Automatic tumour typing based on patterns of somatic passenger mutations

by

Gurnit Singh Atwal

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Molecular Genetics
University of Toronto

Abstract

Automatic tumour typing based on patterns of somatic passenger mutations

Gurnit Singh Atwal
Doctor of Philosophy
Graduate Department of Molecular Genetics
University of Toronto
2021

In cancer, a tumour's cell of origin is the strongest determinant of its clinical behaviour. While cell of origin is typically clear at the time of diagnosis, 3-5% of cancer patients present with a metastatic tumour and no obvious corresponding primary tumour. Despite advances in molecular testing, imaging, and pathology, the primary tumour site cannot be inferred in the majority of these cases. Recent large-scale analysis of cancer genomes has uncovered strong associations between cancer type and somatic mutations, prompting the use of somatic mutations as a tool for identifying cancer type. While existing approaches have attempted to use cancer-associated mutations, which may be more common in specific cancer types to infer the primary tumour type from the metastatic tissue, these methods have had only limited success. A more promising alternative is to use the association between patterns of somatic passenger mutations and cancer type, by exploiting the relationships between both regional mutation density and cancer type, and mutational processes and cancer type. Somatic point mutations accumulate in regions of closed chromatin, and so mutation density provides information about chromatin state, which in turn offers hints about the underlying cell type. As some mutational processes are highly cell-type specific, mutational processes also provide clues about cancer type. In this thesis, I describe a number of deep learning systems for automatic tumour typing based on patterns of somatic passenger mutations. I then address challenges for translating the classifier into clinical scenarios through the use of multiple algorithmic improvements. First, I make use of modern advancements in deep learning to extend the classifier to accurately discriminate between 29 cancer types. I then use a number of statistical methods for assessing the uncertainty in the model's predictions, and for improving uncertainty quantification. Finally, I make use of information theoretic metrics to use the model's predictive uncertainty to automatically detect cancer samples that come from rare cancer types that the model was not trained to classify. These studies demonstrate the utility of passenger mutations as a tool for identifying cancer type, and address challenges for translating the deep learning classifier into clinical settings.

Acknowledgements

First off, I would like to say that I've had a wonderful time working on this PhD. My biggest thanks to my co-supervisors, Drs. Quaid Morris and Gary Bader, for their guidance and support. Quaid and Gary provided me with the freedom to make my own (plentiful) mistakes. The environment they provided for me allowed me to gain the confidence to investigate the ideas I thought were most interesting. While most of my ideas were not fruitful, the opportunity to make these mistakes has had a profound effect on my personal and professional development. I am fortunate to have had their mentorship, and it has been a pleasure working in their labs.

I have been fortunate to have received the guidance and mentorship of a number of incredible scientists throughout my PhD. I am thankful to my supervisory committee members: Drs. Daniel Schramek and Trevor Pugh. Thank you for challenging me, for the difficult questions, and the wonderful advice. I have been fortunate to receive guidance from Drs. Lincoln Stein and Michelle Brazas. I thank them for opening up professional opportunities for me, and for providing me with advice about my next steps as a scientist.

I had the pleasure of collaborating with many talented scientists for the work presented in this thesis. I thank my main collaborator, Wei Jiao for the insightful discussions, for phone calls to discuss data processing, and for helping me advance this work. I thank Dr. Paz Polak for his contributions to my project, and Dr. Linda Sundermann for taking on the task of extending my thesis work into clinical scenarios. I would also like to thank the wonderful undergraduate students that worked with me throughout my PhD: Jacob Chmura, Conor Vedova and Yoonsik Park. It was a pleasure to work with these three young scientists.

My sincere thanks to members of both the Morris and Bader labs. I have greatly enjoyed the opportunity to interact with, and learn from the talented intellectuals in both labs. I am especially thankful for the insightful discussions and guidance when studying machine learning that was provided to me by Chris Cremer, Alex Sasse, Adamo Young, and Amir Khasahmadi. I will not forget discussing deep learning over hard-boiled eggs at NeurIPS, or discussions about discrete latent variable models at a bar in Vancouver. I would also like to thank Drs. Kevin Ha and Rozy Razavi for helping me navigate all stages of my PhD. It would not have been possible to arrive at this point without their guidance, and I am truly thankful they took the time to answer any questions I had.

In writing this, I had hoped to avoid drawing from too many common tropes. However, the treasures of this journey really were the friends I made along the way. To my friend Jeff Wintersinger. Thank you for the discussions, for being an awesome office-mate, and, despite getting us lost on multiple occasions, for being the best conference travel-buddy in the world. To my friend Kaitlin Laverty. Thank you for trying on hats with me in Manhattan, for listening to me whine about literally everything, and for the monthly gossip sessions. And my friend Nil Sahin. Thank you for pushing me to become the person that I am today, for your unwavering support, and for the semi-regular pep talks. I am forever grateful that these three incredible people have graced me with their friendship.

I am, of course, forever indebted to my parents, Manjit and Inderjit, who arrived in Canada from Punjab over 30 years ago. Their hardwork, sacrifice and determination to build a life in Canada provided me with the foundation upon which all my achievements lie. And thank you to my siblings, Abneet, Joti and Ranjit, my sister-in-law, Sukhi and my aunt Kuldip for their unconditional love and support.

Finishing a thesis is a grueling task, and finishing this thesis during a once in a lifetime pandemic has been especially draining. Fortunately, my baby nephew, Savr Singh served as an instant source of happiness and energy. I am thankful that this bundle of joy was here to lift my spirits when I was down, and to keep me going throughout the pandemic. It has been wonderful watching this little guy grow up, and I am looking forward to watching what he gets up to next.

I would like to recognize four individuals I lost during this journey: my aunt Manjit who bravely fought cancer; my aunt Sukhninder who always lived life on her own terms; my friend Vijay whose humour and rants about jiu-jitsu are solely missed; and my grandfather, Gurbax who pushed me to pursue a PhD.

Last but not least, I dedicate this thesis in loving memory of my grandmother, Nachhatar who lost her battle with glioblastoma, which ultimately started me on my academic journey. Her love, care and support will always be missed.

Contents

List of Tables

List of Figures

List of Abbreviations

BO	Bayesian optimization
CUPS	Cancers of unknown primary
cfDNA	Cell-free DNA
CNV	Copy-number variant
CV	Cross-validation
DNN	Deep neural network
DUQ	Deterministic uncertainty quantification
EMT	Epithelial-to-mesenchymal transition
ECE	Expected calibration error
HER2	Human epidermal growth factor receptor-2
IHC	Immunohistochemistry
INDEL	Insertions and deletions
miRNA	micro-RNA
ncRNA	non-coding RNA
NGS	Next-generation sequencing
PARP	Poly-ADP ribose polymerase
PCAWG	Pan-cancer analysis of whole genomes
RBF	Radial basis function
RF	Random forest
RT-PCR	Real time polymerase chain reaction
SBS	Single-base substitution signature
SNV	Single-nucleotide variants
SV	Structural variant

VAF	Variant allele frequency
WES	Whole-exome sequencing
WGS	Whole-genome sequencing

Chapter 1

Introduction

Somatic cells accumulate multiple mutations over cycles of cellular division. Mutations in somatic cells range from single nucleotide variants (SNVs), short insertions or deletions, and large-scale mutations such as structural variants and copy-number variants. Point mutations or SNVs are the most studied variant in somatic cells, and represent the majority of all mutations (Gerstung et al. 2017). The effects of mutations exist on a continuum that depends on the selective pressures placed upon the cells (Martincorena et al. 2017). Some mutations are disadvantageous to the cell and result in cell-death, senescence, or impaired growth. A small subset of the mutations, termed driver mutations, provide a selective advantage to the cells that contain them by, for example, disrupting mechanisms of normal homeostatic regulation. The majority of all point mutations are selectively neutral and are termed passenger mutations (Martincorena et al. 2017). Through a process of genetic diversification and selection, the aggregate effect of these mutations contributes to a process of somatic evolution that can result in the development of cancer (Gerstung et al. 2017).

Importantly, mutations in somatic cells are the result of a number of distinct mutational processes, some of which are highly specific for certain cell types (Alexandrov et al. 2020). Moreover, regional mutation density varies significantly, both across a single genome, and across different cell types (Stamatoyannopoulos et al. 2009; Hodgkinson, Chen, and Eyre-Walker 2012; Schuster-Böckler and Lehner 2012; Supek and Lehner 2015; Polak et al. 2015; Lee, Abd-Rabbo, and Reimand 2020).

Tumour typing is the diagnostic practice of identifying the cell of origin for a given tumour. The cell of origin for a tumour is defined by anatomical site from which the tumour is derived, and the histology of the tumour. While the use of precision medicine in cancer aims to supplement or replace traditional tumour typing by targeting molecular characteristics of a tumour instead of characteristics specific to a tumour's cell of origin, currently, a tumour's cell of origin is the strongest indicator of disease progression and clinical presentation. Moreover, a tumour's cell of origin is the single strongest predictor of therapeutic response (Hyman et al. 2015; Penson et al. 2019; Jiao et al. 2020). Correctly identifying the cell of origin, therefore, is a critical task for guiding decision making for a patient, and forms the basis instructing the use of cancer-specific therapy, which has been shown to be more effective than broad-spectrum chemotherapy (Greco 2013). The past few decades have seen remarkable advances in diagnostic protocols, which have resulted in overall improvements in the accuracy of traditional tumour typing approaches. Despite these advances, challenges arise when determining the primary tumour of origin for metastatic lesions, and in differentiating between a late metastatic recurrence or the emergence of a secondary primary tumour (Pavlidis and Pentheroudakis 2012; Vogt et al. 2017). Correctly leveraging the information that is contained in the mutational history of a tumour is a possible route for accurately determining tumour type. Here I study the use of patterns of somatic mutations for the accurate discrimination of cancer type.

My thesis introduction highlights current practices in cancer diagnostics and treatment. I also highlight two challenging diagnostic scenarios that are not adequately addressed by current practices in cancer diagnostics - cancers of unknown primary (CUPS) and multiple primary tumours. I provide an overview of an avenue for addressing difficult to diagnose tumours, by exploring the association between mutational processes, somatic mutation rate and cell of origin features. Finally, I review machine learning methods that have utility in identifying cancer type, and I review methods for quantifying uncertainty in predictive models that are pertinent to the thesis.

1.1 Tumour typing: traditional diagnostic approaches, molecular diagnostics

Together, the anatomical location from which a tumour is derived, and the histological properties of the cells comprising the tumour, define broad categories of cancer types. As such, the process of tumour typing is the process of determining the histopatholgy and organ of origin for a tumour. Traditionally, this is done through radiographic and pathologic examination by physicians, but more recently, the use of molecular tests and genomics have been integrated in a form of molecular diagnostics. In this section, I provide an overview of tumour types, traditional diagnostic approaches, and molecular diagnostic approaches.

1.1.1 Cancer histopathology

Most commonly, cancers are categorized by their histology or tissue type and their organ of origin. Tumours with the same cell of origin can be further categorized into subtypes based on shared molecular characteristics of the tumour (Hayward et al. 2017). While molecular subtyping of tumours has a role to play, particularly when considering precision or personalized medicine, the cell of origin of a tumour forms the basis for oncological therapy, and is a crucial piece of information when trying to understand the clinical progression of a tumour. Consequently, the bulk of diagnostic work focuses on identifying the organ from which a tumour originates, and the histology of the tumour.

Organ of origin can typically be identified through a combination of blood tests, physical examination, biochemical tests and radiographic imaging. The bulk of diagnostic work comes down to pathological examination to identify the histology of the tumour (Fu et al. 2020). The histology of a tumour is typically determined through pathological assessment of tissue samples. Samples can be obtained through biopsy in the case of solid tumours, and blood tests for liquid cancers. Histologically, cancers are grouped into six broad categories. These categories consist of carcinomas, sarcomas, myeloma, lymphoma, leukaemia and mixed type.

Carcinoma represents the majority of all cancer cases and is defined by cancer of epithelial tissue (Siegel, Miller, and Jemal 2020). Carcinomas are further divided into two major subtypes: adenocarcinoma and squamous cell carcinoma. Adenocarcinomas develop from glandular or secretory cells. These cells make up organs and some tissues. Squamous cell carcinoma develops from squamous epithelium which is comprised of thin, flat cells. This tissue type is found in the skin and lining of some organs. Most cancers of the head and neck tumours, and cervical tumours are squamous cell carcinomas. Squamous cell carcinomas also represent the second most common cancer of the skin. The biological and clinical characteristics of adenocarcinomas and squamous cell carcinomas of matched organ of origin can have significant differences. For example, oesophagal adenocarcinomas are very strongly associated with a history of Barret's oesophagus, while squamous cell carcinoma of the oesophagus is mostly associated with tobacco smoke or excess consumption of alcohol (Kim et al. 2017; Falk 2015).

Cancers that originate in connective or supportive tissues are called sarcomas, which consist of a highly diverse set of mesodermal malignancies (Ceyssens and Stroobants 2011) that include cancers that develop in skeletal muscle, adipose tissue, blood, lymphatics, peripheral nerves and in the central nervous system. Broadly, sarcomas are split into soft tissue sarcoma, primary bone sarcoma, and certain malignancies that arise in the central nervous system. Primary bone sarcomas consist of a number of cancer types, including cancer of bone-forming osteoblasts (osteosarcoma), Ewing sarcoma, cancers that form in the bones of the spine (chordoma) and cancers of cartilaginous tissues (chondroblastomas). Soft tissue sarcomas are similarly diverse and include cancers of the blood or lymphatic vessels (angiosarcoma), cancers that arise in nerve cells lining the gastrointestinal system (gastrointestinal stromal tumour), and cancers that arise in skeletal muscle (rhabdomyosarcoma) (Vodanovich and M Choong 2018). As the central nervous system contains supportive tissue, many cancers of the central nervous system form a third category of sarcomas. These include gliomas and glioblastomas, which both arise in glial tissue and represent the most common central nervous system malignancies in adults (Carlsson, Brothers, and Wahlestedt 2014).

The hematopoietic system is responsible for producing a variety of highly specialized blood cells and consists of organs and tissues involved in producing blood cells. This includes bone marrow, spleen, thymus and lymph nodes. Given the wide variety of organs and tissues involved in this system, a variety of cancer types arise in the hematopeietic system. Myeloma is a malignant disease of plasma cells that arises from post-germinal centre plasma cells that migrate back to the bone marrow and is a relatively rare cancer type (Al-Farsi 2013). Leukaemia is a liquid cancer of hematopoietic stem cells in bone marrow (Mak, Saunders, and Jett 2014). Broadly speaking, leukaemia can be subdivided into four cancer types: acute lymphoblastic leukaemia, acute myelogenous leukaemia, chronic lymphocytic leukaemia and chronic myelogenous leukaemia. Tumours that develop in the lymphatic system are called lymphomas. Lymphomas can develop in any part of the lymphatic system including lymphatic vessels, lymph nodes, the spleen, tonsils or thymus. Unlike leukaemia, lymphomas are solid tumours. These tumours are broadly split into Hodgkin's lymphoma and Non-Hodgkin's lymphoma.

A final category of cancer consists of tumours that contain a mix of the five categories discussed above. This category, called mixed type, contains a mix of cells from the different categories of cancer (Moran et al. 1994). For example, adenosquamous carcinoma contains both squamous cells and gland-like cells. These tumours tend to either be poorly differentiated or contain cells from multiple histologically distinct cancer types. Occasionally, histological transformation can occur in a tumour which can lead to switching of tumour histology. For example, oncogene-driven lung adenocarcinomas can undergo small-cell transformation following exposure to tyrosine kinase inhibitors (Lin et al. 2020). A similar phenomenon is seen in prostate cancer when given androgen deprivation therapy (Volta et al. 2018). These results suggest that tumour cells possess a significant degree of plasticity which can be drawn upon to escape selective pressures imposed on them due to anti-cancer therapy.

1.1.2 Histopathologic and cytopathologic examination

Histopathologic examination is the process of examining disease using a biopsy that is fixed to a glass slide. Specimens can come from tumour sections/biopsies or blood tests. Visualization of different components is done using several stains that reveal specific cellular components. Hematoxylin-Eosin (H&E) staining is one of the most commonly used staining methods. Hematoxylin stains cell nuclei blue, and Eosin stains cytoplasm and connective tissue pink (Painter, Clayton, and Herbert 2010). Cytology involves the study of the structure, function and chemistry of cells. Cytology commonly involves the presence of isolated cells and cell clusters in images. Still, it will lack a higher-level organization of a biopsy sample found in a histopathology sample (Al-Abbadi 2011). Simple staining of sections and cytology can differentiate between malignant and benign neoplasms by highlighting the tissue sample's cellular organization and morphological differences amongst the cells. It can also differentiate squamous cell carcinomas from adenocarcinomas by comparing the tissue sample to the expected organization and structure of normal squamous or gland-like tissue.

Immunohistochemistry (IHC), the use of antibodies to detect antigens of interest, is a standard procedure for the pathologic examination of tumour samples (Yatabe et al. 2019; Duraiyan et al. 2012; Rosai and Ackerman 1979). In IHC, sectioned biopsies are incubated with antibodies targeting antigens of interest. Enzymatic reactions, fluorescent dyes, radioactive elements or colloidal gold is then used to confirm and visualize antibody binding (Matos et al. 2010). IHC with specific markers can be used for cancer diagnosis, tumour staging and identification of tumour histopathology. Through the use of antibodies for tumour-specific antigens, oncogenes and other cancer biomarkers, IHC can also be used as a prognostic tool. Antibodies allow IHC markers to be used to identify molecular subtypes of cancer, and also allows IHC to be used as a tool for guiding treatment decisions. A good example of this is found when looking at breast cancer. When patients present with breast cancer, tumour samples are commonly assessed for the expression of three cell-surface markers: estrogen receptor, human epidermal growth factor receptor (HER)-2 and progesterone receptor (Yin et al. 2020) (Figure 1.1). The use of these markers can differentiate breast cancer into broad subtypes and inform the use of specific treatments that target the over-expression of the markers being stained for. For example, the presence of human epidermal growth factor receptor (HER)-2 suggests the use of trastuzumab, a monoclonal antibody that targets HER-2 (Boekhout, Beijnen, and Schellens 2011). Some other commonly used diagnostic IHC markers include cytokeratins for identifying carcinomas, CD15 and CD30 for Hodgkin's lymphoma, CD20 and CD3 for differentiating between B-cell lymphomas and T-cell lymphomas (Orakpoghenor et al. 2018).

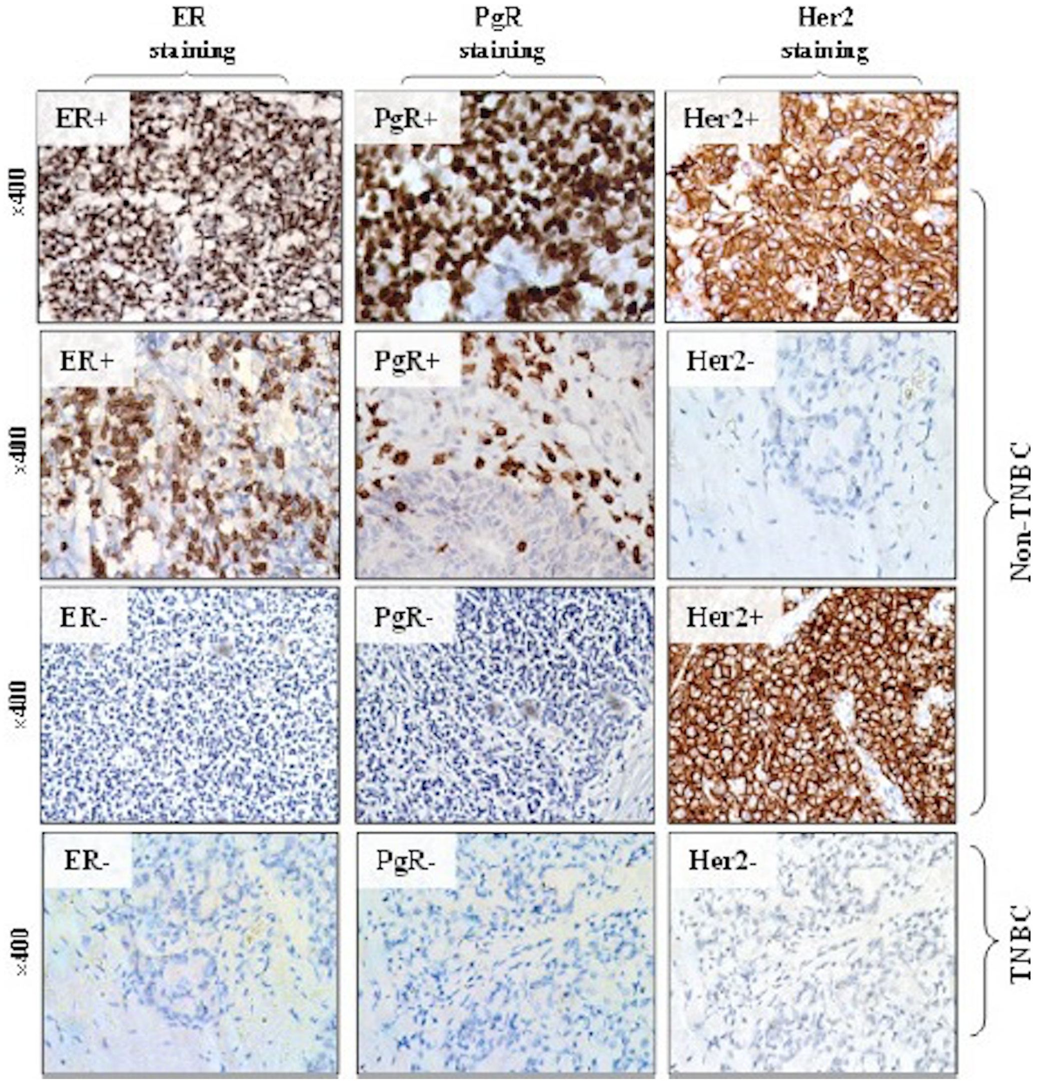


Figure 1.1: Staining for breast cancer cell-surface markers. Representative micrographs showing staining for estrogen receptor (ER), progesterone receptor (PgR) and human epidermal growth factor receptor 2 (Her2). Figure from: (Sikandar et al. 2017), licensed under Creative Commons Attribution-NonCommercial 4.0 International License (`https://creativecommons-org/licenses/by-nc/4.0/`).

More recently, developments in novel IHC markers have allowed for increased resolution when differentiating tumour types. Cellular lineages are often differentiated or defined by gene expression programs controlled by lineage-specific transcription factors (Fueyo et al. 2018). Identifying the presence of these transcription factors in a tumour can provide information for identifying the lineage of cells constituting that tumour. Lineage specifying antibodies can be used to differentiate between cell types of similar origin such as differentiating between gland-like or squamous cells, and differentiating between types of mesenchyme derived cells in the case of a suspected sarcoma. Multiple transcription factors can now be targeted with antibodies, allowing for lineage-specific factors to be stained for (Kei and Adeyi 2020; Hornick 2014). For example, in lung cancer, the use of adenocarcinoma specific antibodies TTF-1 and

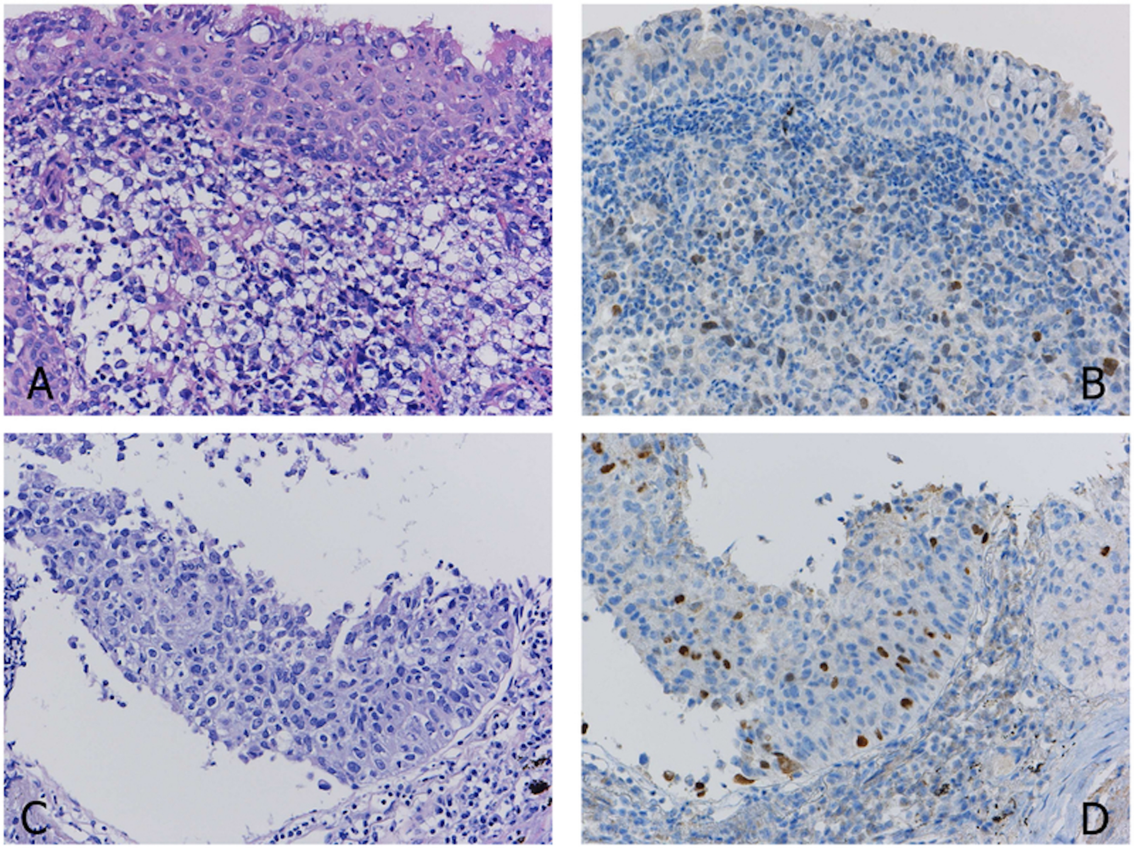


Figure 1.2: Immunohistochemical differentiation of lung adenocarcinoma and lung squamous cell carcinoma. (A) Tumour cells with unclear morphology (B) TTF1 staining identifying adenocarcinoma (C) Tumour cells with unclear morphology (D) p40 staining identifying squamous cell carcinoma. Figure from: (Yatabe et al. 2019), licensed under Creative Commons Attribution 4.0 International License (`https://creativecommons-org/licenses/by/4.0/`).

