC1	CTNNB1	CALM1	CDK7	CD247	GNAO1
GNAO1	LHX1	DLG4	PLK1	ZNF124	ARNT	ACTN2	CTNNB1	ARF1	CDC42
CD247	CALM1	CREB1	ARNT	CDH1	ARF1	CDC42	CREBBP	BCAR1	ARNT
CTNNB1	ARNT	FYN	PIK3CA	CDK7	AR	ARNT	ACTG1	GNG4	ACTG1
HSP90AA1	CASP3	ACTR1A	HDAC1	ARF1	GNB1	GNB1	GNB1	GRB2	ARF1
ESR1	ABL1	ESR1	ACTB	ESR1	FYN	ARF1	BRCA1	ZNF124	HDAC1
ARNT	ARF1	CDH1	CDK7	FYN	CDC42	CASP3	ARF1	AKT3	ESR1
FYN		ABL1	CD247		JUN	CREBBP	MAPK1	POU2F1	CTNNB1
BRCA1			ESR1		DLG4	EGFR	ESR1	GNAL	CEP164
JUN			BCAR1		ACTG1	CD247	FYN	FYN	PIK3CA
CASP3			JUN		PLK1	DLG4	SMAD4	DLG4	DLG4
			BRCA1		ESR1	CDH1		TRADD	ETS1
			GNB1		CRK	ESR1		ESR1	CDK7
			DLG4		DHX38				RPL13
			NDUFAB1		CD247				CRK
					HDAC1				AR
					MAPK1				
					EIF3E				
					CALM1				
					GNG4				
					BRCA1				
					EGFR				
					ZNF124				
					HSP90AA1				

Table A.2 – DriverNet's predicted driver genes (datasets include only point mutations and gene expression profiles)

TCGA (SNP + EXPR)							METABRIC (SNP + EXPR)				
All subtypes		Basal	Her2	LumA		LumB	All subtypes	Basal	Her2	LumA	LumB
TP53	AKAP9	TP53	TP53	PIK3CA	MAP2K4	TP53	TP53	TP53	TP53	PIK3CA	PIK3CA
PIK3CA	ZNF121	ZIM2	PIK3CA	CDH1	ZNF595	PIK3CA	PIK3CA	PIK3CA	PIK3CA	TP53	TP53
CDH1	RPL13A	CREBBP	ZNF17	TP53	SMAD4	ZNF135	CDH1	EGFR	AKT1	CDH1	CDH1
AKT1	EEF1A1	ZIM3	ERBB3	AKT1	EIF3A	CDH1	AKT1	APC	EP300	AKT1	EP300
CREBBP	GATA3	ZNF175	ZNF226	MAP3K1	CDH10	AKT1	EP300	PIK3R1	ERBB2	EP300	AKT1
ZNF135	APC	ZNF100	ZNF157	LHX1	ZNF180	EIF5B	EGFR	ERBB3	CDH1	MAP3K1	MAP3K1
EGFR	PRDM7	BRCA1	EGFR	ZFP30	CDKN1B	EGFR	MAP3K1	AKAP9	SMAD4	EGFR	COL6A3
ACTB	FLNA	ZNF251	JAK2	ZNF138	ZNF583	ZNF28	AKAP9	EP300	COL6A3	NCOR2	AKAP9
EP300	GNAI2	PIK3CA	ZFP37	EP300	ZNF417	ZNF354B	COL6A3	NCOR2	AKAP9	COL6A3	ERBB3
ZNF100	ZNF540	EEF1A1	ZIM3	NCOR1	HKR1	ZNF43	APC	COL6A3		GATA3	NCOR2
BRCA1	ERBB2	EGFR	ZNF107	PIK3R1	HSP90AA1	AHCTF1	ERBB3	NOTCH1			APC
PIK3R1	ZNF184	EP300	RPL13A	ZNF431	ACTB	GATA3	NCOR2	RB1			EGFR
MAP3K1	FLNB	ZFP30	PIK3R1	ZNF121	ACTG1	ZNF33A	ERBB2	ARID1A			ARID1A
CPSF1	CEP290	PRDM7	ZNF160	ERBB3	ZNF492	ZNF41	ARID1A	ERBB2			NOTCH1
ZFP30	CDH23	PIK3R1	ZNF208	ZNF248	DHX9	ZNF134	NOTCH1	AKT1			ATR
ZIM3	CBL	HDAC1	ZNF266	ESR1	ZNF250	EEF2	PIK3R1	BRCA1			NCOR1
ERBB3	ARID1A	ZNF224	ERBB2	ZFP90	EIF3E	ESR1	ATR				GATA3
ZIM2	COL4A3	CPSF1	SMAD4	PRDM9	HDAC1	DMD	RB1				
AHCTF1	ATM	ZNF425	CENPE	ZNF135	COL7A1	AURKA	NCOR1				
ZNF208	ZNF175	ACTB	CDH1	CPSF1	CEP290	ZNF425	LAMA2				
HDAC1	PRDM9	ZNF234	ZFP28	FOXA1	ZNF420	AR					
ZNF251	EEF2	RB1	ZNF551	GATA3	CDK2	CTNNB1					
NCOR1	LAMA1	HNRNPU	CDKN1A	ZIM3	COL4A3	NCOR1					
LHX1	ZFP90	ASH2L	PLK1	ANAPC5	CD247	GNG2					
ZNF28	ANAPC5	ZNF12	FLNB	CREBBP	AR	ZFP28					
ZNF41	NFKB1	CDH23	EP300	ERBB2	APC	ALMS1					
ZNF425	CASP8	ITGAV	POLR2B	ZNF28	CTNNB1	ZNF35					
CTNNB1	GLI3	ZNF208	CREBBP	ZNF275	ATM	CREBBP					
ZNF138	CDH10	CTNNB1	ALMS1	RBAK	CDK5RAP2	RB1					
RB1	ITGAV	ZNF211	ZNF587	ZNF20	CENPE	RHOA					
CENPE	EIF3A	ZNF239	GLI3	ARID1A	ZNF287	COL5A3					
ALMS1	ZNF546	GNAI2	ZFP30	ZNF100	EGFR	ZNF442					
ZNF275	RELN	ZNF133	CDH12	ZNF208	PLCG1	ZNF473					
RBAK	ZNF354B	GNGT1	HDAC1	COL4A5	ITGB2	IL12RB2					
ESR1	ZNF431	ZNF157	ZNF350	DMD	ZNF155	ZFP37					
HNRNPU	ATR	PCDH10		CSNK1A1	PCDH15	CAV1					
ACTN2	SMAD4	NCOR1		AHCTF1	EEF2	PIK3R1					
ZNF20	BTRC	CTNNA2				COL14A1					
AR	APOB	RBAK				COL6A3					
ZNF180	ZNF239	ZNF517				ANAPC7					
DMD	MED12	RPL13A				EP300					
ZFP28	CTNNA2	BUB1				HIF1A					
CCAR1	CDC20	HNRNPA1				ACTA2					
		ZNF415				CCAR1					
		ZNF540				PRDM9					
		CCAR1				ZFP3					
		NFKB1				ZNF160					
		COL5A2				CDH23					
		CUL5				CPSF1					
		ZNF121				RELN					
		COL4A3				ARID1A					
		ARNT				GNAI2					
		ARHGEF2				GNAS					
						MAP2K4					

Table A.3 – DriverNet's predicted driver genes (datasets include only copy number and gene expression profiles)

TCGA (CNV + EXPR)					METABRIC (CNV + EXPR)				
All subtypes	**Basal**	**Her2**	**LumA**	**LumB**	**All subtypes**	**Basal**	**Her2**	**LumA**	**LumB**
PRDM7	*ZFP2*	*ARRB2*	*ZFP1*	*PRDM7*	*ARRB2*	*ARRB2*	*ARRB2*	*ARRB2*	*ARRB2*
ARRB2	*CTNNB1*	*ZFP3*	*ARRB2*	*ARRB2*	*ZFP90*	*CTNNB1*	*TP53*	*ZFP90*	*ZFP90*
EP300	*ARRB2*	*CTNNB1*	*CREBBP*	*CTNNB1*	*TP53*	*ZNF141*	*ZFP90*	*CDH1*	*CBL*
AURKB	*AKT1*	*CRK*	*CDH1*	*GRB2*	*AURKB*	*AKT1*	*AURKB*	*CREBBP*	*EP300*
ACTB	*ACTB*	*TP53*	*AURKB*	*EP300*	*CDH1*	*ACTB*	*CRK*	*BCAR1*	*TP53*
GRB2	*GRB2*	*CREBBP*	*TP53*	*ACTG1*	*EP300*	*TP53*	*EP300*	*AURKB*	*AURKB*
TP53	*EP300*	*AURKB*	*CRK*	*TP53*	*CRK*	*AURKB*	*ACTG1*	*TP53*	*CDH1*
AKT1	*TP53*	*ACTG1*	*DHX38*	*AURKB*	*AKT1*	*CDK7*	*CDH1*	*DHX38*	*ACTG1*
CTNNB1	*AURKB*	*GRB2*	*RPL13*	*AKT1*	*CREBBP*	*PIK3R1*	*LHX1*	*GNAO1*	*AKT1*
DHX38	*CDK7*	*HDAC1*	*EP300*	*CBL*	*DHX38*	*EP300*	*AR*	*EP300*	*GNB1*
CDK7	*EEF2*	*EFTUD2*	*ACTB*	*CDC42*	*CBL*	*EIF1AX*	*RPL13*	*RPL13*	*FYN*
CDC42	*MAX*	*CDC42*	*MYC*	*DHX38*	*ACTG1*	*GRB2*	*POLR2A*	*CBL*	*DHX38*
CREBBP	*CDC42*	*CBL*	*CBL*	*JUN*	*RPL13*	*MAX*	*CBL*	*ACTN2*	*JUN*
RPL13	*CPSF2*	*BRCA1*	*GNAO1*	*DNAJC3*	*GNAO1*	*ACTC1*	*GRB2*	*AKT1*	*EEF1A1*
ZFP3	*JUN*	*LHX1*	*ACTA1*	*MYC*	*CTNNB1*	*GNAI2*	*DLG4*	*PLK1*	*CREBBP*
CBL	*HSP90AA1*	*MYC*	*AKT1*	*ZFP3*	*CDK7*	*FOS*	*CREBBP*	*GRB2*	*CDC42*
MYC		*AKT1*	*GRB2*	*GNAI3*	*ASH2L*	*LHX1*	*AKT1*	*ACTA1*	*GNAO1*
CDH1		*RPL12*	*CDC42*	*HDAC1*	*AR*	*AR*	*CTNNB1*	*MAPK3*	*CRK*
HDAC1		*GNAI3*	*ZFP3*		*GRB2*	*CALM1*	*BRCA1*	*FYN*	*ASH2L*
ACTN2			*HDAC1*		*ACTN2*	*CDC42*	*MAPK1*	*ACTB*	*CTNNB1*
GNAO1			*PLK1*		*GNB1*	*CASP3*	*GNB1*	*ESR1*	*ACTA1*
HSP90AA1			*ARF1*		*MAX*		*CDK7*	*MYC*	*ETS1*
FYN			*MAPK3*		*CDC42*		*ACTN2*	*TRADD*	*GRB2*
ARF1			*ACTN2*		*JUN*		*FYN*	*GNAL*	*MYC*
			ARNT		*FYN*		*SMAD4*	*ARNT*	*HDAC1*
					DLG4				*CEP164*
					ACTB				*ESR1*
					ESR1				*DLG4*
									CDK7