Napsin A are often used to differentiate between lung adenocarcinoma and lung squamous cell carcinoma (Yatabe et al. 2019; Inamura 2018) (Figure 1.2). In soft tissue sarcomas, lineage-restricted transcription factors are commonly used to identify cancer type. This is done by looking for transcription factors only expressed in a small number of cell types. For example, myogenin is a lineage-specific transcription factor that is only expressed in skeletal muscle. Expression of myogenin in a soft-tissue tumour provides evidence that the tumour originated in skeletal muscle (Hornick 2014). Although these antigens are used to identify specific cell types, there are varying degrees of specificity for each antibody, leading to potential misdiagnosis. For example, TTF-1 is expressed in 17% of lung squamous cell carcinomas (Inamura 2018).

1.1.3 Molecular profiling and clinical tumour sequencing

The initial Human Genome Project was a decade-long effort requiring substantial financial investment to sequence a single genome (Lander et al. 2001). Technological developments since the initial project, however, have allowed for multiple genomes or transcriptomes to be sequenced in both time and cost-efficient manner. These technical advancements, referred to as next-generation sequencing (NGS) technologies have enabled the discovery of genomic markers that can be used to identify cancer type, inform treatment and provide prognostic information (Tsimberidou et al. 2020). Whole-genome sequencing (WGS), whole-exome sequencing (WES) and RNA sequencing provide a relatively unbiased look into a tumour's genomics. These methods aim to sequence the entire genome, exome or transcriptome, and, consequently, not biased towards pre-defined disease-associated genes. An additional benefit of not focusing on pre-defined genes is that these approaches provide an opportunity for discovery, and can associate previously unknown elements of tumour genomics with clinical characteristics. By characterizing the entire genome or transcriptome, these approaches can improve the sensitivity of clinical tools that use genomics as a feature. For example, WGS of a tumour provides a complete characterization of the mutational signatures (discussed in detail in a later section) that were active within that tumour. As such, clinical tools that utilize mutational signatures have increased sensitivity using WGS compared to WES. HRDetect, as an example, is a tool that detects homologous recombination deficiency from patterns of somatic mutations. When HRDetect is applied to WGS data, it has a sensitivity of 86%, but sensitivity drops to 46% when applied to WES (Davies et al. 2017). Since HRDetect can be used as the basis for treatment with poly-ADP ribose polymerase (PARP) inhibitors, the sensitivity provided by WGS provides a significant benefit in the clinic.

Targeted DNA sequencing

Although there are advantages to using unbiased approaches such as WGS in a clinical setting, unbiased approaches are not commonly utilized in the clinic. Until recently, these methods were prohibitively expensive and required significant computational investment to analyze. Clinical practice has instead focused on gene panels, which target a small number of pre-selected genes. Targeted approaches have some advantages compared to methods such as WGS. First, they are more cost-effective and require less computational investment to analyze. Second, by focusing on well-characterized genes, they can quickly provide a clinician with important information about actionable targets and other biological characteristics of a tumour. Targeted approaches include those that make use of NGS such as the MSK-IMPACT panel which involves the targeted deep sequencing of protein-coding exons from 468 cancer-associated genes (Zehir et al. 2017). This assay has been deployed in the clinic and used for informing diagnosis, prognosis and therapy (Penson et al. 2019; Stadler et al. 2020). While gene panels are more commonly available in clinical settings, they risk missing opportunities for discovery. Our current understanding of cancer-associated genes further limits them.

RNA Sequencing and microarrays

RNA sequencing allows for the cancer transcriptome's characterization and can provide information about biological pathways active within a tumour. Occasionally, biological pathways that are active within the tumour can be the target of therapeutic agents, which may provide an avenue for tumour treatment (Uzilov et al. 2016). RNA sequencing can also provide information about gene fusions within a tumour. Gene fusions are an important oncogenic process, and gene fusions are key developmental markers in certain cancers such as the BCR-ABL1 fusion in chronic myeloid leukaemia (Haas et al. 2019). Identification of gene fusions can guide the use of therapeutic agents targeting gene fusions. For example, identifying BCR-ABL1 fusions can suggest the use of imatinib mesylate, and identification of EML4-ALK fusions can suggest treatment with crizotinib (Druker 2004; Shaw et al. 2011). To this end, targeted assays have been developed, specifically identifying cancer-associated fusion genes using targeted RNA sequencing (Heyer et al. 2019). RNA sequencing also allows for profiling of multiple types of non-coding RNAs (ncRNA) such as microRNAs (miRNA) which are implicated in the development of cancer and has been used as a feature for cancer diagnostics (Laplante and Akhloufi 2020; Anastasiadou, Jacob, and Slack 2018). An alternative to RNA sequencing, which provides some of the same information is the gene expression microarray. A gene expression microarray works by creating cDNA from the extracted mRNA of a sample. cDNA is labelled with a fluorescent molecule that will emit light when hybridized to DNA. The cDNA is then hybridized to DNA fragments on a slide, and the fluorescent emitted from each pre-selected well on the slide is used to quantify the expression of different genes (Govindarajan et al. 2012). Microarrays allow for the expression of thousands of genes to be assessed and have found use in the clinic for diagnosis and instructing therapy (Kurahashi et al. 2013; Sparano et al. 2018).

Methylation assays

Cancer development is associated with large changes to a cell's normal phenotype. Phenotypic changes in cancer development are often accompanied by changes in DNA methylation within a cell. The widespread changes in DNA methylation allow differentially methylated regions to be used as features for distinguishing cancer from non-cancerous tissue and identifying cancer type (Locke et al. 2019). An advantage of methylation-based diagnostics is that methylation can reliably be detected from cell-free DNA (cfDNA), allowing for it to be used to diagnose and identify cancer from non-invasive liquid biopsies (Chu and Park 2017). To this end, several studies have examined the use of methylation from cfDNA for cancer identification and even molecular subtyping of tumours (Paemel et al. 2020; Shen 2018). While these approaches show promise, studies examining the use of methylation from cfDNA as a tool for tumour typing are currently limited to a small number of cancer types and have not shown the ability to generalize to large sources of data.

Real-time PCR

Whole-genome sequencing, gene panels, and RNA sequencing use NGS technologies to profile expression and/or mutations in the tumour. An alternative to these technologies is the use of Real-time Polymerase Chain Reaction (RT-PCR). RT-PCR allows for the quantification of gene expression or mutation status for specifically targeted genes. RT-PCR based methods are the most commonly used methods for quantifying gene expression in clinical settings, and multiple RT-PCR based diagnostic tools are currently in use (Mocellin et al. 2003; Bender and Erlander 2009; Sokolenko and Imyanitov 2018).

Limitations

Regardless of the technology used, genomics cancer testing has several limitations. Genomics tests are often applied to small, formalin-fixed samples subject to degradation and artefact mutations introduced by the fixation process (Prentice et al. 2018). Gene panels focus on sequencing cancer driver genes. Still, it is often difficult to determine the role of putative passenger mutations. Knowledge of cancer-associated genes depends heavily on the ability to detect positive selection signals from a limited number of sequenced tumour samples (Sabarinathan et al. 2017). As cancer is a process of somatic evolution, tumours typically have a large degree of heterogeneity which is not easily captured or accounted for in single-sample NGS approaches, including WGS (Jamal-Hanjani et al. 2017; Gerstung et al. 2017).

1.1.4 Detecting mutations from sequencing data

To make use of the information generated through NGS, it is essential to accurately identify the variants present in a tumour sample. To this end, a number of somatic SNV calling algorithms have been developed (McKenna et al. 2010; Koboldt et al. 2012; Kim et al. 2018b). These algorithms vary in the specifics of the statistical models used, but fundamentally operate similarly. An SNV can be directly observed from NGS data and using this data, and it is possible to estimate the relative abundance of an SNV in a given sample. This is done using the number of sequencing reads containing each SNV, and the total number of reads at that locus. The ratio of variant reads to total reads represents the variant allele frequency (VAF) of the SNV. To separate germline and somatic SNVs, SNV calling is typically done by utilizing aligned reads from both a tumour and a matched non-cancerous sample.

In theory, all mutations regardless of VAF can be observed given sufficient read depth. In practice, however, the process of SNV calling is made difficult by the presence of many types of biases, errors and noise present in NGS data (Cibulskis et al. 2013). A variety of different sequencing errors can be produced, including CG deletions, short deletions and SNVs. Generally, SNVs are the most frequent error across sequencing platforms (Fox et al. 2014). Sequencing errors can be mistaken for low prevalence mutations. Even with a relatively low error rate of one in 1000 bases, sequencing across the three billion bases of the human genome will produce variant reads that contain the same sequencing error at the same locus (Travis 2011). Common sequencing errors include the introduction of substitutions, short insertions or deletions, and misaligned reads. Distinguishing between real, low VAF mutations and sequencing errors require a precise model of the noise distribution of NGS data. The statistical models used to differentiate noise from true variants are among the core differences between variant calling pipelines. Although significant work has been done to improve SNV calling accuracy, the sensitivity of commonly used pipelines ranges from 80 to 90%, and approximately 95% of variants called by these pipelines are true SNVs (Campbell et al. 2020). This result suggests that any individual method will produce erroneous variant calls. One way to improve SNV calling sensitivity is to use a consensus of multiple pipelines, which has been shown to increase the sensitivity of variant calling without reducing specificity (Campbell et al. 2020).

1.2 Challenges in tumour typing: cancers of unknown primary, and multiple primary tumours

Traditional approaches to tumour typing, including histopathologic examination and molecular diagnostic methods, have remarkable success in correctly identifying cancer type. Despite this success, certain challenging diagnostic scenarios exist where traditional approaches struggle to identify cancer type correctly. Two particularly difficult diagnostic challenges are cancers of unknown primary (CUPS) and patients who present with multiple primary tumours. CUPS is a heterogeneous category of cancer where patients present with recognizable metastatic lesions and no obvious or identifiable primary tumour (Greco 2013). Multiple or secondary primary tumours describe the scenario in which a patient has more than one tumour in the same or different organs (Vogt et al. 2017).

In this section, I provide an overview of CUPS and multiple primary tumours. This is followed by a brief overview of diagnostic approaches for CUPS, and an examination of the effects of correctly identifying cancer type on prognosis for patients with CUPs.

1.2.1 Multiple primary tumours

Recent decades have seen remarkable improvements in diagnostic techniques and treatment, resulting in increases in patients' long-term survival with malignancy. However, the cost of this clinical success is an increased incidence of patients presenting with multiple primary tumours. Multiple primary cancers are defined as multiple primary malignant tumours of different histopathologic origins in a single individual (Vogt et al. 2017). Incidence of multiple primary cancers is high and represents 16% of incident cancers

(Travis 2006). While the prevalence of multiple primary tumours has increased due to advances in cancer care, multiple primary tumours are not a new phenomenon. Reports from the 1920s suggest that approximately 5% of malignancies present with multiple primary tumours (Vogt et al. 2017). The increase in the rate of multiple primary tumours is likely the result of both improvements in care and improvements in detecting and diagnosing cancer. Diagnostic challenges exist when patients present with multiple primaries, particularly when a patient with previous cancer history and exposure to therapy presents with an additional primary. There is potential for new metastases to be from the second primary tumour or the first diagnosis. This will have important implications in patient prognosis and therapeutic management. Incidence of multiple primaries varies according to the initial primary tumour identified in a patient and ranges from 1% in primary liver cancer to 16% in primary bladder cancer (Hayat et al. 2007). Synchronous cancers occur when a patient is diagnosed with more than one histopathologic tumour within an interval of fewer than 2 months. If an interval of greater than 6 months passes between diagnoses, the patient has metachronous multiple primary (Amer 2014). In both synchronous and metachronous cases, accurately identifying cancer type forms the basis for guiding clinical decision making.

There are two major diagnostic challenges associated with multiple primary tumours. The first occurs when one of, or both of, the tumours metastasize. In this scenario, it is essential to identify which primary tumour gave rise to each metastatic lesion. As metastatic lesions can be highly undifferentiated, determining origin can be difficult. In the metachronous case, the older primary tumour may not be available for comparison, which can make the process of identifying the origin of a metastatic lesion more difficult. The second challenging case involves differentiating between a new primary tumour, and a late metastatic recurrence. In this case, it is essential to correctly determine if a new lesion is the result of a newly forming primary tumour, or if it represents a metastatic clone from the first primary tumour. Correctly identifying cancer type in either of these scenarios will form the basis for guiding treatment and understanding the clinical presentation and progression of the malignancies.

1.2.2 Cancers of unknown primary

CUPS represents a diagnostic scenario in which clinicians are presented with a poorly differentiated metastatic tumour that cannot be identified using imaging, pathological or molecular examination. CUPSs represents 3-5% of new cancer diagnoses, and constitute a highly heterogeneous set of cancers that are currently the seventh most common cancer diagnosis, and the fourth leading cause of cancer-related mortality (Pavlidis and Pentheroudakis 2012; Pavlidis et al. 2003). These tumours tend to either be completely undifferentiated or lack the characteristics of any primary tumour. In most cases, pathologists cannot identify primary tumour site post-humously during the autopsy, which suggests significant regression of the primary tumour in many CUPS patients (Ferracin et al. 2011). The majority of CUPS patients present with carcinomas, or cancers that originate in epithelial tissue. Of CUPS cases that are carcinomas, cancer is most often categorized as adenocarcinomas (Greco et al. 2010). CUPS is associated with a short history of symptoms and aggressive behaviour (Pavlidis, Khaled, and Gaafar 2015). In most cases, multiple organs are involved in the metastatic spread, with liver, lymph nodes, bone and lung being the most common metastatic sites (Greco 2013). The patterns of metastatic spread differ between CUPS and known cancer of the same type. As an example, Pancreatic adenocarcinomas have a 4-fold higher incidence of metastasizing to the bone compared with pancreatic adenocarcinomas of known origin (Pavlidis, Khaled, and Gaafar 2015). While the organ specificity of metastases remains poorly understood, this fact may suggest that CUPS is molecularly distinct from metastases of known origin from the same cancer type. Correctly identifying primary tumour site for CUPS may form the basis for understanding the molecular characteristics of CUPS syndrome. This can help understand the factors that differentiate CUPS from metastases of known origin. It may help understand the mechanisms involved in primary tumour recession, and it can help understand the molecular mechanisms involved in patterns of metastatic spread for CUPS. Standard practice currently splits patients into two clinicopatholigcal categories: those with favourable prognosis (15-20% of patients), and unfavourable (Pavlidis et al. 2003). Patients in both categories have a poor prognosis, but the favourable subset typically has survival times that are twice as long as the average survival time for the unfavourable set. Identification of primary tumour site for these cases presents a potential avenue for improving prognosis. Patients for whom primary tumour site can accurately be determined also can potentially be treated with cancer-specific therapy which generally provides survival benefits (Greco 2013).

As mentioned, in most CUPS occurrences, the primary tumour cannot be identified during the autopsy, which suggests that metastatic spread in CUPS syndrome is accompanied by recession or suppression of the primary tumour. The biological processes involved in tumour recession of the primary tumour, and the development and metastatic seeding of CUPS are poorly understood. This stems, in part, from an inability to correctly identify primary tumour site. Without knowing the primary tumour site, it may be difficult to understand the selective mechanisms that suppress primary tumour development, while allowing or promoting metastatic seeding. Although CUPS develops without a detectable primary tumour, it is possible that the developmental patterns and metastatic seeding of CUPS follow those of tumours of known origin.

Tumourigenesis typically involves a process of invasion and intravasation of cancer cells from the primary tumour into neighbouring tissues, lymphatic vessels or other vasculature (Chaffer and Weinberg 2015; Chiang, Cabrera, and Segall 2016). Metastatic spread occurs when the invading cells from the primary tumour disseminate through lymphatic vessels or other vasculature. This process involves an epithelial-to-mesenchymal transition (EMT) where cancer cells discard epithelial markers, allowing them to lose adhesion to epithelial cells, and acquire mesenchymal markers which allow them to traverse vasculature (Fares et al. 2020).

For the development of CUPS, many of these processes likely contribute to metastatic spread. Multiple explanations exist for the timing of metastatic spread and apparent recession of the primary tumour. One possible explanation for the metastatic spread of CUPS proposes that early metastatic spread occurs before a detectable primary tumour develops, or even the development of a primary malignancy (Rassy, Assi, and Pavlidis 2020; Hu and Curtis 2020). In this model, early metastatic cells alter the microenvironment at the metastatic site and acquire malignant characteristics upon colonization. In this scenario, the primary tumour does not have the time to develop to a detectable size before detecting the metastatic lesions. Another explanation suggests that metastatic spread occurs in a similar pattern to cancers of known primary, but the microenvironment at the primary tumour site prohibits primary tumour growth or eliminates the primary tumour site (Rassy, Assi, and Pavlidis 2020; Hu and Curtis 2020). In this scenario, the microenvironment clears the primary tumour while either allowing or promoting the spread of metastatic cells (Figure 1.3). Given that primary tumour site can be uncovered in autopsy for approximately 30% of CUPS cases, some patients may fall into the first explanation of CUPS spread and a different set of CUPS cases spread through the second mechanism.

A few studies have uncovered some common molecular characteristics of CUPS. By utilizing a gene panel, recent work has shown that many of the most commonly mutated genes in CUPS resemble those of primary or metastatic tumours of known origin (Rassy, Assi, and Pavlidis 2020). On average, CUPS had 3.1 mutations of known oncogenic consequence, but many samples lacked mutations to clinically actionable genes (Varghese et al. 2017). One common characteristic of CUPS is the presence of chromosomal instability. Approximately 70% of CUPS have up-regulated activity of multiple DNA damage response networks, suggesting that these tumours have increased levels of DNA damage and therefore, chromosomal instability (Hedley, Leary, and Kirsten 1985). The presence of chromosomal instability may also provide some insight into metastatic mechanisms involved in these tumours. Chromosomal instability can result in cytosolic DNA which has been shown to promote a pro-inflammatory and pro-metastatic program through activating cGAS-STING pathway (Bakhoum et al. 2018).

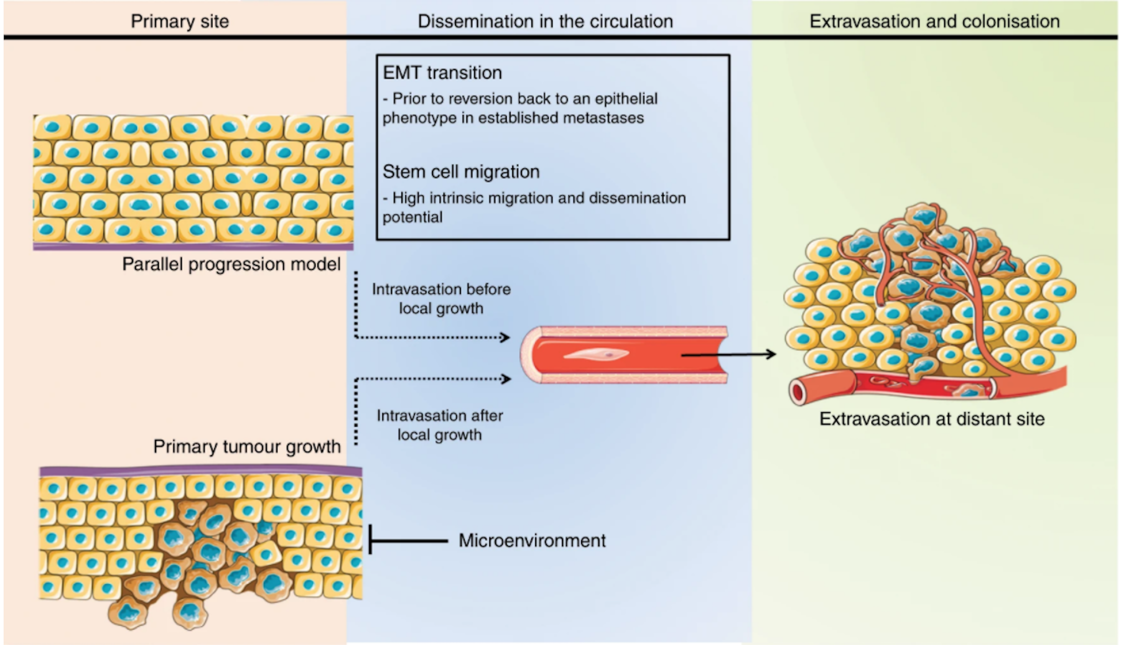

Primary site
Dissemination in the circulation
Extravasation and colonisation
EMT transition
- Prior to reversion back to an epithelial phenotype in established metastases
Stem cell migration
- High intrinsic migration and dissemination potential
Parallel progression model
Intravasation before local growth
Intravasation after local growth
Primary tumour growth
Microenvironment
Extravasation at distant site

Figure 1.3: Competing models for carcinogenesis in cancers of unknown primary. Two competing models of carcinogenesis and metastatic seeding for cancers of unknown primary. Figure from: (Rassy, Assi, and Pavlidis 2020), licensed under Creative Commons Attribution 4.0 International License (`https://creativecommons-org/licenses/by/4.0/`).

1.2.3 Traditional diagnostic approaches

Identifying the primary tumour site can significantly improve patient survival, and may provide information about the underlying biology of this disease. Assessing the accuracy of any CUPS diagnostic program is particularly challenging; however, as the true primary tumour site is rarely known. A potential workaround is to use retrospective studies that examine the results of cancer-specific therapy when a CUPS site is determined through a diagnostic procedure. Diagnosing a patient with CUPS involves identifying the metastatic lesion, and searching for the primary tumour site. Traditionally, this process begins with imaging to identify metastatic lesions and is then followed by examination for determining the primary tumour site. This involves a thorough physical examination and analysis of the patient's medical history. The medical history may be able to provide clues about the primary tumour site. For example, patients with familial cancer syndrome may have a higher risk of developing certain cancers. Similarly, if a patient has a known history of exposure to smoking, cancers more strongly associated with smoking, such as lung cancer, may be likely culprits. Standard blood tests and biochemical examination follow up physical examination and medical history. Finally, multiple imaging studies, histological examination of the metastatic lesions and immunohistochemistry (IHC) are used to determine the primary tumour site.

Radiology and imaging

Typical diagnostic procedures for cancer in general and for CUPS is done using a number of radiological methods including CT scan, MRI and PET-scans. While these methods are standard practice, they do not provide highly accurate identification of primary tumour site. Diagnostic accuracy using fluorodeoxyglucose-PET scans is under 50% (Sève et al. 2007). Similarly, CT scans provide only modest improvements with accuracy of about 55%, heavily biased for a few cancer types (Keller et al. 2011).

Immunohistochemistry

Histopathology represents the most commonly used approach for CUPS diagnosis (Greco 2013). The use of IHC for identifying cancer type for CUPS is a stepwise procedure. First, IHC is used to assign the CUPS into one of the broad categories of cancer - carcinoma, sarcoma, melanoma or lymphoma. Next, identify subtypes of these broad categories: adenocarcinoma, squamous carcinoma, neuroendocrine et cetera. The final categorization stage is to try to identify the primary tumour site. Typically, IHC for CUPS can be used to split CUPS into one of five CUPS subtypes: well-differentiated adenocarcinoma, undifferentiated carcinoma, squamous-cell carcinoma, poorly differentiated neoplasm and neuroendocrine tumour (Alshareeda et al. 2020). While IHC provides the potential for differentiating between different primary sites of these CUPS subtypes, a study examining the ability for IHC to differentiate between 11 metastatic adenocarcinomas suggests that pathologists are typically unable to identify primary site accurately (Sheahan 1993). Improvements in IHC have resulted in modest increases in accuracy for identifying CUPS site of origin, but these results tend to be on a relatively limited number of cancer types.

The main limitation of IHC is that IHC is not well-suited for the identification of the poorly-differentiated tumours that represent many CUPS cases. Despite the availability of antibodies for lineage-specific transcription factors, IHC often cannot identify primary tumour site, even in well-differentiated adenocarcinomas (Varadhachary 2007). This suggests that CUPS lesions have diverged significantly from primary tumours of the same type. Significant divergences from primary tumours in terms of lineage-specific transcription factors and morphology suggest that the chromatin state, and gene expression programs in CUPS may also be significantly different from those found in primary tumours. These differences may significantly impair the utility of gene expression or chromatin features as tools for identifying the primary tumour site.

Biochemical tests

Biochemical tests often test for the presence of serum tumour markers. Studies suggest, however, that serum tumour markers may offer little predictive power for CUPS. At least two markers are overexpressed in approximately 70% of CUPS, but these often encompass multiple markers that provide no specific diagnostic value when co-expressed (Pavlidis, Khaled, and Gaafar 2015).

Molecular diagnostics

In addition to the standard diagnostic protocols described above, multiple molecular diagnostic methods are being deployed or tested for identifying primary tumour site. Many of these methods rely on assays for gene expression, including RT-PCR and microarrays for miRNA expression. Overall, these methods provide an improvement compared to typical histopathologic examination, but are often unable to identify cancer type, and are limited to a relatively small number of cancer types. However, as gene expression profiles can describe cell-specific features, gene expression has been used as a feature for several tumour typing methods. A recent study comparing five commercially available expression-based tests shows that accuracy ranges from 76% to 87% for differentiating between 6 and 49 different cancer types, with accuracy tending to drop as the number of cancer types increases (Ferracin et al. 2011; Monzon and Koen 2010; Bridgewater et al. 2008). Some more recent work utilizing the structure of miRNA stem-loops and deep learning for cancer type identification has achieved 97% accuracy for differentiating 20 anatomical sites (Laplante and Akhloufi 2020). Importantly, this study grouped tumours from the same anatomical site into the same category, suggesting that their model has difficulty differentiating between distinct tumour types from the same anatomical location. Tumours with significant morphological differences from those used for developing these methods are more difficult to identify. As such, methods that use gene expression profiles to identify cancer type tend to struggle with poorly differentiated tumours, which have lost many cell type-specific markers. One way to address this is to use somatic mutations as a feature for identifying cancer type. These methods typically use mutations to cancer-associated genes, copy number profiles, and other mutational features that can easily be accessed from small-scale sequencing. One example is a model using the MSK-IMPACT gene panel as features for identifying cancer type. This model shows modest performance with an overall accuracy of approximately 75% (Penson et al. 2019).

1.2.4 Therapeutic management

The majority of patients with CUPS are treated with broad-spectrum chemotherapy such as platinum-based therapy, gemcitabine or fluorouracil (Greco 2013). While broad-spectrum therapy provides survival benefits compared to forgoing treatment, median survival for these patients is still relatively poor at approximately 9 months (Greco 2013). Cancer-specific therapies, in contrast to broad-spectrum regimens, have some evidence of improving prognosis. Still, the use of cancer-specific therapy for CUPS is currently limited by the inability to identify the primary tumour site.

Prospective studies for directly assessing cancer-specific therapy's ability to improve survival for patients with CUPS are limited for multiple reasons. Firstly, CUPS is a highly heterogeneous disease comprised of multiple cancer types (Greco 2013). This fact makes it difficult to enrol sufficiently many patients from the diverse set of cancer types that constitute CUPS. Secondly, prospective studies are made difficult because the true cancer type is rarely determined for CUPS (Ferracin et al. 2011). This means that it is often unknown if cancer-specific therapy targets the correct cancer type. As such, most evidence favouring cancer-specific therapy for the treatment of CUPS comes from a small number of retrospective studies. In a study that used molecular testing to identify primary tumour site, CUPS cases that were judged to be colorectal cancer based on molecular signatures had significantly higher survival when given a colorectal cancer-specific treatment than patients given empiric therapy (Ma et al. 2006). Median survival for those receiving site-specific therapy in presumed colorectal CUPS ranged between 21 and 30 months, compared to approximately 8 months for those receiving broad-spectrum chemotherapy (Varadhachary et al. 2008; Hainsworth et al. 2012; Greco et al. 2012). This suggests that cancer-specific therapy for colorectal cancer can improve prognosis when the molecular characteristics match those of colorectal cancer. Similar results have also been observed in CUPS presumed to be renal cell carcinoma.

For advanced and/or metastatic renal cell carcinomas, empiric therapy provides no survival benefits. When CUPS cases match the molecular characteristics of renal cell carcinoma and were given cancer-specific therapy, a modest improvement in prognosis was observed (Sorscher and Greco 2012). A review paper examining the effects of cancer-specific therapy compared to empiric therapy suggests that median survival for patients with CUPS receiving empiric therapy was approximately 9 months, compared to 12.5 months for those treated with cancer-specific therapy (Hainsworth et al. 2013). Generally, when a putative site of origin can be identified, survival for patients with CUPS is similar to that of known advanced cancers of the same type (Kim et al. 2018a; Hainsworth et al. 2013).

The availability of cancer-specific therapy does not always translate to increases in patient survival; however, many patients with CUPS do not respond to any form of therapy (Greco 2013; Hainsworth et al. 2013). For cancer-specific therapy to be effective for treating CUPS, an efficacious, cancer-specific therapy must exist for the cancer type being considered. The absence of a suitable therapy for a given cancer type will limit the success of therapeutic approaches. For example, patients that have CUPS identified to be pancreatic adenocarcinoma will continue to have poor overall prognosis as efficacious regimens do not exist for pancreatic adenocarcinomas (Hall et al. 2018). These patients may receive some initial benefit from empiric therapy, but improvements in cancer-specific treatment regiments could improve these results.

This result is not always limited to specific cancer types. By being advanced cancers, cancer-specific therapies for CUPS are often limited when compared to earlier stage malignancies. Therefore, the lack of effective treatment for advanced cancers limits the therapeutic approaches that are effective for addressing CUPS (Morgan, Ward, and Barton 2004). Advances in treating advanced forms of currently unresponsive cancer types, such as pancreatic adenocarcinomas, may increase the efficacy of cancer-specific therapy for CUPS. An alternative or complementary avenue for treatment comes from advances in personalized medicine and immunotherapy. Therapy directed at specific, actionable mutations or pathways has shown some evidence of sustained disease stability and recession in CUPS, suggesting that the use of NGS for profiling tumours, and potential advances in targeted therapy may improve prognosis (Varghese et al. 2017). In cases where the molecular signatures of a CUPS don't match those of the primary tumour, precision medicine approaches that target key oncogenic processes in a tumour may provide survival benefits when compared with cancer-specific therapy. Similarly, work in immunotherapy has recognized potential mutational signatures that are associated with responsiveness to immunotherapy. This includes mutational signatures of UV-radiation and tobacco smoke, which are present in a subset of CUPS (Varghese et al. 2017). Once more, this result suggests that the use of NGS for profiling, and advances in treatment may provide significant improvements for treating CUPS.

1.3 Determinants of mutation rate in somatic cells

Somatic cells are exposed to multiple mutational processes throughout their developmental history. Mutational processes can generate SNVs, copy-number variants (CNVs), structural variants (SVs) and, insertions and deletions (indels). Through these processes, genetic alterations in somatic cells over successive cell divisions. SNVs represent the best studied, and most abundant mutations in somatic cells, with the typical human cell containing thousands of somatic SNVs (Lee-Six et al. 2018; Martincorena et al. 2015; Brunner et al. 2019). The majority of SNVs are innocuous to somatic cells, with no noticeable impact on cellular fitness. As a consequence of having little effect on cell fitness, these mutations, often called passenger mutations, are passed on to descendent cells and record the mutational history of a cellular lineage. However, a small number of somatic SNVs are capable of providing a strong selective advantage (Martincorena et al. 2017). These mutations, termed driver mutations, are implicated in driving clonal expansions which can result in the development of cancer, and are implicated in human aging and neurodegeneration (Lodato et al. 2018).

Each mutational process generates characteristic changes within the genome, called mutation types. Mutational processes can generate multiple mutation types, and many mutational processes generate similar mutation types in varying proportions. Mutational signatures were constructed to detect the signal of mutational processes in the genome. Mutational signatures identify and group mutation types based on the mutational process responsible for generating them (Nik-Zainal et al. 2012a; Alexandrov et al. 2013; Alexandrov et al. 2020) (Figure 1.4). Regardless of the mutational process, all mutations are fundamentally the result of DNA damage that fails to be correctly repaired. Local features of the genome including DNA base composition, nucleosome occupancy, chromatin state and level of transcription affect both the ability for DNA damage to occur and the efficiency in which DNA damage is repaired (Sabarinathan et al. 2016; Polak et al. 2014; Supek and Lehner 2015; Tomkova and Schuster-Böckler 2018; Volkova et al. 2020). As these genome features vary heavily across the genome, the distribution of mutation rate is non-uniform across the genome (Stamatoyannopoulos et al. 2009; Hodgkinson, Chen, and Eyre-Walker 2012; Schuster-Böckler and Lehner 2012; Supek and Lehner 2015; Polak et al. 2015; Lee, Abd-Rabbo, and Reimand 2020). In this section, I will provide an overview of mutational signatures and their characteristic patterns in the genome, the relationship between mutational processes and cell-type, and on the impact of chromatin features on observed mutation rate.

1.3.1 Mutation types and mutational signatures

Multiple exogenous and endogenous mutation-generating processes contribute to mutations in somatic cells. Exogenous mutagens include several carcinogens such as UV radiation and tobacco smoke and include cancer therapies such as chemotherapy and radiation therapy (Pich et al. 2019; Behjati et al. 2016). Endogenous mutational processes range from spontaneous deamination of methylated bases to defects in DNA damage response pathways such as homologous recombination (Duncan and Miller 1980; Polak et al. 2017). Each of these mutational processes leaves characteristic patterns or footprints of mutations. These patterns are composed of different mutation types. The distribution of mutation types caused by mutational processes can be grouped into mutational signatures. By examining mutational signatures in NGS of somatic tissues, the exposure of somatic cells to different mutational processes can be determined.

Mutation types for single-nucleotide variants

The characteristic patterns that different mutational processes leave behind in the genome can be described by both the base-pair changes induced by the mutational process and short-range sequence context surrounding the mutation. Together, these characteristic changes makeup mutation types, and represent the footprint or base-composition spectra of mutational processes on the genome (Alexandrov et al. 2013; Alexandrov et al. 2020). Mutation types result from either the DNA damage generated by a mutational process, or differential efficiencies of DNA damage response pathways that repair the generated damage. An example of this can be seen when looking at the mutations generated by UV-radiation. DNA damage from UV-radiation can result in the formation of dimers of adjacent pyrimidine bases on the same DNA strand and preferentially create thymine-thymine dimers. Pyrimidine dimers

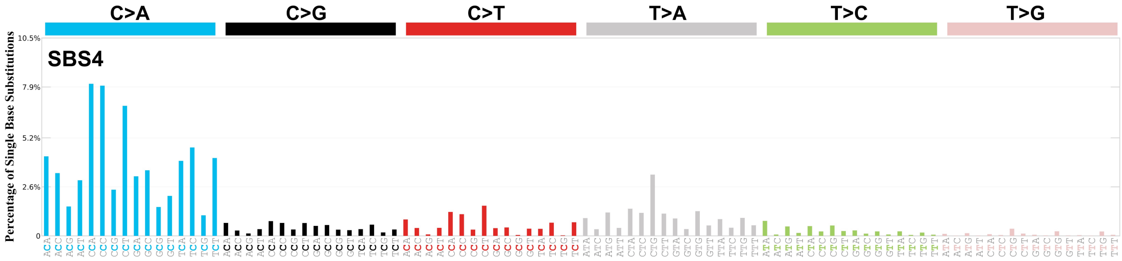


Figure 1.4: Mutational spectra for SBS4. The distribution of mutation types associated with single-base substitution signature 4 (SBS4). The aetiology of SBS4 is associated with exposure to tobacco. Figure from: (Alexandrov et al. 2020), licensed under Creative Commons Attribution 4.0 International License (`https://creativecommons-org/licenses/by/4.0/`).

are bulky lesions that can cause replication fork collapse if not repaired. To prevent replication fork collapse, low-fidelity translesion polymerases are recruited to the region. Repair of the pyrimidine dimer by low-fidelity polymerases often results in the production of DNA mismatches. In the case of UV-radiation, translesion polymerases tend to produce either cytosine to thymine or cytosine-cytosine to thymine-thymine changes (C >T or CC >TT) (Setlow and Carrier 1966; Waters et al. 2009). Another common environmental mutagen, tobacco, has a similar process of mutation generation. Carcinogens in tobacco smoke create bulky DNA adducts at guanines. Once again, low-fidelity translesion polymerases are used to address the bulky DNA lesions. In this case, the mismatches introduced during the repair process tend to be cytosine to adenosine (C>A) substitutions (Rodin and Rodin 2005; Wiencke 2002). In both of these examples, the mutational processes' characteristic patterns of substitutions are the mutation types associated with the mutational process. The presence of a specific mutation type can be used to assess the associated mutational processes' activity. For example, a tumour with a large number of mutation types associated with UV-radiation has evidence of greater exposure to UV-radiation.

Initial studies aimed to associate mutation types with specific mutational processes focused on examining the patterns of mutations found in highly mutated cancer genes, such as *TP53* (Hollstein et al. 1999; Hollstein et al. 1991). In addition to revealing associations between various environmental mutagens such as UV-radiation, aflatoxin and smoking with specific mutation types, these studies demonstrated the mutation types found in *TP53* varied significantly across different cancer types. This provided some initial evidence that mutational spectra can provide information about cell-type. For example, the *TP53* mutation types in skin carcinomas exhibited the characteristic pattern of UV-radiation described previously. In contrast, the *TP53* mutations in lung cancers from smokers tended to have the characteristic mutation type associated with tobacco smoke. Interestingly, not all mutation types initially studied in *TP53* displayed evidence of cell-type specificity. All of the cancer types analyzed in this early work had evidence of C >T mutations at CpG dinucleotides. As this mutation type is present across many cancer types, it is likely related to an endogenous mutational process active across most cell types. The C >T mutations at CpG dinucleotides likely result from the spontaneous deamination of 5-methylcytosine (Pfeifer 2006).

While initial studies focused on mutation types in some well-characterized cancer genes, advances in sequencing technologies enabled the analysis of mutation types in cancer genomes through NGS. Analyzing mutation types using data from WES or WGS has the distinct advantage of uncovering mutation types that are not necessarily biased by signals of positive or negative selection in the genome. Furthermore, using WGS and WES also allowed for examining the footprint of mutational processes across a wider array of nucleotide contexts. These studies started to reveal the diversity of mutation types across different cancer types. Also, they provided evidence that the activity of specific mutational processes can account for differences in mutation rate between tumour genomes. For example, studies in lung cancer demonstrated that tumours from smokers harboured upwards of a 10-fold increase in the overall number of SNVs, with much of the increase being directly attributed to mutation types associated with tobacco smoking (Govindan et al. 2012). Chemotherapy is associated with DNA damage, and studies examining WES and WGS of tumours exposed to treatment provide direct evidence for the mutation generating role of chemotherapy. Glioblastoma multiforme tumours treated with alkylating agents had elevated SNVs compared to non-treated tumours. Tumours treated with alkylating agents tended to have an elevated number of C>T mutations in various contexts (Parsons et al. 2008; Hunter et al. 2006). In addition to unveiling the diversity of mutation types associated with environmental or exogenous sources, early studies using NGS to study mutation types helped characterise mutation types associated with endogenous mutation generating processes. For example, studies in leukaemia revealed that the immunoglobulin genes had a large number of T >G transversions which has been attributed to somatic hypermutation associated with polymerase-η (Puente et al. 2011). Interestingly, many of these mutation types have specificity for a limited number of cancer types, suggesting that examining mutational processes may provide information about cancer type.

Mutational signatures are derived from patterns of mutation types

The study of mutation types using NGS data provided a significant advance in understanding the effect of mutational processes across cancer types. Despite significantly advancing the understanding of mutational processes, studying mutation types alone did not address the issue that arises when we consider that most somatic cells, particularly those found in cancer, are exposed to a combination of multiple mutational processes. This means that the observed mutation types in a tumour genome represent a superposition of many mutational processes that were active throughout the developmental history of the cells within the tumour. To address the challenge of examining mixtures of mutation types generated by many different mutational processes, the mutation types associated with a specific mutational process need to be separated from other mutation types and summarized based on common features. To do this, mutational signatures are constructed by decomposing the distinct patterns of mutation types in a set of genome samples.

Mutational signatures in the form described above were initially demonstrated by examining SNVs derived from WGS of 21 breast cancer patients (Nik-Zainal et al. 2012a). Mutation types were expanded by considering the short-range sequence context around every SNV (the base immediately $5'$ and $3'$ for every SNV). This results in a total of 96 mutation types (although this value can be extended by considering different ranges of sequence context). Including sequence context around SNVs greatly expands the number of mutation types compared to looking at only the substituted base. This increases the resolution at which mutation types can be examined, making it easier to differentiate mutational processes that may have similarities in the mutation types they produce. For example, temozolomide, deamination of 5-methylcytosine and UV-radiation tend to produce C >T mutations. Without considering sequence context around the single-base substitution, it is difficult to determine which of these three mutational processes generated each C >T mutation. The initial analysis provided mutational signatures associated with several endogenous processes. One of these highly relevant processes for breast cancer is the mutational signature associated with mutations in *BRCA1* and *BRCA2*. This mutational signature had a relatively uniform distribution over mutation types, but its activity could accurately discriminate

BRCA1/2 wild-type from mutant tumours (Nik-Zainal et al. 2012a) (Figure 1.5). The strong association of this mutational signature with *BRCA1/2* mutations formed the basis for the development of algorithms like HRDetect, which use patterns of mutations to identify if tumours have deficiencies in homologous recombination (Nik-Zainal et al. 2012b). Presence of homologous recombination deficiencies can be targeted with specific therapies, underscoring the clinical utility of examining mutational signatures in tumours (Davies et al. 2017).

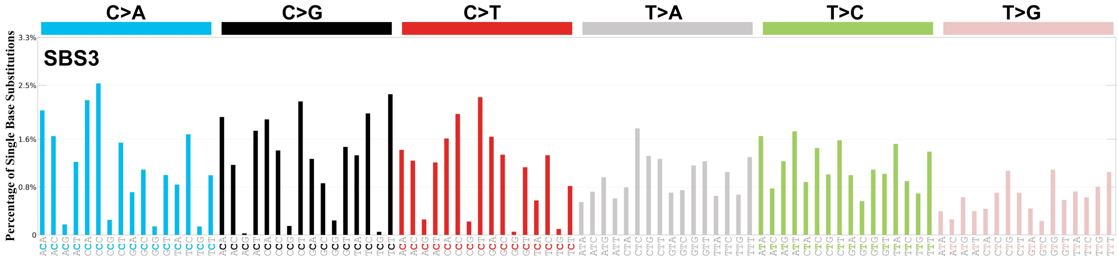


Figure 1.5: Mutational spectra for SBS3. The distribution of mutation types associated with single-base substitution signature 3 (SBS3). The aetiology of SBS3 is associated with defects in homologous recombination. This signature is often found in tumours that have BRCA1/BRCA2 mutations. Figure from: (Alexandrov et al. 2020), licensed under Creative Commons Attribution 4.0 International License (`https://creativecommons-org/licenses/by/4.0/`).

Mutational signatures were further extended by including information about the transcriptional strand on which an SNV resides (Alexandrov et al. 2013). This allowed for the resolution of mutational signatures to effectively be doubled, allowing greater ease when differentiating mutational processes with similar mutation types. For example, if a mutational signature contains the mutation type C >A on the transcribed strand and not the untranscribed strand, it provides evidence that this mutational signature may result from a transcription-coupled mutational process such as the recruitment of transcription-coupled nucleotide excision repair machinery (Fousteri and Mullenders 2008). The computational framework used for constructing mutational signatures in breast cancer has since been applied to multiple pan-cancer datasets, uncovering initially 30 and later at least 60 mutational signatures in total (Alexandrov et al. 2013; Alexandrov et al. 2020). By doing this in a pan-cancer fashion, mutational signatures that have relatively uniform activity across different cancer-types were discovered. Included in this set is the mutational signature for spontaneous deamination of 5-methylcytosine discussed previously. Interestingly, analysis of the burden of mutations contributed by this mutational signature demonstrated a correlation between the activity of this mutational signature and the age of the patient at diagnosis. The relatively large collection of tumour samples in this dataset allowed for increased resolution when deciphering mutational patterns. This work has demonstrated several mutational signatures likely associated with sequencing artefacts, which may be useful when trying to determine if a clinically actionable mutation is actually present or if its due to sequencing quality. This work has also uncovered multiple mutational signatures associated with APOBEC enzymes, consisting of several cytidine deaminases traditionally used to protect mammalian cells from viral infection (Swanton et al. 2015). Despite extensive study into the aetiology of mutational signatures, many mutational signatures have uncertain or unknown origin (Alexandrov et al. 2020).

While some early work uncovered mutational signatures associated with chemotherapy exposure, such as temozolomide treatment in glioblastoma multiforme, mutational signatures associated with treatment have been difficult to discover. This difficulty can be contributed to a limited number of post-treatment tumour genomes. Recent studies examining mutational spectra in post-treatment metastatic tumours has allowed for mutational signatures associated with chemotherapy to be uncovered. A recent study examining metastatic breast cancer demonstrates multiple treatments associated with mutational signatures (Angus et al. 2019). This study uncovered a novel mutational signature associated with cisplatin exposure, which characteristically results in CC >AA mutations. This study also showed that, regardless of treatment type, metastatic breast cancer samples sequenced following treatment had a significant enrichment for mutation signature 17. More recent work studying pan-cancer treatment associated mutational signatures in a large cohort of metastatic tumours has further expanded the mutational spectra associated with chemotherapy (Pich et al. 2019). This study describes four mutational signatures associated with three platinum-based drugs and two signatures associated with multiple drugs. Furthermore, this work showed that mutational signatures associated with platinum-based chemotherapy had transcriptional strand bias. This work also uncovered a novel signature associated with exposure to fluorouracil and capecitabine. Interestingly, the mutational signature associated with exposure to fluorouracil and capecitabine displayed a mutational footprint highly similar to that of mutational signature 17b, a signature of unknown aetiology, but was thought to be associated with oxidative damage to DNA. The identification of therapy associated signatures also allowed for the mutational burden of chemotherapy to be assessed. On average, exposure to chemotherapy contributed thousands of mutations to exposed tumours, but this value varies based on the patient's tumour type. Overall, the percentage of mutations resulting from chemotherapy ranged between 1% and 65% of all mutations within a tumour, suggesting that both the mutation types observed in a tumour and the overall tumour mutation burden can be strongly impacted by exposure to chemotherapy. Additional work for uncovering mutational spectra associated with chemotherapy has been done with experimental models. This work provides evidence for additional mutational signatures associated with six chemotherapy agents. It has also provided experimental evidence for mutational signatures associated with a large range of environmental agents (Kucab et al. 2019).

1.3.2 Association between mutation rate and chromatin-features

Mutations are the result of the interplay between DNA damage generation and DNA damage repair. DNA damage or nucleotide mismatches are incorporated into DNA. DNA damage repair mechanisms then address the presence of DNA damage or mismatches. This results in either faithful repair of the lesion/mismatch or the introduction of a mutation. Each step of this process may be impacted by local genomic features including chromatin state, nucleosome occupancy, transcription factor binding, level of transcription and long-range sequence context. As these local features vary across the genome, mutation rate is non-uniform across the genome. In fact, different regions of the genome vary by up to five-fold somatic mutation density (Lawrence 2013).