2. DawnRank

Table A.4 – DawnRank's predicted driver genes (datasets consist of all 'omics data, including point mutation, copy number, and gene expression profiles)

TCGA (SNP + CNV + EXPR)							METABRIC (SNP + CNV + EXPR)				
All subtypes	Basal		Her2	LumA	LumB		All subtypes	Basal	Her2	LumA	LumB
TP53	TP53	HRAS	TP53	CDH1	TP53	PAFAH1B1	TP53	TP53	TP53	CDH1	TP53
ARF1	MYC	AURKB	ARRB2	BCAR1	MYC	MAX	ARF1	MYC	ARRB2	BCAR1	CBL
ARNT	ARNT	ETS1	BRCA1	ARF1	ARF1	ESR1	CDH1	PIK3R1	DLG4	E2F4	MYC
MYC	PIK3R1	CEBPA	CRK	ARNT	ARRB2	MAPK3	BCAR1	CDK7	ERBB2	ARF1	ETS1
ARRB2	CDK7	GNB1	DLG4	E2F4	ARNT	FOS	ARRB2	APC	CRK	GNAO1	CDH1
BCAR1	ARRB2	MAPK14	ERBB2	CREBBP	CRK	PAK1	POU2F1	CCNB1	NCOR1	ARNT	BCAR1
CDH1	MAX	PPP2CA	ARF1	CD247	CBL	FYN	ARNT	ARNT	AURKB	POU2F1	ARF1
POU2F1	ARF1	CDC42	MYC	POU2F1	POU2F1	HSP90AA1	DLG4	POU2F1	MYC	CD247	ARRB2
CD247	HIF1A	TRAF2	GRB2	TP53	CEBPB	CRKL	MYC	RASA1	BRCA1	PIK3CA	DLG4
CREBBP	FOS	E2F1	AURKB	MAPK3	EP300	HDAC2	CD247	CCNH	DVL2	SHC1	CRK
CRK	AKT1	CEBPB	ARNT	GNAO1	BRCA1	HIF1A	CRK	MAX	PAFAH1B1	CREBBP	E2F4
E2F4	APC	SRC	GNAL	ARRB2	BCAR1	GNAI3	E2F4	FOS	MAP2K4	ACTA1	CD3E
EP300	POU2F1	NFKBIA	PAFAH1B1	AXIN1	GRB2	SMAD4	SHC1	ARRB2	ARF1	TP53	POU2F1
DLG4	HSP90AA1	SP1	STAT3	SHC1	ETS1	MAP2K4	PIK3CA	ARF1	STAT3	AXIN1	CD3D
CBL	RB1	EGF	NCOR1	EP300	DLG4	CSNK2A1	NCOR1	HIF1A	PIK3CA	MAPK3	ARNT
SHC1	CALM1	ESR1	POU2F1	CRK	CDH1	PLCG1	AURKB	CD247	RB1	PTPRC	NCOR1
BRCA1	CCNH	CSNK2A1	CBL	PIK3CA	CD247	JAK1	GNAO1	CALM1	JAK2	ARRB2	CD247
ETS1	CASP3	CTBP1	CD247	PLK1	NFATC2	CALM1	ACTA1	DLG4	POU2F1	TRADD	GNAO1
AURKB	CD247	RHOA	JAK2	MYC	CDC42	GNAL	DVL2	SHC1	RARA	DLG4	AURKB
MAPK3	SMAD3	ARF6	MAP2K4	CBL	MAPK1	MYB	CBL	AURKB	ITGB3	PLK1	DVL2
GRB2	GRB2	INS	PIK3CA	DLG4	CREBBP	CCND1	PAFAH1B1	EGR1	GRB2	CRK	CD3G
PIK3CA	CRK	CTNNA1	RB1	RNPS1	AURKB	ERBB2	MAP2K4	RB1	SMAD4	RNPS1	SHC1
RB1	CCNB1	B2M	DVL2	MAPK1	AURKA	AXIN1	ETS1	GNAI2	CLTC	EP300	EP300
MAPK1	EGR1	HDAC1	HDAC1	ETS1	SRC	ARRB1	EP300	CASP3	CD247	ACTN2	PAFAH1B1
PAFAH1B1	CD4	JUN	NGFR	AURKB	JUN	PLK1	RB1	PPP2CA	CDH1	NCOR1	RB1
CEBPB	EP300	NCOA2	ESR1	ACTA1	SHC1	LCK	CREBBP	CRK	LYN	FCER1G	MAP2K4
AXIN1	RASA1	MYB	ITGB3	PDPK1	E2F1	DVL2	ERBB2	RHOA	ARNT	PDPK1	ACTA1
NCOR1	BRCA1	BCL2	ETS1	PTPRC	RB1	STAT3	BRCA1	B2M	BCAR1	AURKB	LYN