Determinants of mutation rate at the megabase scale

Large-scale WGS of cancer has enabled the study of mutation rate variability across the genome. Early work demonstrated significant variability in mutation rate across the genome at the megabase scale (Hodgkinson, Chen, and Eyre-Walker 2012). Early work also systematically examined the association between mutation rate and chromatin features focusing on correlating SNV density in cancer genomes with genomic features including, base composition, CpG content, gene density, DNA replication timing, nucleosome occupancy, and levels of histone acetylation (Schuster-Böckler and Lehner 2012). This study found that cancer SNV density was associated with features describing chromatin organization and structure at the megabase scale. The strongest positive relationship was observed between SNV density and H3K9me3, a repressive histone modification, which is a mark of inaccessible chromatin. Similar, albeit weaker, correlations between SNV density and genomic features were observed for several repressive chromatin marks. Gene density, early replication timing and histone marks associated with open chromatin, all marks of accessible euchromatin-like domains, were shown to have a negative correlation with SNV density. Overall, these results provided evidence that mutation rate is strongly associated with chromatin organization, and in particular, SNV rate is highest in heterochromatin.

Later work demonstrated that these differences in mutation rate between heterochromatin and euchromatin are explained partly due to the differential efficiency of DNA mismatch repair across the genome. Mismatch repair mechanisms tend to have greater activity in euchromatic, early-replicating regions, leading to fewer observed mutations (Supek and Lehner 2015). The differential efficiency of mismatch repair across the genome is thought to result from multiple factors, including its coupling to DNA replication and differences in DNA accessibility to the repair machinery. A similar result has been demonstrated for nucleotide excision repair and base excision repair. Both tend to have lower efficiency in heterochromatic regions due to differences in DNA accessibility to repair machinery. Fully assembled nucleotide excision repair machinery, as an example, occupies approximately 100bp, which is larger than the length of linker DNA between nucleosomes (Polak et al. 2014). This results in an inability for repair machinery to be assembled in condensed chromatin, hindering repair functionality. Similarly, base excision repair complexes preferentially assemble in euchromatic regions, leading to decreased efficiency of DNA damage repair in heterochromatin (Amouroux et al. 2010).

The relationship between chromatin state and mutation density suggests that mutation rate may directly provide information about chromatin accessibility. The plasticity in cancer can result in significant chromatin heterogeneity in a single tumour. Furthermore, cellular plasticity in cancer can result in chromatin state shifting significantly as a tumour develops (Gomes et al. 2019). Consequently, the chromatin state of a tumour may differ from the cell of origin for that tumour. As the mutations in a tumour record the mutational history of that cell lineage, the regional mutation density of a tumour may contain mutations that match the chromatin state of the cell of origin for that tumour.

The relationship between regional mutation density and cancer cell chromatin state was investigated by computing the correlation between SNV rate in 1Mb bins and chromatin accessibility (Polak et al. 2015). In this study, DNAse I hypersensitivity data was used to determine chromatin accessibility. Using chromatin accessibility data from matched cancer cell lines and the putative normal cell-of-origin for the cancer sample in question. This study demonstrated that SNV density most strongly correlates with chromatin features from the normal cell-of-origin for the tumour (Figure 1.6).

A potential explanation for mutation density being more strongly associated with cell-of-origin chromatin accessibility than cancer cell chromatin accessibility is that passenger mutations accumulate in somatic cells' normal life history before malignant transformation. Based on this result, the regional mutation density of a tumour genome contains information associated with the pre-malignant state of the tumour. Consequently, regional mutation density provides information about ancestral cell-states. Using this fact as intuition, recent work has demonstrated that SNV density can be used as a record of pre-malignant state to infer the putative normal cell-of-origin for many cancer types (Kübler et al. 2019).

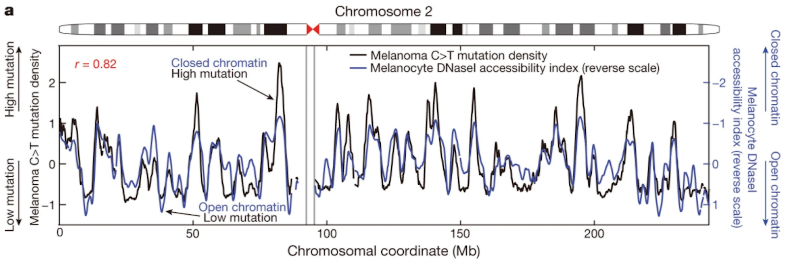


Figure 1.6: The association between mutation density and chromatin accessibility. Reverse scale chromatin accessibility in 100-kb windows(blue line) assessed with DNAse I hypersensitivity (high values correspond to less accessible chromatin) from normal melanocytes compared to the number of C >T mutations aggregated across multiple melanoma samples. Figure from: (Polak et al. 2015), licensed under Creative Commons Attribution 4.0 International License (`https://creativecommons-org/licenses/by/4.0/`).

1.4 Machine-learning as a tool for tumour-typing

The increased volume of data generated by NGS of tumour genomes and imaging analysis has presented the opportunity to exploit patterns found in these data to identify cancer type. To process and mine these rich data, statistical and machine learning methods have been used to model the data, discover patterns, provide diagnostic predictions and generate biological insight. These models can be used to make predictions of cancer type, and can be used to assist pathologists in making cancer diagnoses, and have the potential for providing useful insight into cancer biology. For example, Su et al. described one of the first methods for identifying cancer type using supervised machine learning approaches, which could differentiate between 11 cancer types using gene expression data (Su et al. 2001).

There is significant diversity amongst machine learning algorithms. Some methods, such as deep neural networks (discussed in detail later), consist of compositions of linear transformations and non-linear activations. Other models, such as random forest classifiers (discussed in detail later), use binary decision trees to learn decision boundaries for performing classification.

Regardless of the method being used, however, some commonalities exist amongst most machine learning methods. First, all methods require data. Data for machine learning methods can be split into training data, validation data and test data. Training data are the data being used to teach or train the machine learning model. These data are used to determine the function that maps from the input space to the output space, and are not used for evaluation. Validation data are data samples used for tuning non-learnable parameters of the model, sometimes called hyperparameters (discussed in detail later). Test data are the data examples used for evaluating model performance. Test data should be held-out from training and validation procedures, and the machine learning model should not receive any information about test examples. Second, all machine learning methods consist of some function or model that aims to model or describe the relationship between the input features and the output or target space. As mentioned above, the function being used can be a very simple function or arbitrarily complex, as is the case with deep neural networks. Third, all machine learning models require a function that can assess how well the model is doing at approximating the output as a function of the input features. This is referred to interchangeably as an objective function, loss function or cost function. A well-performing model will typically find a value of the objective function that lies in the vicinity of an optima of the objective function when evaluated on training data. Finally, all machine learning methods require some optimization procedure which uses the features and the corresponding output targets to optimize the objective function. This optimization proceeds by altering the learnable parameters of the model to either maximize or minimize the objective function. The objective function can also be used to determine if a model has overfit to the training data. When a machine learning model overfits, it learns to model the noise associated with training data in a way that does not generalize to different datasets. In this scenario, the model has memorized the training data and will produce very low error on training data, but will have failed to learn a sufficient amount of information regarding the true signal in the data, which will result in significantly higher validation or test error (Goodfellow, Bengio, and Courville 2016).

Machine learning has extensively been applied for predicting cancer type based on tumour genomics. Generally, this involves using NGS data and/or imaging, and training a classifier to predict cancer type from a pre-defined set of cancer types that the model is trained to identify. Some of the earliest classification systems are trained on data from RT-PCR or expression microarrays, and have either limited overall accuracy or are only trained to identify a small number of cancer types (Ma et al. 2006; Bender and Erlander 2009). More recently, classifiers have been trained on data generated by RNA sequencing, gene panels and WGS (Penson et al. 2019; Grewal et al. 2019; Yuan et al. 2016; Salvadores, Mas-Ponte, and Supek 2019). A variety of different algorithms have been employed for identifying cancer type, including support vector machines, random forests and artificial neural networks. The performance of these classifiers varies based on the number of cancer types being identified, and the best performing model has an overall accuracy of 92% for discriminating between 18 cancer types (Salvadores, Mas-Ponte, and Supek 2019). Notably, the wide range of features used for training models suggests that features derived from a wide range of sequencing modalities carry cancer-specific information. Unfortunately, large cohorts containing data from multiple modalities do not currently exist.

Despite efforts towards machine-learning-based tumour typing, it is unclear whether machine learning models can discriminate between a relatively large number of cancer types with strong predictive performance. In Chapter 2, I attempt to address this by building a predictive model of cancer type. In this section, I provide a brief overview of random forests, a commonly used supervised learning method for tumour typing, a review of two commonly used hyperparameter optimization methods, Bayesian optimization and k-fold cross-validation, and finally, a review of feed-forward deep neural networks, a class of machine learning models which I used for developing my models.

1.4.1 Random forests

Random forests is a machine learning method that models data using a collection of binary decision processes or decision trees (Breiman 2001). A decision tree aims to partition or split the data based on a set of features or variables. For example, a decision tree that aims to partition images of polar bears and black bears, may partition the dataset based on features such as size and colour. For classification, the optimal split is determined by the degree to which data is separated into different classes. The Gini impurity index (G) can be used as a tool to assess how well the classifier has partitioned the data. Formally, the Gini impurity index is defined as follows:

$$G = \frac{n_l}{N} \sum_{i=1}^{C} p_{iL}(1 - p_{iL}) + \frac{n_R}{N} \sum_{i=1}^{C} p_{iR}(1 - p_{iR}) \tag{1.1}$$

where N is the number of examples, C is the number of classes, n_L and n_R are the number of examples in the left and right child nodes, respectively, and p_{iL} and p_{iR} are the fraction of examples of class i in each child node of the decision tree. Intuitively, the Gini impurity index is the weighted average between the two daughter nodes in the decision tree.

A random forest model often consists of many decision trees, where each tree makes a prediction given new data. A random forest model is considered an ensemble method. The ensemble's output or prediction is the average or majority vote of the many decision trees in the ensemble. To create variability in each decision tree's predictions in the ensemble, each decision tree is given different data. Datasets for each decision tree are generated through a process called bootstrap aggregation, or bagging. Bootstrapping is a procedure that generates a dataset by randomly sampling the training data with replacement. For random forests, bootstrapping is supplemented by choosing random subsets of features during each bootstrap instance. The bagging procedure described here produces variability in each dataset. As each decision tree gets a different dataset to train on, each decision tree can learn different decision rules, providing variability to the random forest. This procedure helps prevent overfitting of the model, which occurs when models are trained to memorize the training data, and fail to generalize well to new data.

A useful feature of random forests is the ability for random forests to measure feature importance. Feature importance measures the predictive value of a feature in the data. It allows for the relative contribution to the predictive power of each feature to be assessed. This allows the model to be interpretable because the decisions the model makes can easily be followed using feature importance values. This can allow for important features to be further examined, which has the potential to reveal important biological characteristics of the system being studied.

1.4.2 Deep neural networks

Deep learning refers to computational models that consist of multiple processing layers that learn hierarchical representations of the data. The use of multiple processing layers allows deep learning models to learn representations of the data at multiple levels of abstraction (Goodfellow, Bengio, and Courville 2016). Deep learning overcomes the problem of carefully engineered feature sets by learning directly from data with limited or no feature engineering. Rather than explicitly doing feature engineering, deep learning models employ architectures that are well-suited for exploiting the inductive biases implied by the domain (Wilson and Izmailov 2020). Deep neural networks use a succession of simple, non-linear transformation of the data to transform input data into increasingly abstract representations. Through this process, deep learning models use simple transformations to learn highly complex functions that map input data to the output space. The past decade has seen deep-learning-based approaches significantly outperform traditional machine learning methods in various tasks, including speech and image recognition, and text generation (Devlin et al. 2019; He et al. 2015a). Deep learning methods have also demonstrated success in identifying tumour type, with several methods showing performance comparable to pathologists when identifying cancer type from imaging data (Yoon et al. 2019; Esteva et al. 2017).

Deep feed-forward networks or multilayer perceptrons are the quintessential deep learning models (Figure 1.7). The goal of a feed-forward network is to learn a function f^*. In the case of a classifier, $y = f^*(x)$ is a function that outputs a category or class, $y \in K = \{1, 2..., k\}$ for any input value, x. A neural network defines a mapping $y = f(x; \theta)$, and learns the parameters θ that provide the best approximation. These models are called feed-forward because information flows from the input to the output through intermediate transformations used to define f. Feed-forward networks lack feedback connections through which information from y can flow backwards through the model (Goodfellow, Bengio, and Courville 2016). Feed-forward neural networks form the basis for neural networks used for several tasks, including neural networks used for learning generative probabilistic models, such as variational autoencoders and generative adversarial networks, and transformers, which show state of the art performance in natural language processing (Kingma and Welling 2014; Goodfellow et al. 2014; Vaswani et al. 2017).

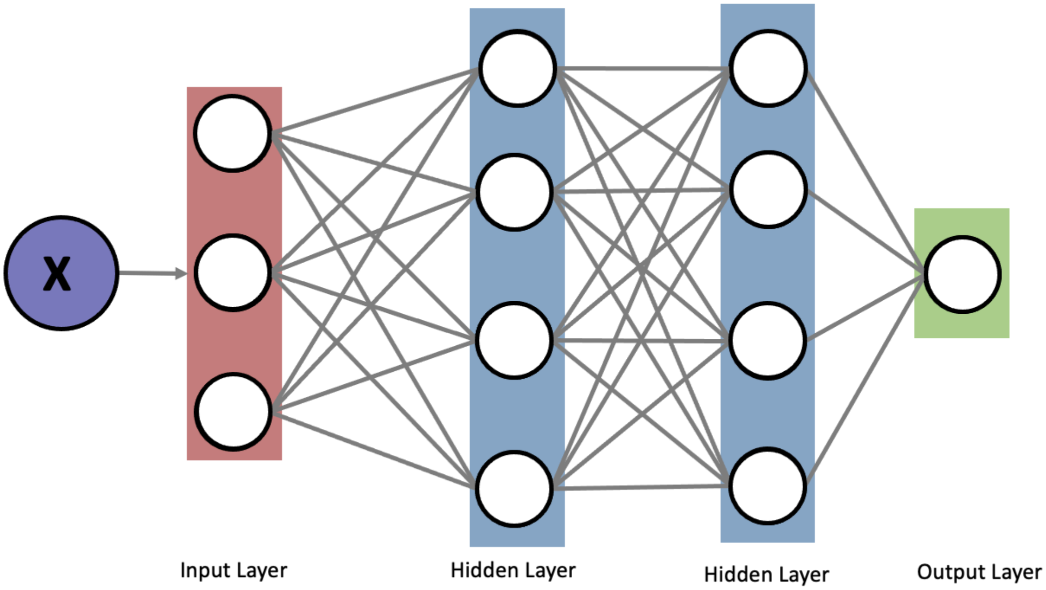


Figure 1.7: A deep multilayer perceptron. A schematic showing a deep multilayer perceptron or feed-forward neural network. Input signal or data enters through the input layer. The hidden layers perform non-linear transformation of the data, which allow for abstract representations of the data to be learned. The last layer of the feed-forward network is responsible for providing the model's outputs.

Feed-forward neural networks are referred to as networks as they are the composition of many different functions. The functions that compose a neural network form a chain such that the output of one function is the input or argument for the next. For example, $f^*(x) = f^{(3)}(f^{(2)}(f^{(1)}(x)))$ is a function representing a 3 layer network composed of the chaining of $f^{(3)}$, $f^{(2)}$, and $f^{(1)}$. In this example, $f^{(1)}$ is called the first layer of the network, $f^{(2)}$ is the second layer of the network, and more generally, $f^{(j)}$ is the j^{th} layer of the neural network. The depth of a neural network is the length of the chain of layers that compose the neural network. The final layer of a neural network is referred to as the output layer.

During training, data provide noisy approximates of $f^*(x)$ that are evaluated at different training points. Each data point is accompanied by a label or target value, y, and training aims to match the network's output with the target, y. The learning algorithm alters the functions in the intermediate layers to produce the desired output. Instead of explicitly defining the functional form of the intermediate layers, the learning algorithms allow the intermediate layers' functions to be determined by the data during model training. The intermediate layers in a neural network are sometimes referred to as hidden layers. In contrast to an output layer, the output of a hidden layer is not directly accessed, but is instead, used as the input for a subsequent layer.

Each layer of a neural network contains at least one, and typically many, "neurons" or "units". The number of neurons in a hidden layer specifies the width of a neural network. Each neuron receives a vector-valued input consisting of the neurons from the previous layer that it forms connections with. As such, each layer of a neural network is composed of many vector-to-scalar units, which takes in a vector of outputs from the previous layer, and outputs a scalar value.

A convenient way to understand feed-forward networks is to think of them as in the context of linear models, such as linear regression and logistic regression. In logistic regression, the output of the model is defined as:

$$\hat{y} = \sigma(\theta^T x) \tag{1.2}$$

Where σ is a sigmoid function, and θ is a vector of coefficients or weights. As with neural network classifiers, logistic regression models are trained to match predictions, $\hat{y}$ with true labels, y. These models can be extended to represent increasingly complex, nonlinear functions of the inputs by applying a nonlinear transformation to the inputs:

$$\hat{y} = \sigma(\theta^T \phi(x)) \tag{1.3}$$

Where ϕ is a nonlinear transformation. ϕ provides a new representation of the input features x While there are many ways to pick ϕ, in deep learning, ϕ is learned from the data. In this approach, we have:

$$y = f(x; \lambda, w) = \phi(x : \theta)^T w \tag{1.4}$$

Where λ are parameters that we use to learn ϕ, and w are parameters that map from $\phi(x)$ to the desired output. While this approach depends heavily on learning ϕ from the input data, domain-specific knowledge can be encoded to restrict the function learned in a way that is expected to improve performance. In general, this can be written as the output of the first layer as follows:

$$h^{(1)} = f(W^{(1)T} x + b^1) \tag{1.5}$$

and for any arbitrary layer as:

$$h^{(l)} = f(W^{(l)T} h^{l-1} + b^l) \tag{1.6}$$

In these equations, W represents a matrix of learned parameters or weights, b is a bias or offset value, and f is an arbitrary non-linear transformation. Prior to model training, W and b are typically initialized randomly (Glorot and Bengio 2010).

The non-linear transformations that compose hidden layers and the output layers of a neural network are formed by applying activation functions to a neural unit's inputs. Many activation functions exist, with some functions being more useful for output layers, and others tending to be used for hidden layers. The choice of activation function for the hidden layers can affect model performance and training dynamics. One of the most commonly used activation function for hidden layers is the rectified linear activation function which produces rectified linear units (ReLU):

$$f(x) = max(0, x) \tag{1.7}$$

Rectified linear activation functions are standard defaults that have shown performance improvements when compared to alternatives. This function is differentiable everywhere except 0 and has several computationally desirable properties. Some generalizations to the rectified linear function exist which aim to help improve training dynamics. A commonly used example is the softplus function:

$$f(x) = ln(1 + e^x) \tag{1.8}$$

A less commonly used activation function that has recently been adapted for certain tasks is the radial basis function:

$$f(x) = \exp\left(\frac{-1}{\sigma_i^2} \left\|(W_{:,i} - x)\right\|^2\right) \tag{1.9}$$

The radial basis function is a commonly used kernel metric that measures the similarity between two input arguments. The application of this activation function in modern deep learning tasks will be touched on in a later section. Activation functions used on the output layer are one of the ways in which the task of a neural network can be defined. For the purposes of classification, the softmax function is the most commonly used output activation function. The softmax function takes as input a logit vector which is a vector consisting of the unnormalized log probabilities for each class:

$$z = W^T h + b \tag{1.10}$$

Where h is the output of the previous hidden layer, and b is the bias or offset value, and z_i represents the unnormalized log-probability that an input sample belongs to class i. The softmax function can convert the unnormalized log probabilities into class-specific probabilities. It does so by both exponentiating and normalizing the elements of the logit vector. The softmax function is defined as follow:

$$softmax(z)_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)} \tag{1.11}$$

The output of a $softmax$ function is a class probability vector which can be interpreted as a categorical distribution over K classes. The use of a $softmax$ function for classification allows for neural networks to be interpreted as probabilistic classifiers. For an input sample x, a neural network will output a class probability vector $\hat{f}(x) = (\hat{f}_1(x), \hat{f}_2(x), ..., \hat{f}_k(x))$ where $\sum_{k=1}^{K} \hat{f}_i(x) = 1$ and $\hat{f}_i(x) \geq 0$. The elements of the class probability vector represent the probability that an input x belongs to each of the K classes that the classifier is trained to identify. The input x is then assigned to the class with the largest probability. This value is also called the confidence value.

To train a neural network, an appropriate cost or loss function must be selected. The cost function is typically selected based on the data generating process that the network is modelling, and then derived based on principles of maximum likelihood estimation. In principle, this means that the cost function of a neural network is the negative log-likelihood of the data:

$$J(\theta) = -\mathbb{E}_{x,y \sim p_{data}}[\log p_{model}(y|x)] \tag{1.12}$$

Neural networks are typically trained using a combination of two algorithms: backpropagation and stochastic gradient descent. Recall that information flows forward from input x to the outputs $\hat{y}$. This is called forward propagation. During training, each forward pass produces a value for $j(\theta)$. To allow information to flow backwards from the cost through the rest of the network, the backpropagation algorithm is used (Rumelhart, Hinton, and Williams 1986). The backpropagation algorithm is an algorithm for efficiently computing the gradient of the cost with respect to the network's inputs. Stochastic gradient descent is then used to perform learning by using gradient information to update the neural network's weights. The weights of the neural network are updated along the direction of the gradient so that the updated weights would reduce training loss on the same input data. An overview of backpropagation is provided by Rumelhart (Rumelhart, Hinton, and Williams 1986), and an overview of gradient descent is provided by Ruder (Ruder 2017).

1.4.3 Hyperparameter optimization

Neural networks, like most machine learning methods, have several hyperparameters. Hyperparameters are typically related to architectural choices (the type of hidden layer, activation function, number of hidden layers, number of units in a hidden layer), and related to regularization methods and training. Architectural choices are often made based on prior knowledge of the structure or inductive biases present in the data. For example, if the input data to a model consist of images, neural network architectures that are well suited to capturing the structure of images, such as convolutional layers are often employed. By using domain-specific knowledge to inform architectural choices, overall performance and generalizability can be increased (Wilson and Izmailov 2020).

Multiple methods for optimizing or selecting the best hyperparameters exist. Regardless of the method used, k-fold cross-validation (CV) is commonly used in the optimization procedure to evaluate model performance (Figure 1.8) (Hastie, Tibshirani, and Friedman 2009). In CV, training data are randomly segregating into k partitions or folds of the data. Each $k-1$ partitions are used as training data to build a model with the specified hyperparameters. The trained model is then evaluated on the held out partition. This procedure is then repeated for each partition, and for each setting of the hyperparameters. Performance of the model as a function of hyperparameters is taken to be the average of the performance on each of the k held out partitions (Figure 1.8). Typically, the accuracy or the value of the loss function are used as metrics to evaluate model performance.