CDC42	ABL1	ERBB3	SHC1	NCOR1	NCOA2	BCL2	PTPRC	NCOR1	NGFR	DVL2	NCOA2
LYN	SMAD4	CDKN1A	CLTC	PAFAH1B1	AKT1	FOXA2	AXIN1	CTNNA1	CBL	PAFAH1B1	PAK1
PLK1	PIK3CA	CDK2	GNB1	CRKL	E2F4	PIK3CA	LYN	AR	NCOA2	CBL	CCND1
ESR1	SHC1	PAFAH1B1	CREBBP	BRCA1	GNB1	CLTC	NCOA2	ERBB2	ACTA1	MAP2K4	BIRC3
NCOA2	CREBBP	GNGT1	FYN	RB1	LYN	NFKB1		HSP90AA1	SHC1	MYC	PTPRC
GNAO1	CTNNB1	HDAC2	SMAD4	MAP2K4	NCOR1	NGFR		HDAC3	BCL2	AKT3	ARRB1
SRC	JAK2	CUL1	NFKB1	DVL2	JAK2	HRAS		AKT1	MAP2K3	MAPK1	CLTC
SMAD4	GNAL	FYN	BCAR1	TRADD	HDAC1	CD3E		ACTN1	CDK7	ETS1	CEBPB
E2F1	NFKB1	BCAR1	HDAC2					DVL2	CDKN2A		GRB2
GNAL	HDAC3	ITGB1	LYN					PIK3CA	PIK3R1		BRCA1
AKT1	CBL	GNAI2	HSP90AA1					CTBP1	E2F4		SDC2
NFATC2	LYN	MAP3K1	CDC42					MAP3K1	ETS1		FYN
FYN	DLG4	MAPK1	CASP3					BRCA1	SDC2		FOXO1
MAP2K4			CEBPB					PAFAH1B1	MED1		
GNB1			ABL1					ACTA1	SMAD2		
ACTA1			PAK1					MAP2K4	ESR1		
HDAC2			JUN					CTNNB1			
MAX			MAPK3					BTK			
AURKA			CDH1					CDC42			
DVL2			AURKA					NFIC			
FOS			MYB					JAK2			
			GNAI3					RBBP7			
			APP					NFKBIA			
			E2F4					SMAD3			
			FOS					CD4			
			PIK3R1					IL2RG			
			EP300					ETS1			
			BCL2					ESR1			
			CSNK1D					CDH1			
			NCOA2					BCAR1			
			CTNNB1								
			AKT1								
			CSNK2A1								

Table A.5 – DawnRank's predicted driver genes (datasets include only point mutations and gene expression profiles)

TCGA (SNP + EXPR)					METABRIC (SNP + EXPR)				
All subtypes	**Basal**	**Her2**	**LumA**	**LumB**	**All subtypes**	**Basal**	**Her2**	**LumA**	**LumB**
TP53	*TP53*	*TP53*	*PIK3CA*	*TP53*	*PIK3CA*	*TP53*	*TP53*	*PIK3CA*	*PIK3CA*
PIK3CA	*BRCA1*	*PIK3CA*	*CDH1*	*PIK3CA*	*TP53*		*PIK3CA*		*TP53*
CDH1	*CREBBP*	*ERBB3*	*MAP3K1*	*CDH1*					
MAP3K1	*PIK3CA*	*HSP90AA1*	*TP53*	*NCOR1*					
NCOR1	*RB1*	*ERBB2*	*NCOR1*	*DMD*					
PIK3R1	*RELN*	*RB1*	*MAP2K4*	*MAP2K4*					
MAP2K4	*EP300*	*EGFR*	*AKT1*	*ATM*					
BRCA1	*PIK3R1*	*STAT4*	*FOXA1*	*RB1*					
ERBB2	*CASP8*	*JAK2*	*ERBB2*	*EGFR*					
AKT1	*SP1*	*PIK3R1*	*PIK3R1*	*NFATC2*					
RB1	*NCOR1*	*TYK2*	*CREBBP*	*AKT1*					
CREBBP		*DMD*	*BRCA1*	*JAK1*					
ERBB3			*ATM*	*MAP3K1*					
EGFR			*APC*	*RELN*					
ATM			*ESR1*						
RELN									
FOXA1									
DMD									
EP300									
TRRAP									
APC									