Neural networks have a large number of hyperparameters, and training of neural networks is computationally expensive. This means that simply performing a grid search over all possible combinations of hyperparameters is not feasible except in very limited scenarios. To counteract this, Bayesian optimization (BO) is often used for hyperparameter optimization (Snoek, Larochelle, and Adams 2012). BO is a sequential, model-based optimization method that aims to perform global optimization with a minimum number of trials. BO involves two models or functions: a Bayesian probability surrogate model, which is used to model the objective function (in this case, the performance of a neural network as a function of the hyperparameters), and an acquisition function which is used to determine the set of hyperparameters to sample for the next trial. The algorithm is as follows: A prior distribution of the surrogate model is built; the acquisition function used to sample a set of hyperparameters; the neural

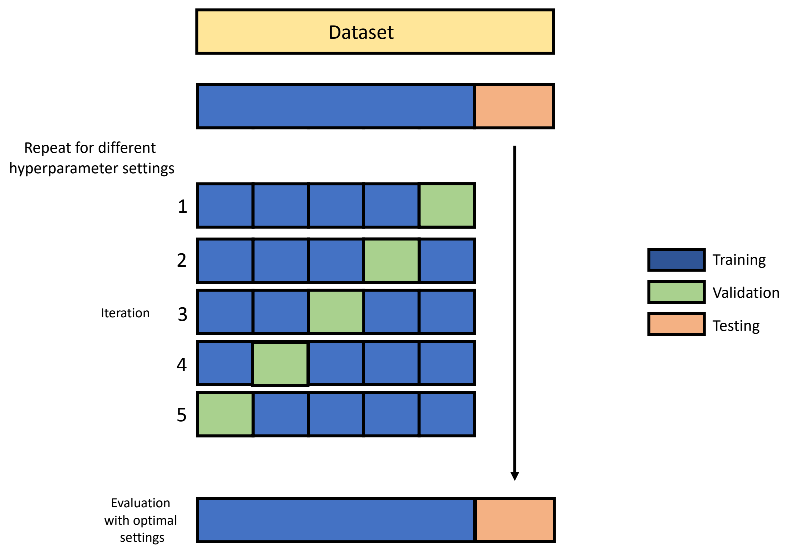

Dataset
Repeat for different hyperparameter settings
1
2
Iteration 3
4
5
Evaluation with optimal settings
Training
Validation
Testing

Figure 1.8: *k*-fold cross-validation. Schematic illustrating the cross-validation procedure. Training data is randomly split into k partitions ($k = 5$ in this example). Each partition is a held-out validation set for a model trained on the other $k - 1$ partitions. This procedure is repeated for all combinations of hyperparameter settings being assessed. A final model is then trained with the optimal hyperparameters, and performance is evaluated on an independent test set that was not used during cross-validation.

network is trained and evaluated with the sampled hyperparameters; the posterior distribution of the surrogate model is computed. This process is repeated until the optimal value is found, or until the number of trials reaches its limit (Figure 1.9).

Many probability models have been used as surrogate models in BO. Gaussian processes are the most commonly used surrogate probability models in BO. A Gaussian process describes a stochastic process where any finite subset of random variables $x_1, x_2, ...x_n \in X$ jointly follows a Gaussian distribution. Gaussian processes are capable of approximating any Lipschitz continuous function arbitrarily well, and therefore, can approximate any smooth function. Gaussian processes are Bayesian non-parametric models, and as such, the number of parameters in a Gaussian process does not need to be decided beforehand. It is determined by the dataset size and inductive biases in the data. For a Gaussian process, the kernel function must be specified beforehand. The kernel function describes the covariance of the Gaussian process random variables, and, together with the mean function, completely describes a Gaussian process. The choice of kernel function determines most generalization properties of a Gaussian process. For hyperparameter optimization, the Matérn (5/2) is widely used due to its flexibility and learnable hyperparameters. In addition to choosing a kernel function, an acquisition function must be chosen for Bayesian optimization. The acquisition function is used to determine the next set of hyperparameters to sample. This function needs to weigh exploring the space of possible hyperparameter combinations and moving towards hyperparameters that will improve the model. The most commonly used acquisition function is the expected improvement algorithm, which tries to find regions of hyperparameter space that will, on average, lead to improved model performance. More recently, acquisition functions that

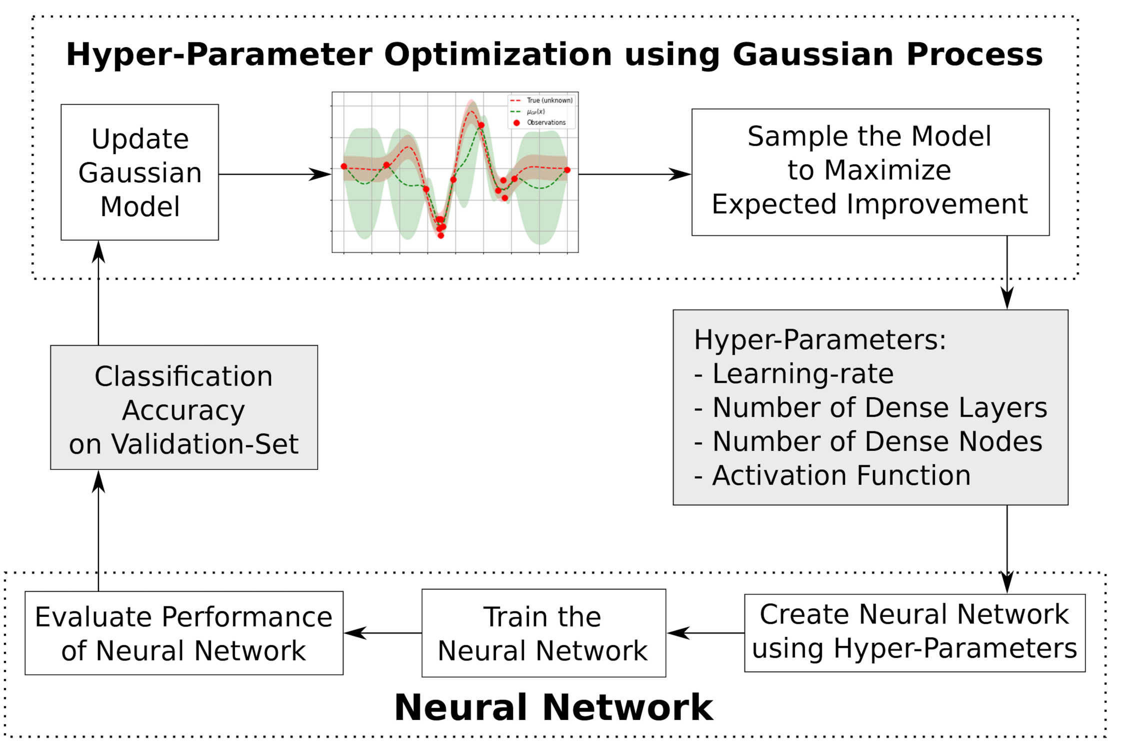

Hyper-Parameter Optimization using Gaussian Process
Update Gaussian Model
True (unknown)
$\mu_{GP}(x)$
Observations
Sample the Model to Maximize Expected Improvement
Hyper-Parameters:
- Learning-rate
- Number of Dense Layers
- Number of Dense Nodes
- Activation Function
Classification Accuracy on Validation-Set
Evaluate Performance of Neural Network
Train the Neural Network
Create Neural Network using Hyper-Parameters
Neural Network

Figure 1.9: Bayesian hyperparameter optimization. Schematic illustrating Bayesian optimization for neural network hyperparameters. Neural network performance is approximated with a Gaussian process model. Hyperparameters are sampled from the Gaussian process by maximizing an acquisition function. Hyperparameters are used to train and evaluate a neural network. Model performance on a held-out validation set is then used to update the posterior distribution of the Gaussian process. This procedure is repeated iteratively until an optimal set of hyperparameters is found. Figure from: (Pedersen 2020)

use a combination of multiple functions have been shown to offer improved performance on some tasks (Brochu, Hoffman, and De Freitas 2011).

1.5 Quantifying uncertainty in deep neural networks

Deep neural networks provide highly accurate predictions on several machine learning tasks. In real-world decision-making systems, such as identifying cancer type, robust estimates of predictive uncertainty must accompany highly accurate models. However, modern deep neural networks are poor at providing uncertainty estimates and tend to be overconfident in their predictions (Guo et al. 2017). In cost-sensitive scenarios, such as the ones encountered when physicians use the predictions of a neural network to make cancer diagnosis, overconfident predictions may lead to adverse outcomes. For example, an overconfident but inaccurate prediction for a specific cancer type may form the basis for therapy targeting the incorrect prediction, which may be ineffective or harmful. In these scenarios, it is crucial to accurately represent a model's predictive uncertainty to estimate the reliability of a model's predictions.

When quantifying uncertainty, it is important to recognize two sources of uncertainty in deep learning models. The first, called epistemic uncertainty, is the uncertainty that exists in the parameters of the model. This uncertainty should ideally be high for out-of-distribution data points. A second source of uncertainty is called aleatoric uncertainty. This is the uncertainty that is inherent in the data. For example, an image of 3 is similar to an image of 8. In this case, there is aleatoric uncertainty making it difficult to differentiate between these two samples (Amersfoort et al. 2020). The uncertainty in a model's predictions often consists of a combination of both aleatoric and epistemic uncertainty.

Predictive uncertainty in machine learning can refer to multiple distinct notions of uncertainty. One quantity could be in-distribution uncertainty. Sometimes referred to as model calibration or confidence calibration, this refers to how well the uncertainty in the model's predictions reflects the true uncertainty for classes the model has been trained to identify (Guo et al. 2017). Put differently, the uncertainty in the model's prediction for an input sample should reflect the ground truth correctness likelihood. A second notion of uncertainty involves a machine learning model's ability to identify whether an input sample belongs to a class that the model is not trained to classify (Lakshminarayanan, Pritzel, and Blundell 2017). For example, if a classifier is trained to differentiate between images of dogs and images of cats, can it automatically determine if an input image is that of an owl? This is referred to as out-of-distribution detection, and it also has implications for deploying machine learning systems in clinical settings.

The majority of the work adapting deep neural networks to represent uncertainty involves probabilistic methods and has focused around Bayesian deep learning. In the Bayesian formalism, a prior distribution upon a neural network's parameters (weights) is specified. Given both a neural network architecture (as the likelihood) and training data, a posterior distribution over the neural network weights can be computed. Weights are then sampled from the posterior distribution multiple times. Each set of sampled weights can be used to make a prediction from the neural network. The collection of predictions is used to derive a predictive distribution that can represent the predictive uncertainty of the network (Neal 1996). The complexity of neural network functions means that exact Bayesian inference is not tractable. This has prompted the use of many approximations such as Laplace approximation, Markov chain Monte Carlo methods, and several variational Bayesian methods (Neal 1996; MacKay 1992; Blundell et al. 2015).

Despite these advancements, Bayesian deep neural networks have several computational difficulties and tend to struggle to produce highly well-calibrated uncertainty estimates. The relatively poor performance of Bayesian neural networks for providing well-calibrated uncertainty estimates is, in part, a result of the simplifications required to make these models computationally tractable. Variational Bayesian methods and Laplace approximations, for example, typically learn a unimodal posterior distribution, which may not properly model the true underlying uncertainty (Fort, Hu, and Lakshminarayanan 2020). While Bayesian methods have been the traditional focus for assessing uncertainty in deep learning, more recently, several non-Bayesian methods have been developed which make use of ensembling approaches for deriving predictive distributions, stochastic deactivation of neurons at test time, and methods that are based on deterministic notions of predictive uncertainty (Gal and Ghahramani 2016; Lakshminarayanan, Pritzel, and Blundell 2017).

In Chapter 3, I develop a number of deep learning models that can address both notions of uncertainty discussed above. In this section, I will briefly provide an overview of assessing and improving model calibration and an overview of using the output of neural networks for detecting out-of-distribution samples.

1.5.1 Model calibration

A probabilistic classifier such as a deep neural network is well-calibrated if the predicted class distribution is approximately equal to the true class distribution (Guo et al. 2017). Typically, deep neural networks produce poorly calibrated output probabilities, yielding overly confident predictions. For this overview, I will focus on calibration in the multiclass classification setting, where the classification function is a neural network. To evaluate the calibration of a neural network, several commonly used metrics have been introduced. These evaluation metrics can assess how well a model is calibrated at different scales.

The strictest notion of calibration is multiclass calibration (Kull et al. 2019). A multiclass-calibrated classifier is perfectly calibrated for every single class that the model is trained to classify. More formally, Consider a neural network classifier $\hat{f} : X \to \Delta_k$ that outputs probabilities for k classes. For any input $x \in X$, the classifier outputs a class probability vector $\hat{f}(x) = (\hat{f}_1(x), \hat{f}_2(x), ..., \hat{f}_k(x))$ belonging to $\Delta_k = \{(q_1, q_2, ..., q_k) \in [0,1]^k | \sum_{i=1}^{k} q_i = 1\}$ which is the $(k-1)$-dimensional probability simplex over k classes.

Definition 1 *A probabilistic classifier $\hat{f} : X \to \Delta_k$ is multiclass-calibrated if for any prediction vector $q = (q_1, q_2, ..., q_k) \in \Delta_k$, the proportions of classes among all possible $x \in X$ getting the same predictions $\hat{f}(x) = q$ are equal to the prediction vector q:*

$$P(Y = i|\hat{p}(x) = q) = q_i \; for \; i = 1, ...k. \tag{1.13}$$

A necessary condition for obtaining a multiclass-calibrated classifier is for the classifier to be calibrated for all individual classes (Kull et al. 2019). That is, for any given class, the classifier is perfectly calibrated. Formally, a classwise-calibrated classifier is as follows:

Definition 2 *A probabilistic classifier $\hat{f} : X \to \Delta_k$ is classwise-calibrated if for any class i and any predicted probability q_i:*

$$P(Y = i|\hat{f}(x) = q_i) = q_i \tag{1.14}$$

The notion of calibration that is typically of concern is confidence calibration. When neural networks make predictions, an input x is assigned to the class with the largest element in the output class probability vector. This value is referred to as the model's confidence. A classifier is confidence-calibrated if, for all instances where the confidence is predicted to be c, the expected accuracy of the classifier is c. Formally, a confidence-calibrated classifier is defined as follows:

Definition 3 *A probabilistic classifier $\hat{f} : X \to \Delta_k$ is confidence-calibrated, if for any $c \in [0,1]$:*

$$P(Y = argmax(\hat{f}(x))|max(\hat{f}(x)) = c) = c \tag{1.15}$$

As classifiers are learned from finitely many data with varying levels of noise and uncertainty. This means that, in practice, perfectly calibrated classifiers are not possible. To assess the calibration of a model, many metrics have been introduced. One important metric is the Expected Calibration Error (ECE) (Guo et al. 2017). ECE is the average difference between a model's confidence and accuracy. This is defined as follows:

$$ECE = \mathbb{E}_{\hat{f}}[|P(Y = argmax(\hat{p}(x))|max(\hat{f}(x)) = c) - c|] \tag{1.16}$$

In practice, ECE is calculated using an approximation which partitions predictions in M equally-spaced bins and takes a weighted average of the difference between accuracy and confidence in each bin. Formally this approximation is as follows:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)| \tag{1.17}$$

Where n is the number of samples, B_m is the set of indices of samples whose predictive confidence falls into bin m, $acc(B_m)$ is the accuracy in bin m and $conf(B_m)$ is the confidence for samples in bin m. ECE is a commonly used metric but has an important limitation. Using ECE, it is difficult to evaluate each class's contribution to the overall calibration performance. For example, in a classification task where the number of instances for each class is highly unbalanced, low ECE on instances from the most populated classes will result in low overall ECE even if less populated classes are poorly calibrated. A solution to this issue is to calculate ECE for each class independently and then average the class-specific ECE (Kull et al. 2019). This metric is called classwise-ECE and is defined as follows:

$$ECE_{classwise} = \frac{1}{k} \sum_{j=1}^{k} \sum_{m=1}^{M} \frac{|B_{m,j}|}{n} |y_j(B_{m,j}) - \hat{p}_j(B_{m,j})| \tag{1.18}$$

Where k, m, n are the numbers of classes, bins and instances, respectively, $B_{m,j}$ refers to bin m for instances of class j, $\hat{p}_j(B_{m,j})$ is the average probability of class j, and $y_j(B_{m,j})$ is the true proportion of class j in bin $B_{m,j}$. The contribution of a class j to the classwise-ECE is called class-j-ECE and can be used to examine how well a classifier is calibrated for each class. Reliability diagrams are used to visualize a model's calibration. A reliability diagram is constructed by plotting the empirical accuracy as a function of confidence. Similar to the approximation of ECE, reliability diagrams split confidence values into equal-sized bins. For a perfectly calibrated classifier, the reliability diagram should be a perfect diagonal, and any deviation from the diagonal represents miscalibration. Since reliability diagrams do not factor in the proportion of samples that fall into each confidence bin, they do not replace summary statistics like ECE and classwise-ECE.

Deep neural networks tend to produce highly accurate, but poorly calibrated classifiers (Guo et al. 2017). In general, this trend results in overly confident classifiers that may not be immediately useful when cost-sensitive decisions are being made. The disconnect between calibration and accuracy results from overfitting, where the neural network overfits to the negative log-likelihood without overfitting to accuracy (Guo et al. 2017). This results from classifying most samples correctly, but misclassifying a very small number of samples with increasingly high confidence. More recently, several neural network architectures have been developed that provide improved model calibration.

Deep ensembles and adversarial data

Deep ensembles are a novel class of neural networks which provide significant improvements in accuracy, calibration and out-of-distribution detection (covered in a later section) (Lakshminarayanan, Pritzel, and Blundell 2017). Deep ensembles consist of an ensemble of M independent neural networks, each with the same hyperparameters, but with random initialization. While it is unnecessary to use the same hyperparameters for each neural network in the ensemble, keeping the hyperparameters constant reduces the computational cost of multiple hyperparameter optimization procedures. More formally, consider a training set D consisting of N data points $D = \{x_n, y_n\}_{n=1}^{N}$, where $x \in X$ represents the input features, and $y \in Y = \{1, 2, ..., K\}$ represent the labels for a classification problem. Given the input features, a neural network is used to parametrize the predictive distribution $p_\theta(y|x)$ over the labels, where θ are the neural network parameters. A deep ensemble, then, is a collection of M neural networks where $\{\theta\}_{m=1}^{M}$ are the parameters of the ensemble. In contrast to the Random Forest model, a deep ensemble uses all of the training set to train each neural network. This is because the presence of multiple optima and random parameter initialization in neural networks introduces a sufficient degree of stochasticity to learn a highly expressive predictive distribution (Fort, Hu, and Lakshminarayanan 2020). An additional contribution of the deep ensemble framework is the use of adversarial data during training. Adversarial examples are data points that are very similar to original training examples but are misclassified by the neural network (Goodfellow, Shlens, and Szegedy 2015). The use of adversarial data has been demonstrated to act as a regularization method, helping prevent overfitting. By doing so, including adversarial data during training has been shown to improve model calibration, generalization and robustness (Goodfellow, Shlens, and Szegedy 2015). Adversarial data is included during training using the fast gradient sign method. Given an input x, a target y and a loss function $J(\theta, x, y)$, the fast gradient sign generates an adversarial example x' as follows:

$$x' = x + \epsilon sign(\nabla_x J(\theta, x, y)) \tag{1.19}$$

Where ϵ is a small value which defines the maximum perturbation allowed, intuitively, this method perturbs the input associated with each training example along a direction which increases loss. These examples are used to augment training by created additional training data $D' = (x', y)$.

The ensemble is treated as a uniformly weighted mixture model such that predictions are made by combining predictions from each model:

$$p(y|x) = \frac{1}{M}\sum_{m=1}^{M} p_{\theta_m}(y|x, \theta_m) \tag{1.20}$$

For classification tasks, this corresponds to averaging the predictions from each individual network in the ensemble.

Post-hoc model calibration

Given an already trained classifier, several post-hoc calibration methods exist. These methods perform transformations on a model's output in a way that aims to reduce the negative log-likelihood of the model. These methods have varying parameters or hyperparameters that are tuned by minimizing a negative log-likelihood on a held-out validation set. Four commonly used post-hoc calibration methods are temperature scaling, matrix scaling, vector scaling and Dirichlet scaling (Guo et al. 2017; Kull et al. 2019).

Temperature Scaling

Temperature scaling is one of the simplest post-hoc calibration method (Guo et al. 2017). Recall that the output of a probabilistic neural network is a softmax function. Given the model's logit vector z_x for an input sample x, the confidence prediction is as follows:

$$c = max(\sigma_{sm}(z_x)) \tag{1.21}$$

Where σ_{sm} is the $softmax$ function.

In temperature scaling, instead of working directly with z_x, the logit vector is scaled by a single temperature parameter, $T > 0$ for each class. Consequently, the scaled prediction is given by:

$$c = max(\sigma_{sm}(z_x/T)) \tag{1.22}$$

When the temperature parameter is greater than zero, T will raise the class probability vector's entropy. As $T \to 0$, the confidence value will go up. Temperature scaling has two advantages compared to other approaches. First, only a single parameter needs to be selected, limiting the risk of overfitting when learning a post-hoc method on a small validation set. Second, T will not change the class prediction from the $softmax$ as the same positive number scales each element. This means that a highly accurate classifier can be scaled to have better calibration performance without affecting the model's overall accuracy.

Matrix and Vector Scaling

Similar to temperature scaling, matrix and vector scaling work by performing a transformation on the logit vector (Guo et al. 2017). Matrix scaling works by learning a linear transformation on the logits such that the scaled confidence value is:

$$c = max(\sigma_{sm}(Wz_x + b)) \tag{1.23}$$

W and b are optimized on the negative log-likelihood using a validation set. This approach can be viewed as learning a multiclass logistic regression model using the model logit vector as the input features. Since the size of W grows quadratically with the number of classes, this method has a risk of overfitting when the validation set has a small number of samples. One potential workaround is to use vector scaling which is identical to matrix scaling, but uses a diagonal matrix W. In both matrix and vector scaling, there is no guarantee that the predictions after scaling will be the same as those from the original classifier.

Dirichlet Scaling

Dirichlet scaling is similar to matrix scaling in that it learns a multiclass logistic regression model on some output from the original classifier. The key difference is that Dirichlet scaling uses the class probability vector as a feature (Kull et al. 2019). In its linear parametrisation, Dirichlet scaling is as follows:

$$c = max(\sigma_{sm}(W \ln{(q)} + b)) \tag{1.24}$$

Where q represents the class probability vector of the original neural network, the use of log-transformed class probability vectors as input features compared to logit vectors in the case of matrix scaling means that the input to Dirichlet scaling has reduced information content compared to matrix scaling. In the original work describing Dirichlet scaling, this loss of information typically did not affect the performance of this calibration method (Kull et al. 2019).

Post-hoc calibration methods can also be applied to a classifier when there is a significant shift between a new dataset and the dataset used to train the classifier (Kull et al. 2019). A new dataset may have a different prior distribution over classes, and as such, the class probability vector may change significantly compared to the original dataset. Post-hoc calibration can re-calibrate predictions for this new context, allowing for an already trained classifier to be more easily adapted in the case of dataset shift.

1.5.2 Out-of-distribution detection

A second notion of uncertainty estimation that is important when deploying deep learning models in cost-sensitive scenarios is the ability to automatically detect anomalous or significantly different data from the data used during training. Traditionally, deep neural networks have been shown to classify out-of-distribution data as in-distribution with high confidence. In a cost-sensitive scenario such as in medical diagnosis, this can lead to falsely identifying the type of cancer a patient has, which could alter treatment regiments and prognosis. Here, I provide an overview of two approaches for out-of-distribution detection: the predictive entropy of a deep ensemble, and deterministic uncertainty quantification networks.

Out-of-distribution detection with deep ensembles

The output of a classifier can be used to quantify uncertainty. Recall, that a neural network classifier, $\hat{f} : X \to \Delta_k$, outputs a class probability vector $\hat{f}(x) = (\hat{f}_1(x), \hat{f}_2(x), ..., \hat{f}_k(x))$ belonging to $\Delta_k = \{(q_1, q_2, ..., q_k) \in [0,1]^k | \sum_{i=1}^{k} q_i = 1\}$. The probability vector can be used to quantify the overall predictive uncertainty for an input sample x. This can be done by looking at the entropy of the prediction vector, and in particular, this can be accomplished by looking at the entropy of the predictive distribution of a deep ensemble (Lakshminarayanan, Pritzel, and Blundell 2017; Amersfoort et al. 2020). Recall that the average predictive distribution of a deep ensemble is as follows:

$$p(y|x) = \frac{1}{M} \sum_{m=1}^{M} p_{\theta m}(y|x, \theta_m) \tag{1.25}$$

The entropy of the deep ensemble's predictive distribution is therefore defined as:

$$H(p(y|X)) = -\sum_{m=1}^{M} p_{\theta m}(y|x, \theta_m) \log p_{\theta m}(y|x, \theta_m) \tag{1.26}$$

Predictive distributions with high entropy suggest a greater degree of uncertainty. A deep ensemble's predictive entropy has demonstrated state-of-the-art performance in detecting out-of-distribution samples (Amersfoort et al. 2020). Using the deep ensemble's predictive distribution entropy has the additional benefit of not requiring any additional computational overhead, as out-of-distribution samples can be detected using a simple function of the model's output.

Deterministic uncertainty quantification networks

Deterministic uncertainty estimation networks (DUQ) is a recently proposed solution to identifying out-of-distribution data (Amersfoort et al. 2020). These networks require only a single forward pass of the model to provide uncertainty estimates. Unlike the neural network classifiers discussed thus far, DUQ is not a probabilistic classifier. Instead, DUQ makes predictions by computing a kernel or similarity function between embedded features and class-specific centroids. Uncertainty, then, is measured by the similarity between the model output and the closest centroid. If the feature vector representation of a data point has low similarity to any of the centroids, it is out-of-distribution. DUQ provides competitive performance for both classification and for identifying out of distribution samples.