Table A.6 – DawnRank's predicted driver genes (datasets include only copy number and gene expression profiles)

TCGA (CNV + EXPR)							METABRIC (CNV + EXPR)				
All subtypes	Basal		Her2	LumA	LumB		All subtypes	Basal	Her2	LumA	LumB
ARF1	MYC	PIK3CA	TP53	BCAR1	MYC	MAX	ARF1	MYC	TP53	CDH1	CBL
ARNT	ARNT	ETS1	ARRB2	CDH1	TP53	MAPK3	TP53	PIK3R1	ARRB2	BCAR1	MYC
MYC	PIK3R1	GNB1	BRCA1	ARF1	ARF1	FOS	CDH1	CDK7	DLG4	E2F4	TP53
TP53	TP53	MAPK14	CRK	ARNT	ARRB2	ESR1	BCAR1	TP53	ERBB2	ARF1	ETS1
ARRB2	CDK7	CBL	DLG4	E2F4	ARNT	FYN	ARRB2	APC	CRK	GNAO1	BCAR1
POU2F1	MAX	TRAF2	ERBB2	CREBBP	CRK	PAK1	POU2F1	CCNB1	NCOR1	ARNT	CDH1
BCAR1	ARRB2	CDC42	ARF1	CD247	CBL	CRKL	ARNT	ARNT	AURKB	POU2F1	ARF1
CDH1	ARF1	E2F1	MYC	POU2F1	CEBPB	HDAC2	DLG4	POU2F1	MYC	CD247	ARRB2
CD247	HIF1A	HRAS	GRB2	MAPK3	POU2F1	HSP90AA1	MYC	RASA1	BRCA1	SHC1	DLG4
CRK	FOS	PPP2CA	AURKB	TP53	BRCA1	GNAI3	CD247	CCNH	DVL2	CREBBP	CRK
CREBBP	AKT1	CEBPB	ARNT	GNAO1	BCAR1	SMAD4	CRK	MAX	PAFAH1B1	ACTA1	E2F4
E2F4	APC	SRC	GNAL	ARRB2	EP300	CSNK2A1	E2F4	FOS	MAP2K4	AXIN1	CD3E
DLG4	POU2F1	NFKBIA	PAFAH1B1	AXIN1	GRB2	HIF1A	SHC1	ARRB2	ARF1	MAPK3	POU2F1
CBL	HSP90AA1	EGF	STAT3	SHC1	DLG4	MAP2K4	AURKB	ARF1	STAT3	TP53	CD3D
EP300	RB1	ESR1	NCOR1	EP300	CDH1	PLCG1	GNAO1	HIF1A	JAK2	PTPRC	ARNT
SHC1	CASP3	CREBBP	POU2F1	CRK	CD247	CALM1	NCOR1	CD247	POU2F1	ARRB2	CD247
AURKB	CD247	CTBP1	CBL	MYC	ETS1	GNAL	ACTA1	CALM1	RARA	TRADD	NCOR1
BRCA1	CALM1	CSNK2A1	CD247	PLK1	CDC42	JAK1	DVL2	DLG4	ITGB3	DLG4	GNAO1
ETS1	CCNH	RHOA	MAP2K4	CBL	MAPK1	ERBB2	CBL	SHC1	GRB2	PLK1	AURKB
MAPK3	SMAD3	ARF6	DVL2	DLG4	AURKB	CCND1	PAFAH1B1	AURKB	RB1	CRK	DVL2
GRB2	GRB2	INS	JAK2	RNPS1	CREBBP	MYB	MAP2K4	EGR1	SMAD4	RNPS1	CD3G
RB1	CCNB1	SP1	NGFR	MAPK1	NFATC2	AXIN1	ETS1	GNAI2	CLTC	ACTN2	SHC1
MAPK1	CRK	B2M	ITGB3	ETS1	SRC	PLK1	CREBBP	CASP3	CD247	FCER1G	PAFAH1B1
PAFAH1B1	EGR1	JUN	CLTC	AURKB	AURKA	ARRB1	RB1	PPP2CA	CDH1	PDPK1	MAP2K4
CEBPB	CD4	CTNNA1	GNB1	ACTA1	JUN	LCK	PTPRC	CRK	LYN	AURKB	ACTA1
AXIN1	RASA1	NCOA2	FYN	PDPK1	SHC1	DVL2	EP300	RB1	ARNT	DVL2	RB1
NCOR1	ABL1	MYB	SHC1	PTPRC	RB1	BCL2	ERBB2	RHOA	BCAR1	NCOR1	LYN
CDC42	SMAD4	BCL2	RB1	PAFAH1B1	NCOA2	FOXA2	BRCA1	B2M	NGFR	EP300	NCOA2
LYN	BRCA1	CDKN1A	CREBBP	NCOR1	GNB1	STAT3	AXIN1	CTNNA1	CBL	PAFAH1B1	EP300
PLK1	SHC1	CDK2	ETS1	CRKL	E2F4	CLTC	LYN	AR	NCOA2	CBL	PAK1
GNAO1	CTNNB1	ERBB3	SMAD4	RB1	LYN	NFKB1	NCOA2	HSP90AA1	ACTA1	MAP2K4	CCND1
SRC	EP300	PAFAH1B1	HDAC1	MAP2K4	E2F1	NGFR		HDAC3	SHC1	MYC	BIRC3
NCOA2	JAK2	HDAC1	ESR1	BRCA1	JAK2	HRAS		NCOR1	BCL2	MAPK1	PTPRC
SMAD4	GNAL	HDAC2	HDAC2	DVL2	HDAC1	CD3E		AKT1	MAP2K3	AKT3	ARRB1
GNAL	NFKB1	FYN	NFKB1	TRADD	NCOR1	NFKBIA		ERBB2	CDK7	ETS1	CLTC
FYN	HDAC3	GNGT1	BCAR1	GRB2	PAFAH1B1	ABL1		ACTN1	E2F4		CEBPB
E2F1	LYN	CUL1	CDC42		AKT1			DVL2	ETS1		GRB2
GNB1	DLG4	ITGB1	CASP3					CTBP1	SDC2		SDC2
ESR1	AURKB	MAPK1	CEBPB					PAFAH1B1	MED1		FYN
MAP2K4	CEBPA	GNAI2	MAPK3					ACTA1	SMAD2		
NFATC2			LYN					BRCA1	CDKN2A		
ACTA1			CDH1					MAP2K4	ESR1		
HDAC2			AURKA					CTNNB1			
MAX			ABL1					MAP3K1			
AURKA			MYB					BTK			
AKT1			GNAI3					CDC42			
DVL2			PAK1					NFIC			
FOS			E2F4					JAK2			
			JUN					RBBP7			
			FOS					NFKBIA			
			BCL2					SMAD3			
			CSNK1D					CD4			
			APP					IL2RG			
			AKT1					ETS1			
			EP300					ESR1			
			HSP90AA1					BCAR1			

3. OncodriveFML

Table A.7 – OncodriveFML's predicted driver genes (datasets include only point mutations data)

TCGA (SNP Only)						METABRIC (SNP Only)				
All subtypes		Basal	Her2	LumA	LumB	All subtypes	Basal	Her2	LumA	LumB
SMAD4	RPS6KA3	PTEN	TP53	ARID1A	PTEN	CDKN1B	PIK3CA	SMAD4	CBFB	PTEN
GATA3	EIF1AX	TP53	PIK3CA	CBX1	PIK3CA	NF1	TP53	PIK3CA	ERBB2	TBX3
FBXW7	GRPEL2	RB1	PTEN	PTEN	TP53	ERBB2	PTEN	TP53	MAP2K4	MAP3K1
RB1	G3BP1	INO80D		MAP3K1	RB1	PIK3CA	RB1	PIK3R1	PTEN	PIK3CA
PLCXD3	OR51E2	FBXW7		CBFB	RGS4	PIK3R1		PTEN	GATA3	MLL3
AKT1	EEF1A1	GRPEL2		ARHGAP11A	MLL3	MLL3		MAP2K4	RUNX1	FOXO3
PRRX1	ATF6			NCOR1	MAP2K4	CTCF		CBFB	AKT1	GATA3
SMAD2	RUNX2			RUNX1		PTEN		NF1	CDH1	ARID1A
USP12	THSD4			GATA3		NF2		RUNX1	TP53	TP53
TP53	ATF7IP			CDKN1B		CDH1		SBNO1	PIK3CA	CBFB
NHLRC2	ZXDB			MAP2K4		RB1		AFF2	MAP3K1	AKT1
NCOR1	HCFC2			FOXA1		CBFB			TBX3	SMAD4
LPP	METAP2			MLL3		RUNX1			SF3B1	CDH1
CASP8	SYT2			PIK3CA		SF3B1			MLL3	CDKN1B
DUSP16	KCNH8			AKT1		GATA3			PIK3R1	MAP2K4
CDH1	MPZL1			CDH1		TBX3			ARID1A	RUNX1
SEMA5A	SHISA4			PIK3R1		MAP2K4			MAP3K13	SF3B1
ARID1A	TWSG1			CAMK4		AKT1			ERBB4	CTCF
PGR	PKN2			TP53		SMAD4			CTCF	NF1
CBFB	TRMT5			TBX3		ARID1A			SMAD4	
RUNX1	TBL1XR1			SEMA5A		MAP3K1			KRAS	
MAP3K1	DIAPH1			CCDC88C		TP53			ERBB3	
PTCHD1	AGO1			GOLGA4		FOXO3			SIK2	
PTEN	INO80D			PGR		SBNO1			NF1	
SLFN13	ELMO1			NF1		ERBB4			CDKN1B	
ERLIN2	EIF3F			SPRR2B		ERBB3				
ASXL2	FBXL20			ZNF660		KRAS				
TBX3	LONRF2			TSGA10		NCOR1				
MLL3	MTHFSD			SMAD4		ASXL2				
KLF7	UBE2V2			ERBB2		AGTR2				
MAP2K4	GRTP1-AS1			EIF1AX		FBXW7				
NF1	ZC3H4			LONRF2		CDKN2A				
PIK3R1	HIPK2			GNA13		MAP3K13				
FOXA1	SORCS3			JMY		PRKACG				
PIK3CA	SF3B1			FBXL20		SMAD2				
JMY	MSANTD3			ZNF528		DTWD2				
CAMK4	DIEXF					AFF2				
CDKN1B	KDM5B					SIK2				
FLRT3	TMEM151B					BAP1				
ETV1	SH3GL3					SMARCC1				
LIFR	HELZ									
MR1	LNPEP									
AFF2	CBLN4									
TRAF3	NPR3									
CELF2	ZNF660									
ORC2	CCAR1									
ERBB2	RBBP9									
DCUN1D4	NAMPT									
SETD7	ARHGEF7									
ASPA	CCDC88C									
ZNF716	RAB22A									
DLGAP2	DAB1									
LZTR1	UGGT1									
MAPK8	MIB1									
MAST1	ITK									