DUQ consists of a feature extractor which embeds input data into a lower-dimensional space. This can be any feed-forward neural network provided the softmax layer has been removed. Instead of a softmax layer, DUQ contains a single learnable weight matrix W_k per class, k. DUQ computes a centroid for each class in the embedded space, and using these centroids; an input sample is assigned to the class with the most similar centroid. Formally, DUQ uses the radial basis function (RBF) kernel to compute the similarity between the model output and the centroids:

$$RBF(f_\theta(x), e_k) = \exp\left[-\frac{\frac{1}{n}\left\|W_k f_\theta(x) - e_c\right\|_2^2}{2\sigma^2}\right] \tag{1.27}$$

Where $f_\theta : \mathbb{R}^p \rightarrow \mathbb{R}^d$ is the neural network model, p is the input dimension, d is the dimension of the embedded space, and θ are the parameters of the neural network. e_k is the centroid vector for class k with centroid size n. W_k is a weight matrix, and σ is the length scale of the radial basis function. DUQ is trained by maximizing the similarity to the correct centroid while minimizing the similarity to all other centroids. DUQ provides state-of-the-art performance on several challenging datasets and has also demonstrated comparable accuracy to commonly used probabilistic neural networks. Similar to deep ensembles, it has the benefit of low computational overhead and can be used for both classification tasks and quantifying uncertainty. A limitation of DUQ, however, is that it does not provide an analogue for confidence calibration.

1.6 Overview of thesis research

Diagnostic challenges such as CUPS and multiple primary tumours underscore the need for genomics-based tumour typing methods. To this end, several machine learning models have been used to identify cancer type based on molecular or genomic features derived from tumour samples. These studies have focused on using cancer-associated mutations, gene expression or epigenetic alterations as features to train machine learning models for tumour typing (Penson et al. 2019; Grewal et al. 2019; Yuan et al. 2016). While these methods have had some success, they typically come with some shortcomings. Focusing on cancer-associated mutations, for example, has relatively low accuracy for a number of cancer types, and is overly dependent on the ability to identify oncogenic mutations within a tumour correctly. Methods using gene-expression or chromatin features tend to perform better, but fail to consider the high degree of phenotypic plasticity in tumours. Readouts of cell-state, such as chromatin assays or RNA sequencing do not necessarily provide ancestral information about previous cell-states. Therefore, they may not provide information about the primary tumour that seeded a metastasis. An alternative approach is to focus on passenger mutations - those mutations that are thought to be inconsequential for tumourigenesis. This can be done by using the strong relationship between regional mutation rate and chromatin features, allowing for mutation rate to be used as a proxy for chromatin state (Polak et al. 2015). As most mutations are passed on over multiple cell generations, mutation rate encodes information about ancestral cell-states and can be used as a feature for identifying tumour type. Additional cancer type-specific information can be gained by looking at the mutational signatures within a tumour, which are often specific to only a small number of cancer types. In my thesis, I focused on developing deep learning classifiers for identifying cancer type from patterns of somatic passenger mutations. After establishing the utility of passenger mutations for identifying cancer type, I focused on algorithmic improvements that allow for calibrated uncertainty estimates, a critical need for any machine learning model deployed in a clinical setting.

In Chapter 2, I describe and evaluate a new deep learning model for identifying cancer type based on patterns of somatic mutations derived from WGS of cancer genomes. To accomplish this, I implement and make use of a Bayesian optimization procedure to efficiently search the space of hyperparameters for deep neural network classifiers. Part of my evaluation included testing deep learning classifiers trained on a wide array of features derived from somatic mutations. I evaluated models trained using regional mutation density, mutation types (corresponding to mutational signatures), driver genes (cancer-associated mutations) and information about the biological pathways in which driver genes operate, and various combinations of the features. My evaluation demonstrated that the combination of passenger-mutation-derived features, namely, regional mutation density and mutation types, could accurately discriminate between 24 cancer types. This model had superior performance compared to other tumour-typing models when accounting for the relatively large number of cancer types the model is trained on. Surprisingly, when information about driver genes and pathways was included in addition to the passenger-mutation-derived features, overall model performance failed to improve, suggesting that passenger mutations alone are sufficient for accurately identifying cancer type, and providing further evidence for the relationship between passenger mutations and cancer type. Moreover, a model trained solely on regional mutation density had comparable performance to the best-trained model, suggesting that regional mutation density is strongly associated with cell-type. By investigating the misclassifications the model made, I demonstrated that cell-of-origin or mutational exposures could influence misclassifications. For example, the classifier mistakes Stomach-AdenoCA and Eso-AdenoCA, two cancers that originate in gastric tissues. These tumours also have highly similar patterns of mutational exposures, which also contribute to misclassification.

To demonstrate the utility of the passenger-trained classifier, I applied the classifier to an independent data set of primary tumours. Despite large differences in analysis pipelines and sequencing depth, the classifier was able to identify cancer type with high accuracy, suggesting that the model generalizes across multiple cohorts of tumour WGS. Finally, to determine if the model could identify the primary tumour site of metastases, I tested the model on the largest collections of WGS from metastatic tumours. Although most metastatic samples were sequenced following exposure to chemotherapy, the classifier trained on passenger mutations from primary tumours could accurately identify primary tumour type for the dataset of metastases. Reassuringly, the misclassification patterns match those seen on other datasets and generally were associated with common cell-of-origin or common mutational exposures. In addition to metastases of known primary, the model was applied to 62 CUPS. While the primary tumour site is not available for evaluating the performance on these data, some clinical information lends support to the model's applicability. Although sex chromosomes were not used for training the model, in all but one case when the classifier assigned a CUPS to be a gynaecological malignancy specific to female patients, the patients providing the samples were female.

In Chapter 3, I make several algorithmic advancements to address challenges in translating the classifier described in Chapter 2 into a clinical setting. The classifier described in Chapter 2 provided impressive performance for identifying cancer type, but several challenges exist for translating the model into a clinical setting. Namely, the model needs to be extended to a greater number of cancer types, and it needs to provide calibrated uncertainty estimates. I develop and evaluate a number of deep learning model architectures for extending the model to a greater number of cancer types and ultimately demonstrate that using a deep ensemble neural network architecture allows for the classifier to be extended to 29 cancer types with comparable performance to the model presented in Chapter 2. To improve the model's confidence calibration, the deep ensemble was trained with adversarial data. To provide the best-calibrated model, I implement and assess several post-hoc calibration methods, including Dirichlet scaling, matrix scaling, vector scaling and temperature scaling. Overall, the evaluation suggests that both the deep ensemble and the temperature scaled model provide relatively low calibration error, suggesting that they can provide estimates of reliability when predicting cancer type in a clinical setting. This represents one of the first uses of confidence calibration and post-hoc calibration to deep learning models in genomics and molecular biology. Another notion of uncertainty is the ability to determine if an input sample is highly dissimilar to the data distribution used to train a classifier. This is called out-of-distribution detection, and in a clinical setting, will allow for cancer samples that cannot be reliably identified to be automatically detected. To perform out-of-distribution detection, I used the predictive entropy of the deep ensemble. I ruled samples as out-of-distribution if they had high entropy relative to a validation set of in-distribution samples. Using this method, I developed a threshold hold that can accurately discriminate in-distribution from out-of-distribution samples on several datasets. Reassuringly, the CUPS samples' predictive entropy tended to be lower than the threshold value, suggesting that CUPS samples may be accurately classified by the deep ensemble method. Furthermore, using this cut-off to rule out test samples that cannot be reliably classified, the classifier's overall accuracy was improved. Together, this work represents significant algorithmic advancements that address challenges for translating the classifier into a clinical setting.

Taken together, the research presented in my thesis demonstrates the feasibility of using somatic passenger mutations derived from WGS and deep learning to identify cancer type. It makes important contributions in quantifying and assessing predictive uncertainty in deep learning models. These findings suggest that the classifier I developed has immediate clinical applicability in identifying the primary tumour site for CUPS. Furthermore, my results provide evidence that somatic passenger mutations are sufficient for accurately identifying cancer type across a large set of cancer types. The results of my thesis will provide potential diagnostic tools for clinicians, a methodology for assessing uncertainty for deep learning models in genomics, and evidence for further investigating the use of deep learning for mapping relationships between regional mutation density and chromatin features.

Chapter 2

Tumour typing using patterns of somatic mutations

This chapter is adapted from the following manuscript, which is published under the Creative Commons Attribution 4.0 International Licence (`https://creativecommons-org/licenses/by/4.0/`):

Jiao W*, **Atwal G***, Polak P*, Karlic R, Cuppen E, Danyi A, de Ridder J, van Herpen C, Lolkema MP, Steeghs N, Getz G, Morris QD, Stein LD. A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nat. Commun.*11, 728 (2020)

I developed methods and carried out the experiments using deep learning methods, and contributed to writing the manuscript. Wei Jiao carried out experiments using random forest models, and contributed to writing the manuscript. Paz Polak contributed to revising the manuscript. Lincoln Stein and Quaid Morris supervised the project and revised the manuscript.

2.1 Abstract

In cancer, the primary tumour's organ of origin and histopathology are the strongest determinants of its clinical behaviour, but in 3% of cases a patient presents with metastatic tumour and no obvious primary. Here we train a deep learning classifier to predict cancer type based on patterns of somatic passenger mutations detected in whole genome sequencing (WGS) of 2606 tumours representing 24 common cancer types produced by the PCAWG Consortium. Our classifier achieves an accuracy of 91% on held-out tumour samples and 88% and 83% respectively on independent primary and metastatic samples. Surprisingly, adding explicit information on alterations to cancer-associated genes and cancer-associated pathways reduced accuracy. Our results have clinical applicability, underscore how patterns of somatic passenger mutations encode the state of the cell of origin and can inform future strategies to detect the source of circulating tumour DNA.

2.2 Introduction

Human cancers can be categorized at multiple levels of resolution. Most commonly, human cancers are distinguished by their anatomic organ of origin and their histopathology. For example, lung squamous cell carcinoma originates in the lung and has histology similar to the normal squamous epithelium that lines bronchi and bronchioles. Together these two criteria define the tumour's cell of origin. Despite advances in precision medicine, a tumour's cell of origin is the single major predictor of the disease's natural history, including the age at which the tumour manifests, its factors, growth rate, the pattern of invasion and metastasis, response to therapy, and overall prognosis. Studies have shown that cancer-specific therapy based on the tumour's cell of origin is more effective than broad-spectrum chemotherapy (Greco 2013). Typically, a tumour's cell of origin can be determined through a variety of pathological assessments. However, it is not always straightforward to determine the origin of a metastatic tumour. In the most extreme case, a clinician may be presented with the challenge of determining the source of a poorly differentiated metastatic cancer when multiple imaging studies have failed to identify the primary ("cancer of unknown primary," CUPS) (Pavlidis, Khaled, and Gaafar 2015). In current clinical practice, pathologists use histological criteria assisted by immunohistochemical stains to determine such tumours' histological type and site of origin (D'cruze et al. 2013), but some tumours are so poorly differentiated that they no longer express the cell-type-specific proteins needed for unambiguous immunohistochemical classification.

Based on recent large-scale exome and genome sequencing studies we know that major tumour types present different patterns of somatic mutation (Kandoth 2013; Lawrence 2013; Ciriello 2013; Campbell et al. 2020). For example, ovarian cancers are distinguished by a high rate of genomic rearrangements (Patch 2015), chronic myelogenous leukaemia (CML) carry a nearly pathognomonic structural variation involving a t(9;22) translocation leading to a BCR-ABL fusion transcript (Kurzrock et al. 2003), melanomas have high rates of C >T and G >A transition mutations due to UV damage (Hayward et al. 2017), and pancreatic ductal adenocarcinomas have near-universal activating mutations in the KRAS gene (Biankin 2012). Recent work has also pointed to substantial variability in regional somatic mutation density(Schuster-Böckler and Lehner 2012). These studies provide evidence for a strong relationship between regional mutation density and a number of genomic features, including chromatin accessibility, levels of gene transcription, and histone marks (Supek and Lehner 2015; Polak et al. 2015; Polak et al. 2014). This correlation has suggested that the normal cell of origin for a tumour can be inferred from differences in regional mutation density (Kübler et al. 2019).

The PCAWG Consortium aggregated whole genome sequencing data from 2,658 cancers across 38 tumour types generated by the ICGC and TCGA projects. These sequencing data were re-analysed with standardised, high-accuracy pipelines to align to the human genome (reference build hs37d5) and identify germline variants and somatically acquired mutations, as described in PCAWG Network (Campbell et al. 2020).

This paper asks whether we can use machine learning techniques to accurately determine tumour organ of origin and histology using the patterns of somatic mutation identified by whole genome DNA sequencing. One motivation of this effort was to demonstrate the feasibility of a next-generation sequencing (NGS) based diagnostic tool for tumour type identification. Due to its stability, DNA is particularly easy to recover from fresh and historical tumour samples; furthermore, because mutations accumulate in DNA, they form a historic record of tumour evolution unaffected by the local, metastatic environment. Here we use deep learning techniques to explore whether a simple DNA-based sequencing and analysis protocol for tumour type determination would be a useful adjunct to existing histopathological techniques. Unexpectedly, we find that the regional mutation density of passenger mutations and mutation type are sufficient to discriminate among tumour types with a high degree of accuracy, while information about driver genes and pathways fail to improve classifier performance.

2.3 Results

2.3.1 Training Set

Using the Pan-cancer Analysis of Whole Genomes (PCAWG) data set (Campbell et al. 2020), we built a series of tumour-type classifiers using individual sequence-based features and combinations of features. The best performing classifier was validated against an independent set of tumour genomes to determine overall predictive accuracy and then tested against a series of metastatic tumours from known primaries to determine the accuracy of predicting the primary from a metastasis.

The full PCAWG data set consists of tumours from 2778 donors comprising 34 main histopathological tumour types, uniformly analysed using the same computational pipeline for quality control filtering, alignment, and somatic mutation calling. However, the PCAWG tumour types are unevenly represented, and several have inadequate numbers of specimens to adequately train and test a classifier. We chose a minimum cutoff of 35 donors per tumour type. In a small number of cases, the same donor contributed both primary and metastatic tumour specimens to the PCAWG data set. In these cases, we used only the primary tumour for training and evaluation, except for the case of the small cohort of myeloproliferative neoplasms (Myeloid-MPN; N=55 samples), for which multiple primary samples were available. In this case, we used up to two samples per donor and partitioned the training and testing sets to avoid having the same donor appear more than once in any training/testing set trial. The resulting training set consisted of 2436 tumours spanning 24 major types (Table 2.1 and Appendix A1.1).

Table 2.1: Distribution of tumour types in the PCAWG training and test data sets.

	Tumor Type	Samples
Abbreviation		
Liver-HCC	Liver hepatocellular carcinoma	306
Panc-AdenoCA	Pancreatic adenocarcinoma	235
Breast-AdenoCA	Breast adenocarcinoma	198
Prost-AdenoCA	Prostate adenocarcinoma	189
CNS-Medullo	Medulloblastoma	146
Kidney-RCC	Renal cell carcinoma (proximal tubules)	143
Ovary-AdenoCA	Ovarian adenocarcinoma	112
Skin-Melanoma	Skin melanoma	106
Lymph-BNHL	Mature B-cell lymphoma	105
Eso-AdenoCA	Esophageal adenocarcinoma	98
Lymph-CLL	Chronic lymphocytic leukemia	95
CNS-PiloAstro	Pilocytic astrocytoma	89
Panc-Endocrine	Pancreatic neuroendocrine tumor	85
Stomach-AdenoCA	Gastric adenocarcinoma	70
Head-SCC	Head/neck squamous cell carcinoma	57
ColoRect-AdenoCA	Colorectal adenocarcinoma	52
Lung-SCC	Lung squamous cell carcinoma	48
Thy-AdenoCA	Thyroid adenocarcinoma	48

Continued on next page

	Tumor Type	Samples
Abbreviation		
Myeloid-MPN	Myeloproliferative neoplasm	46
Kidney-ChRCC	Renal cell carcinoma (distal tubules)	45
Bone-Osteosarc	Sarcoma, bone	44
CNS-GBM	Diffuse glioma	41
Uterus-AdenoCA	Uterine adenocarcinoma	40
Lung-AdenoCA	Lung adenocarcinoma	38
		2436

2.3.2 Classification using Single Mutation Feature Types

To determine the predictive value of different mutation features, we trained and evaluated a series of tumour type classifiers based on single categories of feature derived from the tumour mutation profile. For each feature category, we developed a random forest (RF) classifier (See ”Methods” section). Each classifier's input was the mutational feature profile for an individual tumour specimen, and its output was the probability estimate that the specimen belongs to the type under consideration. Each classifier was trained using a randomly selected set of 75% of samples drawn from the corresponding tumour type. To determine the most likely type for a particular tumour sample, we applied its mutational profile to each of the 24 type-specific classifiers and selected the type whose classifier emitted the highest probability. To evaluate the performance of the system, we applied stratified four-fold cross-validation by training on three-quarters of the data set and testing against each of the other quarter specimens. We report overall accuracy as well as recall, precision and the F1 score using the average of all four test data sets (see ”Methods” section for cross-validation methodology and definitions of terms).

We selected a total of seven mutational feature types spanning three major categories (Table 2.2):

Table 2.2: WGS feature types used in classifiers.

	Feature Name	Feature Count	Description
Feature Category			
Mutation Distribution	SNV distribution	2897	Number of SNVs per 1 Mbp bin
Mutation Distribution	CNV distribution	2826	Number of CNAs per 1 Mbp bin
Mutation Distribution	SV distribution	2929	Number of SVs per 1 Mbp bin
Mutation Distribution	INDEL distribution	2757	Number of INDELS per 1 Mbp bin
Mutation Type	SNV type	150	Type of single nucleotide substitution
Mutation Pathway	Gene	554	Presence of mutation in cancer gene
Mutation Pathway	Pathway	1865	Presence of mutation in cancer pathway

Mutation Distribution: The somatic mutation rate in cancers varies considerably from one region of the genome to the next (Lawrence 2013). In whole genome sequencing, a major covariate of this regional variation in whole genome sequences is the epigenetic state of the tumour's cell of origin, with 74-86% of the variance in the mutation density being explained by histone marks and other chromatin features related to open versus closed chromatin (Ciriello 2013). This suggests that tumours sharing similar cells of origin will have a similar topological distribution of mutations across the genome. To capture this, we divided the genome into ∼3000 1 Mbp bins across the autosomes (excluding sex chromosomes) and created features corresponding to the number of somatic mutations per bin. For RF-based models, this feature was normalized to the total number of somatic mutations. For all other models, the mutation counts were used directly. Mutation rate profiles were created independently for somatic substitutions (SNV), indels, somatic copy number alterations (CNA), and other structural variations (SV). Note that the vast majority of variants, e.g., at least 99% of the SNVs in nearly all samples, used for this analysis are non-functional passenger mutations. See Campbell (Campbell et al. 2020) and Li (Li et al. 2020) for descriptions of point and structural variations in the PCAWG dataset.

Mutation Type: The type of the mutation and its nucleotide neighbours, for example, GC >TC, is an indicator of the exposure history of the cell of origin to extrinsic and endogenous factors that promote mutational processes (Alexandrov et al. 2013). This in turn can provide information on the aetiology of the tumour. For example, skin cancers have mutation types strongly correlated with UV light-induced DNA damage. Reasoning that similar tumour types will have similar mutational exposure profiles, we generated a series of features that represented the normalized frequencies of each potential nucleotide change in the context of its $5'$ and $3'$ neighbours. Like the mutation distribution, the variants that contribute to this feature category are mostly passengers. Readers are referred to Alexandrov (Alexandrov et al. 2020) for more information on signature analysis in the PCAWG data set.

Driver Gene/Pathway: Some tumour types are distinguished by high frequencies of alterations in particular driver genes and pathways. For example, melanomas have a high frequency of BRAF gene mutations (Pollock and Meltzer 2002), while pancreatic cancers are distinguished by KRAS mutations (Biankin 2012). We captured this in two ways: (1) whether a gene is affected by a driver event as determined by the PCAWG Cancer Drivers Working Group (Rheinbay et al. 2020), and (2) whether there was an impactful coding mutation in any gene belonging to a known or suspected driver pathway (also see Reyna (Reyna et al. 2020) for cancer pathway analysis performed by the PCAWG Pathway and Networks Working Group). We counted driver events affecting protein-coding genes, long noncoding RNAs and micro-RNAs, but did not attempt to account for alterations in cis-regulatory regions. In all, we created ∼2000 driver pathway-related features describing potential gene and pathway alterations for each tumour.

The accuracy of individual RF classifiers ranged widely across tumour and feature categories, with a median F1 (harmonic mean of recall and precision) of 0.42 and a range from 0.00 to 0.94 (Figure 2.1a,b, Table 2.3). Nine tumour types had at least one well-performing classifier that achieved an F1 of 0.80: CNS-GBM, CNS-PiloAstro, Liver-HCC, Lymph-BNHL, Kidney-RCC, Myeloid-MPN, Panc-AdenoCA, Prost-AdenoCA, Skin-melanoma. Five classifiers performed poorly, with no classifier achieving an accuracy greater than 0.6: Bone-Osteosarc, Head-SCC, Stomach-AdenoCA, Thy-AdenoCA and Uterus-AdenoCA. The remaining eight tumour types had classifiers achieving F1s between 0.60 and 0.80.

Table 2.3: Predictive accuracy of random forest trained on mutational features.

	SNV type	SNV distribution	CNV distribution	INDEL distribution	SV distribution	Gene	Pathway
Tumour Type							
Kidney-RCC	0.95	0.76	0.65	0.70	0.41	0.79	0.28
Liver-HCC	0.94	0.94	0.48	0.75	0.45	0.59	0.39
Skin-Melanoma	0.87	0.94	0.24	0.25	0.25	0.65	0.63
CNS-GBM	0.74	0.90	0.66	0.38	0.51	0.55	0.38
Myeloid-MPN	0.88	0.33	0.26	0.20	0.31	0.00	0.17
Lymph-BNHL	0.77	0.83	0.51	0.75	0.76	0.86	0.54

Continued on next page

Tumour Type	SNV type	SNV distribution	CNV distribution	INDEL distribution	SV distribution	Gene	Pathway
Prost-AdenoCA	0.77	0.85	0.67	0.35	0.68	0.25	0.27
Panc-AdenoCA	0.76	0.76	0.64	0.54	0.50	0.84	0.79
CNS-PiloAstro	0.69	0.71	0.58	0.66	0.81	0.03	0.53
ColoRect-AdenoCA	0.70	0.79	0.28	0.31	0.32	0.79	0.40
Lymph-CLL	0.79	0.65	0.77	0.45	0.75	0.33	0.16
Lung-SCC	0.70	0.78	0.44	0.55	0.18	0.55	0.27
CNS-Medullo	0.77	0.58	0.55	0.29	0.32	0.32	0.28
Breast-AdenoCA	0.54	0.74	0.44	0.52	0.39	0.37	0.26
Eso-AdenoCA	0.69	0.69	0.38	0.41	0.60	0.15	0.31
Lung-AdenoCA	0.45	0.69	0.11	0.12	0.00	0.00	0.06
Panc-Endocrine	0.62	0.54	0.67	0.29	0.38	0.54	0.12
Ovary-AdenoCA	0.64	0.66	0.65	0.37	0.39	0.20	0.22
Kidney-ChRCC	0.51	0.62	0.61	0.20	0.06	0.00	0.05
Thy-AdenoCA	0.53	0.10	0.40	0.54	0.13	0.03	0.09
Head-SCC	0.42	0.48	0.28	0.09	0.19	0.00	0.04
Uterus-AdenoCA	0.23	0.23	0.05	0.07	0.26	0.38	0.17
Bone-Osteosarc	0.37	0.27	0.20	0.04	0.26	0.00	0.03
Stomach-AdenoCA	0.33	0.32	0.16	0.06	0.24	0.00	0.01

The highest accuracies were observed for features related to mutation type and distribution (Figure 1b). Contrary to our expectations, altered driver genes and pathways were poor discriminatory features. Whereas both SNV type and distribution achieved median F1 scores of 0.7, RF models built on driver gene or pathway features achieved median F1s of 0.33 and 0.27, respectively. Only Panc-AdenoCA, Kidney-RCC, Lymph-BNHL and ColoRect-AdenoCA exceeded F1s greater than 0.75 on RF models built from gene or pathway-related features, but we note that even in these cases, the mutation type and/or distribution features performed equally well.

2.3.3 Classification using Combinations of Mutation Feature Types

We next asked whether we could improve classifier accuracy by combining features from two or more categories. We tested both Random Forest (RF) and multi-class Deep Learning/Neural Network (DNN)-based models (See "Methods" section), and found that overall the DNN-based models were more accurate than RF models across a range of feature category combinations (median F1=0.86 for RF, F1=0.90 for DNN, p<1.2e-7 Wilcoxon Rank Sum Test; Figure 2.1C). For the DNN-based models, overall accuracy was the highest when just the topological distribution and mutation type of SNVs were taken into account. Adding gene and/or pathway features slightly reduced classification accuracy; using only gene and pathway features greatly reduced classifier performance. We did not investigate the effect of training the DNN on CNV or SV features as these mutation types were not uniformly available in the validation data sets (see below).

Figure 2.2 shows a heatmap of the DNN classifier accuracy when tested against held out tumours (mean of 10 independently-built models). Overall, the accuracy for the complete set of 24 tumour types was 91% (classification accuracy), but there was considerable variation for individual tumours types (Table 2.4). Recall (also known as sensitivity) ranged from 0.61 (Stomach-AdenoCA) to 0.99 (Kidney-RCC). Precision (similar to specificity but is sensitive to the number of positives in the data set) was comparable, with rates ranging from 0.74 (Stomach-AdenoCA) to 1.00 (CNS-GBM, Skin-Melanoma, and Liver-HCC). Twenty-one of 24 tumour types achieved F1s greater than 0.80, including 8 of the 9 types that met this threshold for RF models built on single feature categories. The three worst-performing tumour types were CNS-PiloAstro (mean F1 0.79 across 10 independently-trained DNN models), Lung-AdenoCA (F1 0.77) and Stomach-AdenoCA (F1 0.67).

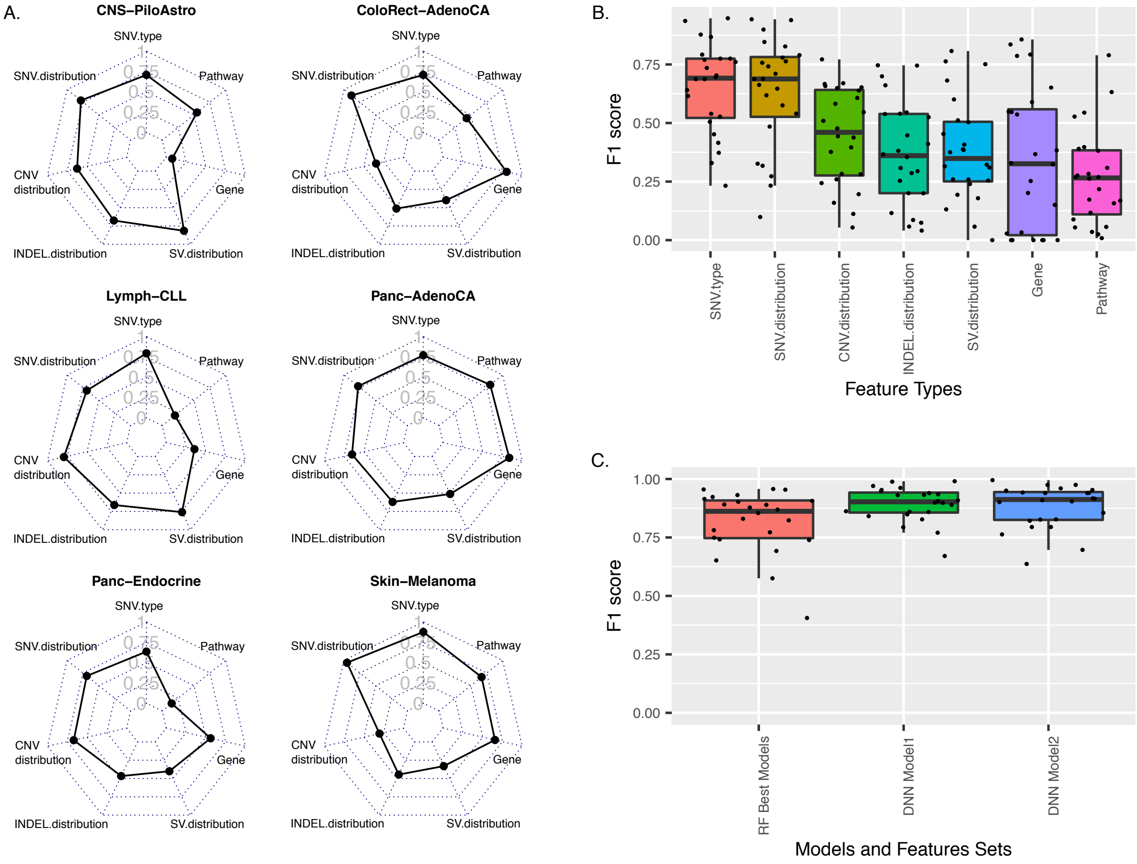


Figure 2.1: Comparison of tumour type classifiers using single and multiple features. (A) Radar plots describing the cross-validation-derived accuracy (F1) score of Random Forest classifiers trained on each of 7 individual feature categories, across six representative tumour types. (B) Summary of Random Forest classifier accuracy (F1) trained on individual feature categories across all 24 tumour types. (C) Accuracy of classifiers trained on multiple feature categories. RF Best Models corresponds to the cross-validation F1 scores of Random Forest classifiers trained on the three best single-feature categories for all 24 tumour types. DNN Model shows the distribution of F1 scores for held-out samples for a multi-class neural network trained using passenger mutation distribution and type. DNN Model+Drivers shows F1 scores for the neural net when driver genes and pathways are added to the training features. The centre line in the boxplot represents the median of the F1 scores. The lower and upper bounds of the box represent the first and third quartile. The whiskers extend to 1.5 IQR plus the third quartile or minus the first quantile. Figure from: (Jiao et al. 2020)

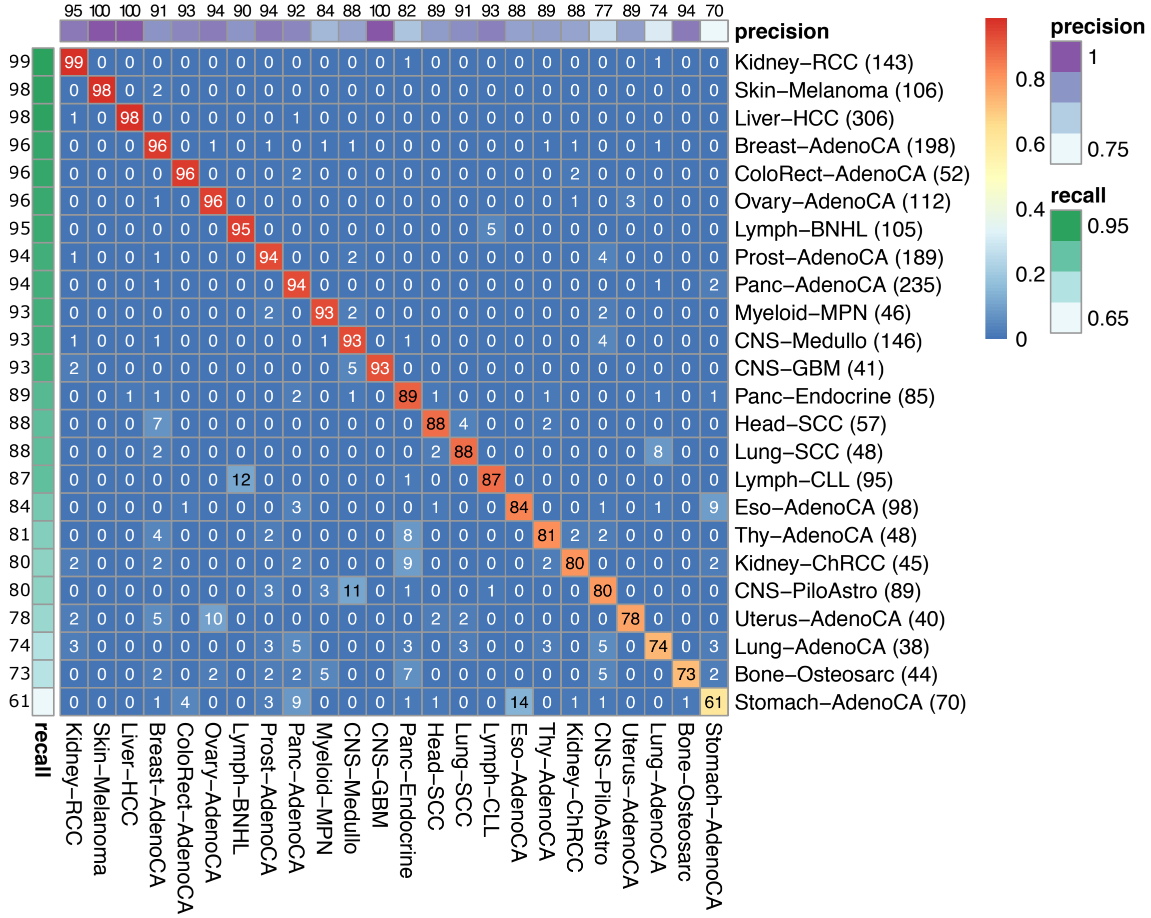


Figure 2.2: Heatmap displaying the accuracy of the merged classifier using a held-out portion of the PCAWG data set for evaluation. Each row corresponds to the true tumour type; columns correspond to the class predictions emitted by the DNN. Cells are labelled with the percentage of tumours of a particular type that were classified by the DNN as a particular type. The recall and precision of each classifier are shown in the colour bars at the top and left sides of the matrix. All values represent the mean of 10 runs using selected data set partitions. Due to rounding of values, some rows add up to slightly more or less than 100%. Figure from: (Jiao et al. 2020)

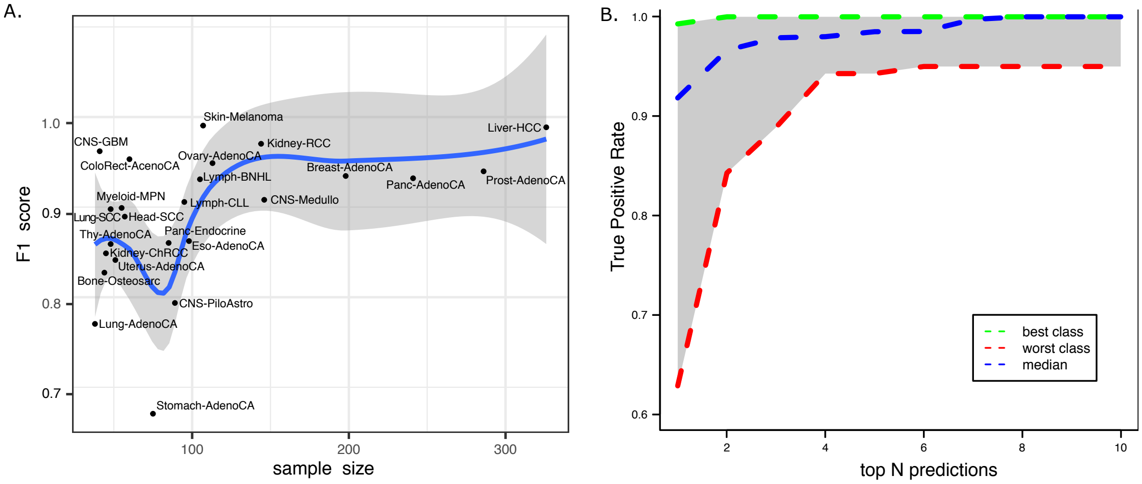


Figure 2.3: Performance of the DNN on held-out PCAWG data. (a) The relationship between training set size and prediction accuracy of the DNN is shown for each tumour type. The blue line represents a regression line fit using LOESS regression, while the grey area represents a 95% confidence interval for the regression function. (b) Accuracy of the classifier when it is asked to identify the correct tumour type among its top N-ranked predictions. The blue dashed line is the median true-positive rate among all 24 tumour classes. The green and red dashed lines correspond to the true-positive rate for the best- and worst-performing tumour classes. Figure from: (Jiao et al. 2020)

Table 2.4: Performance metrics of deep neural network.

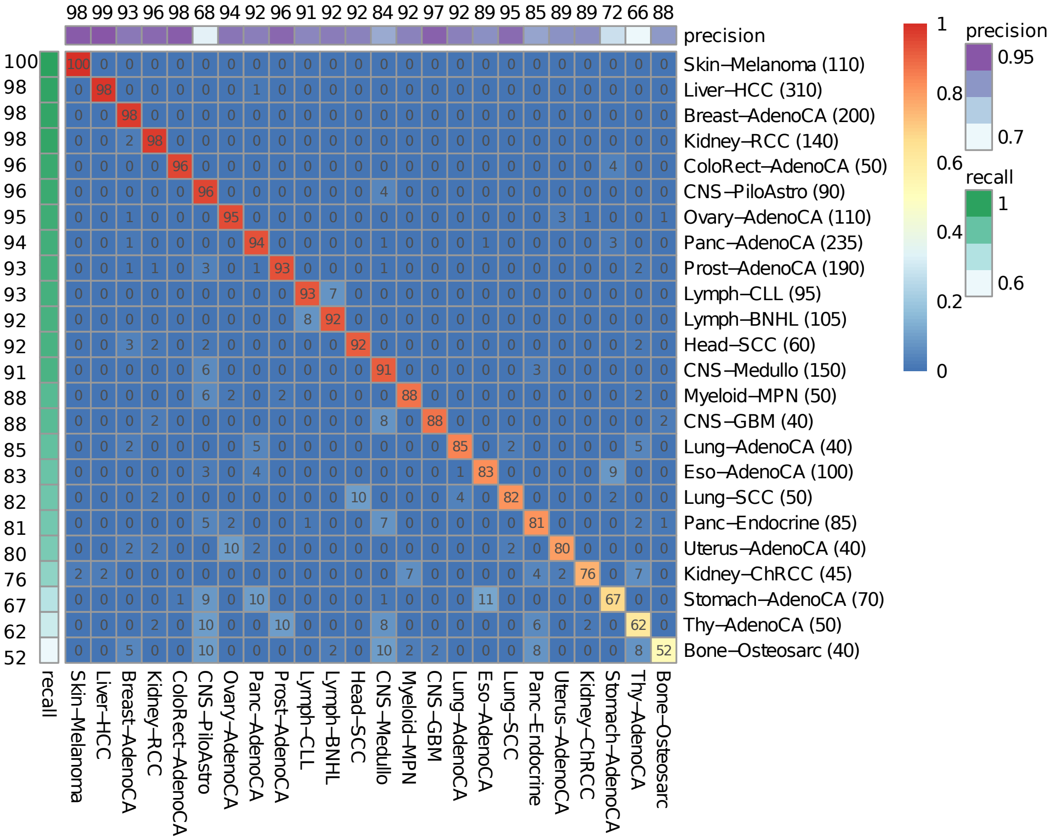
Tumour	Recall	Precision	F1
Kidney-RCC	0.99	0.95	0.97
Skin-Melanoma	0.98	1.00	0.99
Liver-HCC	0.98	1.00	0.99
Breast-AdenoCA	0.96	0.91	0.93
ColoRect-AdenoCA	0.96	0.95	0.95
Ovary-AdenoCA	0.96	0.94	0.95
Lymph-BNHL	0.95	0.91	0.93
Panc-AdenoCA	0.94	0.93	0.93
Prost-AdenoCA	0.94	0.94	0.94
Myeloid-MPN	0.93	0.87	0.90
CNS-Medullo	0.93	0.89	0.91
CNS-GBM	0.93	1.00	0.96
Panc-Endocrine	0.89	0.83	0.86
Head-SCC	0.88	0.90	0.89
Lung-SCC	0.88	0.92	0.90
Lymph-CLL	0.87	0.94	0.91
Eso-AdenoCA	0.84	0.89	0.86
Thy-AdenoCA	0.81	0.91	0.86
Kidney-ChRCC	0.80	0.90	0.85
CNS-PiloAstro	0.80	0.79	0.79
Uterus-AdenoCA	0.78	0.92	0.84
Lung-AdenoCA	0.74	0.81	0.77
Bone-Osteosarc	0.73	0.96	0.83
Stomach-AdenoCA	0.61	0.74	0.67

We investigated the effect of the training set size on classifier accuracy (Figure 2.3a). Tumour types with fewer than 100 samples in the data set were more likely to make incorrect predictions, and tumour types with large numbers of samples were among the top performers. However, several tumour types, including ColoRect-AdenoCA (N=52), Lung-SCC (N=48) and CNS-GBM (N=41) achieved excellent predictive accuracy despite having small training sets.

The DNN emits a softmax output that can be interpreted as the probability distribution of the tumour sample across the 24 cancer types. We ordinarily select the highest probability tumour type as the classifier's choice. If instead we asked how often the correct type is contained among the top N ranked probabilities, we find that the worst-performing tumour type (Stomach-AdenoCA) achieved a true positive rate of 0.88 for placing the correct tumour type among the top-ranked three choices and that the average true positive rate across all tumour types for this task was 0.98 (Figure 2.3b).

2.3.4 Patterns of Misclassification

Misclassifications produced by the DNN in many cases seem to reflect shared biological characteristics of the tumours that are representing in either the mutation distribution or mutation type features. For example, the most frequent classification errors for Stomach-AdenoCA samples were to two other upper gastrointestinal tumours, oesophagal adenocarcinoma (Eso-AdenoCA, 14% misclassification rate), and pancreatic ductal adenocarcinoma (Panc-AdenoCA, 9%). These three organs share a common developmental origin in the embryonic foregut and may share similar epigenetic profiles, which may be reflected in the mutation distribution for these tumours. We also speculate that the high rate of confusion between gastric and oesophagal cancers might be due to similar mutational exposures among the two sites: a subset of C>A, C>G substitutions are commonly seen in stomach and oesophagal (but not pancreatic) cancers and comprise Signature 17 in the COSMIC catalogue of mutational signatures (Forbes 2017). To test this, we assessed the effect of training the DNN with mutation distribution alone, excluding mutation-type features (Figure 2.4). Using just passenger mutation distribution, the overall F1 for stomach tumours increased by 4%, supporting the idea that part of the error is due to shared mutational signatures among stomach and oesophageal cancer. Another possible explanation for the frequent misclassification of gastric and oesophagal tumours is that some of the tumours labelled gastric arose at the gastroesophageal junction (GEJ), which some consider to be a distinct subset of oesophagal tumours (Rüschoff 2012).



	recall	Skin–Melanoma	Liver–HCC	Breast–AdenoCA	Kidney–RCC	ColoRect–AdenoCA	CNS–PiloAstro	Ovary–AdenoCA	Panc–AdenoCA	Prost–AdenoCA	Lymph–CLL	Lymph–BNHL	Head–SCC	CNS–Medullo	Myeloid–MPN	CNS–GBM	Lung–AdenoCA	Eso–AdenoCA	Lung–SCC	Panc–Endocrine	Uterus–AdenoCA	Kidney–ChRCC	Stomach–AdenoCA	Thy–AdenoCA	Bone–Osteosarc
precision		98	99	93	96	98	68	94	92	96	91	92	92	84	92	97	92	89	95	85	89	89	72	66	88
Skin–Melanoma (110)	100	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Liver–HCC (310)	98	0	98	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Breast–AdenoCA (200)	98	0	0	98	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Kidney–RCC (140)	98	0	0	2	98	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ColoRect–AdenoCA (50)	96	0	0	0	0	96	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0
CNS–PiloAstro (90)	96	0	0	0	0	0	96	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0
Ovary–AdenoCA (110)	95	0	0	1	0	0	0	95	0	0	0	0	0	0	0	0	0	0	0	0	3	1	0	0	1
Panc–AdenoCA (235)	94	0	0	1	0	0	0	0	94	0	0	0	0	1	0	0	0	1	0	0	0	0	3	0	0
Prost–AdenoCA (190)	93	0	0	1	1	0	3	0	1	93	0	0	0	1	0	0	0	0	0	0	0	0	0	2	0
Lymph–CLL (95)	93	0	0	0	0	0	0	0	0	0	93	7	0	0	0	0	0	0	0	0	0	0	0	0	0
Lymph–BNHL (105)	92	0	0	0	0	0	0	0	0	0	8	92	0	0	0	0	0	0	0	0	0	0	0	0	0
Head–SCC (60)	92	0	0	3	2	0	2	0	0	0	0	0	92	0	0	0	0	0	0	0	0	0	0	2	0
CNS–Medullo (150)	91	0	0	0	0	0	6	0	0	0	0	0	0	91	0	0	0	0	0	3	0	0	0	0	0
Myeloid–MPN (50)	88	0	0	0	0	0	6	2	0	2	0	0	0	0	88	0	0	0	0	0	0	0	0	2	0
CNS–GBM (40)	88	0	0	0	2	0	0	0	0	0	0	0	0	8	0	88	0	0	0	0	0	0	0	0	2
Lung–AdenoCA (40)	85	0	0	2	0	0	0	0	5	0	0	0	0	0	0	0	85	0	2	0	0	0	0	5	0
Eso–AdenoCA (100)	83	0	0	0	0	0	3	0	4	0	0	0	0	0	0	0	1	83	0	0	0	0	9	0	0
Lung–SCC (50)	82	0	0	0	2	0	0	0	0	0	0	0	10	0	0	0	4	0	82	0	0	0	2	0	0
Panc–Endocrine (85)	81	0	0	0	0	0	5	2	0	0	1	0	0	7	0	0	0	0	0	81	0	0	0	2	1
Uterus–AdenoCA (40)	80	0	0	2	2	0	0	10	2	0	0	0	0	0	0	0	0	0	2	0	80	0	0	0	0
Kidney–ChRCC (45)	76	2	2	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	4	2	76	0	7	0
Stomach–AdenoCA (70)	67	0	0	0	0	1	9	0	10	0	0	0	0	1	0	0	0	11	0	0	0	0	67	0	0
Thy–AdenoCA (50)	62	0	0	0	2	0	10	0	0	10	0	0	0	8	0	0	0	0	0	6	0	2	0	62	0
Bone–Osteosarc (40)	52	0	0	5	0	0	10	0	0	0	0	2	0	10	2	2	0	0	0	8	0	0	0	8	52

Figure 2.4: Heatmap displaying the accuracy of the merged classifier using a held-out portion of the PCAWG data set for evaluation. Heatmap displaying the accuracy of the merged classifier using a held-out portion of the PCAWG data set for evaluation. The classifier was trained using the mutation distribution. Each row corresponds to the true tumor type; Columns correspond to the predictions emitted by each of the classifiers. Cells are labeled with the proportion of tumors of a particular type that were called by each type-specific classifier. The recall and precision of each classifier is shown in the color bars at the top and left sides of the matrix. Figure from: (Jiao et al. 2020)

Other common misclassification errors include misclassification of 12% of chronic lymphocytic leukaemia (Lymph-CLL) samples as B-cell non-Hodgkin's lymphoma (Lymph-BNHL). Both tumours are derived from the B-cell lymphocyte lineage, and likely share a similar cell of origin. Another pattern was occasional misclassifications among the three types of brain tumour CNS-GBM, CNS-Medullo, and CNS-PiloAstro, all three of which are derived from various glial lineages. We speculate that these errors are again due to similarities among the cells of origin of these tissues.

Despite the difficulties described above, the DNN was able to accurately distinguish among several tumour types that arise from the same organ. Renal cell carcinoma (Kidney-RCC) and chromophobe renal carcinoma (Kidney-ChRCC), were readily distinguished from each other, as were the squamous and adenocarcinoma forms of non-small cell lung cancer (Lung-SCC, Lung-AdenoCA), and the exocrine and endocrine forms of pancreatic cancer (Panc-AdenoCA, Panc-Endocrine). The misclassification rate between Lung-SCC and Lung-AdenoCA was just 8%, and all other pairs had misclassification rates of 2% or lower. This is in keeping with a model in which major histological subtypes of tumours reflect different cells of origin.

2.3.5 Validation on an Independent Set of Primary Tumors

A distinguishing characteristic of the PCAWG data set is its use of a uniform computational pipeline for sequence alignment, quality filtering, and variant calling. In real-world settings, however, the data set used to train the classifier may be called using a different set of algorithms than the test data. To assess the ability of the DNN to generalize to this setting, we applied the classifier trained on PCAWG samples to an independent validation set of 1,436 cancer whole genomes assembled from a series of published non-PCAWG projects. The validation set spans 14 distinct tumour types assembled from 21 publications or databases (Table 2.5). We were unable to collect sufficient numbers of independent tumour genomes representing nine of the 24 types in the merged classifier, including colorectal cancer, thyroid adenocarcinoma and lung squamous cell carcinoma. SNV coordinates were lifted from GRCh38 to GRCh37 when necessary, but we did not otherwise process the mutation call sets.