4. OncodriveCLUSTL

Table A.8 – OncodriveCLUSTL's predicted driver genes (datasets include only point mutations data)

TCGA (SNP Only)					METABRIC (SNP Only)				
All subtypes	**Basal**	**Her2**	**LumA**	**LumB**	**All subtypes**	**Basal**	**Her2**	**LumA**	**LumB**
PIK3CA	*TP53*		*FLG*	*PIK3CA*	*PIK3CA*	*PIK3CA*	*PIK3CA*	*PIK3CA*	*PIK3CA*
AKT1	*PIK3CA*		*PIK3CA*	*OR6A2*	*AKT1*	*TP53*	*MUC16*	*AKT1*	*AKT1*
TP53	*FCGBP*		*AKT1*		*SF3B1*	*DNAH2*	*AHNAK2*	*SF3B1*	*SF3B1*
OLFML2B	*ZNF541*		*SF3B1*		*SYNE1*		*AKT1*	*AHNAK2*	*AHNAK2*
FRAS1	*LRP1B*		*SGIP1*		*DNAH2*		*TP53*	*DNAH2*	*SYNE1*
COL3A1	*ARHGAP5*		*URGCP*		*AHNAK*			*MUC16*	*TP53*
BRIP1	*KANK2*		*FRMPD4*		*TP53*			*UTRN*	*CDH1*
SF3B1	*FLG2*		*SPOP*		*MUC16*			*ERBB3*	*GATA3*
FBN3			*SRRM2*		*AHNAK2*			*ERBB2*	*JAK1*
URGCP			*UBR4*		*HERC2*			*ASXL2*	*ACVRL1*
CDH1			*HPD*		*ERBB2*			*GATA3*	*FOXO1*
MGA					*CDH1*			*CDH1*	
RET					*NCOA3*			*KRAS*	
EML6					*ERBB3*			*AHNAK*	
ASPG					*EP300*			*MAP2K4*	
KIF3C					*PDE4DIP*			*SYNE1*	
SRRM2					*GATA3*			*EP300*	
ERBB2					*COL6A3*				
PEG3					*JAK1*				
SPOP					*FOXO1*				
MICAL3					*KRAS*				
					MLL2				
					NOTCH1				
					GLDC				
					UTRN				
					MLL3				
					STAB2				
					MAGEA8				
					NCOR2				
					THADA				
					DNAH11				
					USP28				
					TG				
					DNAH5				
					RPGR				
					OR6A2				
					SHANK2				
					AKAP9				
					MAP2K4				
					LAMA2				
					ARID1B				
					MAP3K1				
					COL22A1				
					MLLT4				
					GPR124				
					ATR				
					SETD2				
					NPNT				
					ZFP36L1				
					CHD1				
					MYO3A				
					CACNA2D3				
					SBNO1				
					ASXL2				
					ROS1				
					SETD1A				
					PTPRM				
					MYH9				
					PTPRD				

5. 20/20+

Table A.9 – 20/20+'s predicted driver genes (datasets include only point mutations data)

TCGA (SNP Only)					METABRIC (SNP Only)				
All subtypes	**Basal**	**Her2**	**LumA**	**LumB**	**All subtypes**	**Basal**	**Her2**	**LumA**	**LumB**
ARID1A	*TP53*		*NCOR1*	*KMT2C*	*CDH1*	*TP53*	*TP53*	*MAP3K1*	*CDH1*
CDH1	*RB1*		*PTEN*	*RB1*	*PTEN*	*RB1*	*PTEN*	*CDH1*	*PTEN*
NF1	*KDM6A*		*CDH1*	*NCOR1*	*MAP3K1*	*BRCA1*	*RUNX1*	*MAP2K4*	*MAP3K1*
MAP3K1	*NF1*		*MAP3K1*	*CDH1*	*MAP2K4*	*PTEN*	*MAP2K4*	*KMT2C*	*KMT2C*
RB1	*BRCA1*		*KMT2C*	*PTEN*	*TP53*	*PIK3CA*	*KMT2C*	*CBFB*	*TP53*
PTEN			*NF1*		*GATA3*	*NF1*	*ARID1A*	*PTEN*	*MAP2K4*
KMT2C			*ARID1A*		*KMT2C*		*NF1*	*RUNX1*	*ARID1A*
NCOR1			*MAP2K4*		*RUNX1*		*MAP3K1*	*GATA3*	*TBX3*
DHX30			*GATA3*		*CBFB*		*PIK3CA*	*TP53*	*GPS2*
BRCA1					*NCOR1*		*CDH1*	*NCOR1*	*GATA3*
MAP2K4					*CDKN1B*		*NCOR1*	*NF1*	*CBFB*
TBL1XR1					*NF1*		*ATR*	*ARID1A*	*AKT1*
TP53					*TBX3*		*AKT1*	*TBX3*	*SF3B1*
					ARID1A		*FBXW7*	*AKT1*	*CDKN1B*
					RB1		*PIK3R1*	*ERBB2*	*PIK3CA*
					SF3B1		*SMAD4*	*PIK3CA*	*NCOR1*
					AKT1		*MLLT4*	*SF3B1*	*USP28*
					PIK3CA			*ERBB3*	*FOXO3*
					SMAD4			*CDKN1B*	*NF1*
					ERBB2			*KRAS*	
					KRAS			*FOXO3*	
					MLLT4				
					FOXO3				
					CDKN2A				
					ERBB3				
					CTCF				
					FBXW7				
					BRCA1				
					PIK3R1				
					PDE4DIP				
					GPS2				
					USP9X				

Complete list of genes included in the targeted sequencing for METABRIC somatic mutation data

Table A.10 – list of genes (n = 173) included in the targeted sequencing panel for METABRIC somatic mutation data

ACVRL1	BAP1	COL12A1	FANCD2	KDM3A	MEN1	NR3C1	PRR16	SIAH1	TG
AFF2	BCAS3	COL22A1	FBXW7	KDM6A	MLL2	NRAS	PTEN	SIK1	THADA
AGMO	BIRC6	COL6A3	FLT3	KLRG1	MLLT4	NRG3	PTPN22	SIK2	THSD7A
AGTR2	BRAF	CTCF	FOXO1	KMT2C	MTAP	NT5E	PTPRD	SMAD2	TP53
AHNAK	BRCA1	CTNNA1	FOXO3	KRAS	MUC16	OR6A2	PTPRM	SMAD4	TTYH1
AHNAK2	BRCA2	CTNNA3	FOXP1	L1CAM	MYH9	PALLD	RASGEF1B	SMARCB1	UBR5
AKAP9	BRIP1	DCAF4L2	FRMD3	LAMA2	MYO1A	PBRM1	RB1	SMARCC1	USH2A
AKT1	CACNA2D3	DNAH11	GATA3	LAMB3	MYO3A	PDE4DIP	ROS1	SMARCC2	USP28
AKT2	CASP8	DNAH2	GH1	LARGE	NCOA3	PIK3CA	RPGR	SMARCD1	USP9X
ALK	CBFB	DNAH5	GLDC	LDLRAP1	NCOR1	PIK3R1	RUNX1	SPACA1	UTRN
APC	CCND3	DTWD2	GPR124	LIFR	NCOR2	PPP2CB	RYR2	STAB2	ZFP36L1
ARID1A	CDH1	EGFR	GPR32	LIPI	NDFIP1	PPP2R2A	SBNO1	STK11	
ARID1B	CDKN1B	EP300	GPS2	MAGEA8	NEK1	PRKACG	SETD1A	STMN2	
ARID2	CDKN2A	ERBB2	HDAC9	MAP2K4	NF1	PRKCE	SETD2	SYNE1	
ARID5B	CHD1	ERBB3	HERC2	MAP3K1	NF2	PRKCQ	SETDB1	TAF1	
ASXL1	CHEK2	ERBB4	HIST1H2BC	MAP3K10	NOTCH1	PRKCZ	SF3B1	TAF4B	
ASXL2	CLK3	FAM20C	HRAS	MAP3K13	NPNT	PRKG1	SGCD	TBL1XR1	
ATR	CLRN2	FANCA	JAK1	MBL2	NR2F1	PRPS2	SHANK2	TBX3	