Table 2.5: Distribution and source of tumour types contained within the validation data sets.

	Type	Year	Genome Version	#Samples
Source				
Primary Tumour WGS				
doi:10.1038/nature12477	Primary	2013	GRCh37	72
doi:10.1038/nature17676	Primary	2016	GRCh37	455
doi:10.1016/j.cell.2012.04.024	Primary	2012	GRCh37	1
doi:10.1038/nature11213	Primary	2012	GRCh38	11
doi:10.1038/ng.2938	Primary	2014	GRCh38	33
doi:10.1038/ng.2611	Primary	2013	GRCh38	16
doi:10.1038/ng.2699	Primary	2013	GRCh38	14
doi:10.1038/ng.3547	Primary	2016	GRCh37	1
doi:10.1038/nature09744	Primary	2011	GRCh38	8
doi:10.1038/nature08658	Primary	2009	GRCh38	1
ICGC (https://dcc.icgc.org/)	Primary	Jun2017	GRCh37	551
COSMIC	Primary	Aug2017	GRCh38	73
doi: 10.1158/1078-0432.CCR-17-2994	Primary	2017	GRCh37	200
			Total	1436
Metastatic Tumour WGS				
doi: 10.1101/41513	Metastatic	2018	GRCh37	2028
doi: 10.1101/41513	Metastatic	2018	GRCh37	62
doi: 10.1158/1078-0432.CCR-17-2994	Metastatic	2017	GRCh37	92
			Total	2182

The DNN classifier recall for the individual tumour types included in the validation data set ranged from 0.41 to 0.98, and the precision ranged from 0.43 to 1.0 (Figure 2.4a). The overall accuracy of the classifier was 88% across the 12 cancer types in the validation set. In general, the tumour types that performed the best within the PCAWG data set were also the most accurate within the validation, with Breast-AdenoCA, Ovary-AdenoCA, Panc-AdenoCA, Lymph-CLL, CNS-Medullo, and Kidney-RCC tumour types all achieving greater than 85% accuracy. The Eso-AdenoCA, Liver-HCC, and Pediatric

Gliomas were poorly predicted with recalls below 70%, and the remaining types had intermediate accuracies.

The majority of classification errors observed in the primary tumour validation set mirrored the patterns of misclassifications previously observed within the PCAWG samples, with the exception that Liver-HCC cases were frequently misclassified as CNS-Medullo (13%). We believe this case to be due to a lower than expected mutation burden in the liver tumours from the validation set (median 3202 SNVs per sample in validation set compared to 22,230 SNVs per sample in the PCAWG training set; P <1.5e-15 by Wilcoxon Rank Sum Test; Figure 2.5). This mutation load is more similar to the rates observed in CNS-Medullo (median 2330 per sample) among the PCAWG samples, and might suggest poor coverage of Liver-HCC or another sequencing/analysis artifact in the validation set.

We were initially puzzled that a set of 49 validation data set samples that were identified as CNS glioma overwhelmingly matched to the pediatric piloastrocytoma model rather than to the CNS-GBM model. However, on further investigation, we discovered that these samples represent a mixture of low- and high-grade pediatric gliomas, including piloastrocytomas (Wu 2014; Zhang et al. 2013; Ceccarelli 2016). The SNV mutation burden of these pediatric gliomas is also similar to CNS-PiloAstro and significantly lower than adult CNS-GBM (Figure 2.5).

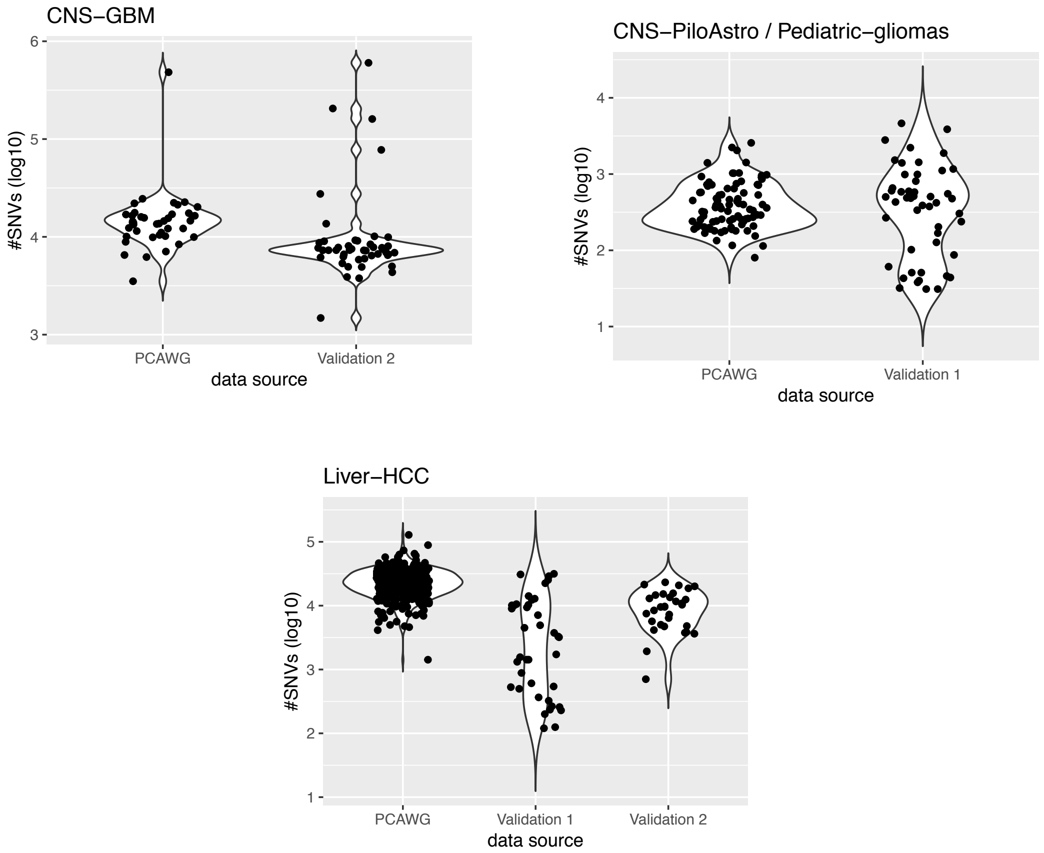
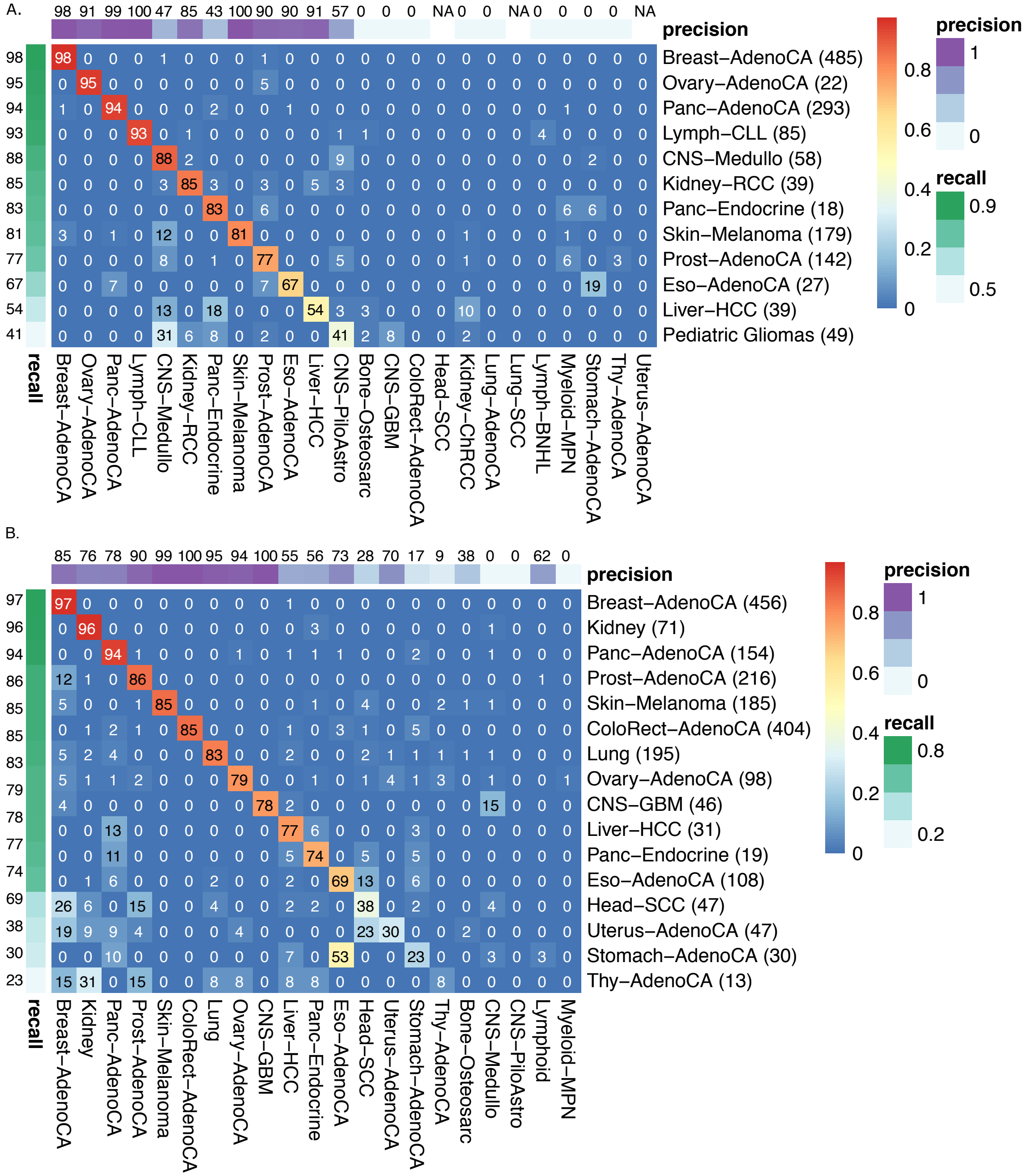


Figure 2.5: Comparison of SNV counts between PCAWG and the validation data set. Violin charts demonstrating the distribution of the number of SNVs in the PCAWG and validation data sets. Note that we have paired the validation set of pediatric gliomas with the PCAWG juvenile piloastrocytoma data set. Figure adapted from: (Jiao et al. 2020).



A.

recall		Breast–AdenoCA	Ovary–AdenoCA	Panc–AdenoCA	Lymph–CLL	CNS–Medullo	Kidney–RCC	Panc–Endocrine	Skin–Melanoma	Prost–AdenoCA	Eso–AdenoCA	Liver–HCC	CNS–PiloAstro	Bone–Osteosarc	CNS–GBM	ColoRect–AdenoCA	Head–SCC	Kidney–ChRCC	Lung–AdenoCA	Lung–SCC	Lymph–BNHL	Myeloid–MPN	Stomach–AdenoCA	Thy–AdenoCA	Uterus–AdenoCA
	precision	98	91	99	100	47	85	43	100	90	90	91	57	0	0	0	NA	0	0	NA	0	0	0	0	NA
98	Breast–AdenoCA (485)	98	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
95	Ovary–AdenoCA (22)	0	95	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
94	Panc–AdenoCA (293)	1	0	94	0	0	0	2	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0
93	Lymph–CLL (85)	0	0	0	93	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0	4	0	0	0	0
88	CNS–Medullo (58)	0	0	0	0	88	2	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	2	0	0
85	Kidney–RCC (39)	0	0	0	0	3	85	3	0	3	0	5	3	0	0	0	0	0	0	0	0	0	0	0	0
83	Panc–Endocrine (18)	0	0	0	0	0	0	83	0	6	0	0	0	0	0	0	0	0	0	0	0	6	6	0	0
81	Skin–Melanoma (179)	3	0	1	0	12	0	0	81	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0
77	Prost–AdenoCA (142)	0	0	0	0	8	0	1	0	77	0	0	5	0	0	0	0	1	0	0	0	6	0	3	0
67	Eso–AdenoCA (27)	0	0	7	0	0	0	0	0	7	67	0	0	0	0	0	0	0	0	0	0	0	19	0	0
54	Liver–HCC (39)	0	0	0	0	13	0	18	0	0	0	54	3	3	0	0	0	10	0	0	0	0	0	0	0
41	Pediatric Gliomas (49)	0	0	0	0	31	6	8	0	2	0	0	41	2	8	0	0	2	0	0	0	0	0	0	0

B.

recall		Breast–AdenoCA	Kidney	Panc–AdenoCA	Prost–AdenoCA	Skin–Melanoma	ColoRect–AdenoCA	Lung	Ovary–AdenoCA	CNS–GBM	Liver–HCC	Panc–Endocrine	Eso–AdenoCA	Head–SCC	Uterus–AdenoCA	Stomach–AdenoCA	Thy–AdenoCA	Bone–Osteosarc	CNS–Medullo	CNS–PiloAstro	Lymphoid	Myeloid–MPN
	precision	85	76	78	90	99	100	95	94	100	55	56	73	28	70	17	9	38	0	0	62	0
97	Breast–AdenoCA (456)	97	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
96	Kidney (71)	0	96	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	1	0	0	0
94	Panc–AdenoCA (154)	0	0	94	1	0	0	0	1	0	1	1	1	0	0	2	0	0	1	0	0	0
86	Prost–AdenoCA (216)	12	1	0	86	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
85	Skin–Melanoma (185)	5	0	0	1	85	0	0	0	0	0	1	0	4	0	0	2	1	1	0	0	0
85	ColoRect–AdenoCA (404)	0	1	2	1	0	85	0	0	0	1	0	3	1	0	5	0	0	0	0	0	0
83	Lung (195)	5	2	4	0	0	0	83	0	0	2	0	0	2	1	1	1	1	1	0	0	0
79	Ovary–AdenoCA (98)	5	1	1	2	0	0	0	79	0	0	1	0	1	4	1	3	0	1	0	0	1
78	CNS–GBM (46)	4	0	0	0	0	0	0	0	78	2	0	0	0	0	0	0	0	15	0	0	0
77	Liver–HCC (31)	0	0	13	0	0	0	0	0	0	77	6	0	0	0	3	0	0	0	0	0	0
74	Panc–Endocrine (19)	0	0	11	0	0	0	0	0	0	5	74	0	5	0	5	0	0	0	0	0	0
69	Eso–AdenoCA (108)	0	1	6	0	0	0	2	0	0	2	0	69	13	0	6	0	0	0	0	0	0
38	Head–SCC (47)	26	6	0	15	0	0	4	0	0	2	2	0	38	0	2	0	0	4	0	0	0
30	Uterus–AdenoCA (47)	19	9	9	4	0	0	0	4	0	0	0	0	23	30	0	0	2	0	0	0	0
23	Stomach–AdenoCA (30)	0	0	10	0	0	0	0	0	0	7	0	53	0	0	23	0	0	3	0	3	0
	Thy–AdenoCA (13)	15	31	0	15	0	0	8	8	0	8	8	0	0	0	0	8	0	0	0	0	0

Figure 2.6: Prediction accuracy for the DNN against two independent validation data sets. (a) Primary tumours. (b) Metastatic tumours. Each row corresponds to the true tumour type; columns correspond to the class predictions emitted by the DNN. Cells are labelled with the percentage of tumours of a particular type that were classified by the DNN as a particular type. The recall and precision of each classifier are shown in the colour bars at the top and left sides of the matrix. Due to rounding of values, some rows add up to slightly more or less than 100%. Figure adapted from: (Jiao et al. 2020).

2.3.6 Validation on an Independent Set of Metastatic Tumors

To evaluate the ability of the classifier to correctly identify the type of the primary tumour from a metastatic tumour sample, we developed an independent validation data set that combined a published series of 92 metastatic Panc-AdenoCA (Aung 2018) with an unpublished set of 2,028 metastatic tumours from known primaries across 16 tumour types recently sequenced by the Hartwig Medical Foundation (HMF) (Priestley et al. 2019), resulting in a combined set of 2,120 samples across 16 tumour types (Table 2.5). All metastatic samples were subjected to paired-end WGS sequencing of tumour and normal at a tumour coverage of at least 65x, but the computational pipelines used for alignment, quality filtering, and SNV calling were different from those used for PCAWG. In addition, samples from the HMF dataset were obtained using a needle biopsy which can limit spatial heterogeneity in the sequenced sample. The rules for matching classifier output to the validation set class labels were developed in advance of the experiment, and the DNN classifier was applied to the molecular data from the validation set in a blind fashion.

When the DNN classifier was applied to these metastatic samples it achieved an overall accuracy of 83% for identifying the type of the known primary (Figure 2.4b), which is similar to its performance on the validation primaries. Seven of the tumour types in the metastatic set achieved recall rates of 0.80 or higher, including Breast-AdenoCA (0.97), Kidney (0.96), Panc-AdenoCA (0.94), Prost-AdenoCA (0.86), Skin-Melanoma (0.85), ColoRect-AdenoCA (0.85), and Lung (0.83). On the other end of the spectrum, four tumour types failed to achieve a recall of at least 0.50: Head-SCC (0.38), Uterus-AdenoCA (0.30), Stomach-AdenoCA (0.23), and Thyroid-AdenoCA (0.08). Overall, the patterns of misclassification were similar to what was seen within PCAWG. For example, the gastric cancers were misclassified as oesophagal tumours 53% of the time.

In contrast to the other tumour types, metastatic thyroid adenocarcinoma was a clear outlier. In this case, the DNN was unable to correctly identify a great majority of the 13 metastatic samples, classifying them instead as other tumour types such as Kidney, Panc-Endocrine, Prost-AdenoCA or Breast-AdenoCA. We lack information on the histological subtype of the metastatic thyroid tumours in the HMF data set, but speculate that the metastatic thyroid tumours in this set are enriched in more aggressive histological subtypes than the PCAWG primaries, which are exclusively of low-grade papillary (N=31), papillary-follicular (N=18) and papillary-columnar (N=1) types.

The HMF data set also included 62 CUPs tumours. While we do not know the corresponding primary for these samples, we did attempt to classify them (Table 2.6). The CUPs cases were most frequently classified as Liver-HCC (N=10; 16%), Lung-AdenoCA (N=9; 15%) and Panc-AdenoCA (N=8; 13%). Reassuringly, despite the fact that information on the sex chromosomes were not used by the classifier, almost all the CUPS tumours classified as gynaecological tumours (Breast-AdenoCA, N=5; Uterus-AdenoCA, N=2) came from female patients except one patient with a low confidence prediction.

Table 2.6: Sample information and top 3 predictions for cancers of unknown primary.

HMF Sample ID	Patient Sex	Rank 1	Rank 2	Rank 3
sample237	FEMALE	Panc.Endocrine (0.48)	Breast.AdenoCA (0.23)	Kidney.RCC (0.12)
sample347	FEMALE	Panc.Endocrine (0.42)	Kidney.RCC (0.28)	Liver.HCC (0.21)
sample396	FEMALE	Lung.AdenoCA (1)	Skin.Melanoma (0)	CNS.GBM (0)
sample474	MALE	Stomach.AdenoCA (0.82)	Panc.AdenoCA (0.09)	Eso.AdenoCA (0.06)
sample487	FEMALE	Head.SCC (0.42)	Breast.AdenoCA (0.16)	Prost.AdenoCA (0.14)
sample492	FEMALE	Head.SCC (0.37)	Breast.AdenoCA (0.17)	Lung.SCC (0.15)
sample584	FEMALE	ColoRect.AdenoCA (0.58)	Stomach.AdenoCA (0.32)	Kidney.RCC (0.1)
sample610	MALE	Lymph.BNHL (1)	Lymph.CLL (0)	Stomach.AdenoCA (0)
sample694	MALE	Stomach.AdenoCA (0.9)	Panc.AdenoCA (0.1)	Eso.AdenoCA (0)
sample713	FEMALE	Kidney.RCC (0.44)	Liver.HCC (0.38)	Prost.AdenoCA (0.07)
sample744	FEMALE	Kidney.RCC (0.7)	Prost.AdenoCA (0.09)	Bone.Osteosarc (0.06)
sample777	NaN	Liver.HCC (0.84)	Panc.AdenoCA (0.08)	Panc.Endocrine (0.02)
sample811	NaN	ColoRect.AdenoCA (0.94)	Lymph.BNHL (0.06)	Liver.HCC (0)
sample896	FEMALE	Bone.Osteosarc (0.42)	Lung.AdenoCA (0.17)	Kidney.RCC (0.12)
sample907	MALE	Lung.AdenoCA (1)	Lung.SCC (0)	Skin.Melanoma (0)
sample934	FEMALE	Lung.AdenoCA (1)	Lung.SCC (0)	Stomach.AdenoCA (0)
sample945	MALE	Liver.HCC (0.64)	Breast.AdenoCA (0.18)	Panc.AdenoCA (0.04)
sample972	MALE	Liver.HCC (1)	Kidney.RCC (0)	Breast.AdenoCA (0)
sample1016	MALE	Liver.HCC (0.78)	Panc.AdenoCA (0.07)	Panc.Endocrine (0.06)
sample1034	FEMALE	Lung.AdenoCA (0.9)	Lung.SCC (0.1)	Stomach.AdenoCA (0)
sample1091	MALE	ColoRect.AdenoCA (0.96)	Stomach.AdenoCA (0.04)	Eso.AdenoCA (0)
sample1145	MALE	Kidney.RCC (0.33)	Liver.HCC (0.23)	Panc.AdenoCA (0.21)
sample1218	MALE	Breast.AdenoCA (0.42)	Head.SCC (0.18)	Uterus.AdenoCA (0.1)
sample1219	MALE	Lung.SCC (1)	Uterus.AdenoCA (0)	Lung.AdenoCA (0)
sample1222	MALE	Lung.AdenoCA (1)	Lung.SCC (0)	Ovary.AdenoCA (0)
sample1254	FEMALE	Uterus.AdenoCA (0.36)	Prost.AdenoCA (0.32)	CNS.Medullo (0.13)
sample1275	MALE	Eso.AdenoCA (0.39)	Panc.AdenoCA (0.29)	Stomach.AdenoCA (0.2)
sample1288	MALE	Panc.AdenoCA (0.35)	Liver.HCC (0.25)	Lung.AdenoCA (0.11)
sample1336	MALE	Head.SCC (0.48)	Lung.SCC (0.24)	Bone.Osteosarc (0.11)
sample1379	FEMALE	Breast.AdenoCA (0.94)	Prost.AdenoCA (0.05)	CNS.Medullo (0)
sample1483	MALE	Lung.SCC (1)	Lung.AdenoCA (0)	Uterus.AdenoCA (0)
sample1529	MALE	CNS.Medullo (0.26)	Breast.AdenoCA (0.22)	Thy.AdenoCA (0.21)
sample1576	MALE	Panc.AdenoCA (0.68)	Stomach.AdenoCA (0.12)	Prost.AdenoCA (0.11)
sample1639	FEMALE	Liver.HCC (1)	Kidney.RCC (0)	Panc.Endocrine (0)
sample1686	MALE	Panc.AdenoCA (0.72)	Eso.AdenoCA (0.18)	Stomach.AdenoCA (0.1)
sample1773	MALE	ColoRect.AdenoCA (0.71)	Panc.AdenoCA (0.29)	Stomach.AdenoCA (0)
sample1808	FEMALE	Breast.AdenoCA (0.98)	Lung.AdenoCA (0)	ColoRect.AdenoCA (0)
sample1817	FEMALE	Lung.AdenoCA (1)	Lung.SCC (0)	Ovary.AdenoCA (0)
sample1877	FEMALE	Liver.HCC (0.47)	Kidney.RCC (0.31)	Prost.AdenoCA (0.15)
sample1927	MALE	Head.SCC (0.54)	Lung.SCC (0.42)	Breast.AdenoCA (0.01)
sample1954	MALE	Lung.AdenoCA (1)	Lung.SCC (0)	Stomach.AdenoCA (0)
sample2077	FEMALE	Uterus.AdenoCA (0.97)	ColoRect.AdenoCA (0.01)	Ovary.AdenoCA (0.01)
sample2080	FEMALE	Lung.AdenoCA (1)	Lung.SCC (0)	Panc.Endocrine (0)
sample2082	FEMALE	Breast.AdenoCA (1)	Head.SCC (0)	Uterus.AdenoCA (0)
sample2102	FEMALE	Stomach.AdenoCA (0.66)	ColoRect.AdenoCA (0.13)	Uterus.AdenoCA (0.1)
sample2140	FEMALE	ColoRect.AdenoCA (0.62)	Stomach.AdenoCA (0.2)	Panc.AdenoCA (0.11)
sample2215	MALE	Panc.AdenoCA (0.87)	Stomach.AdenoCA (0.05)	Eso.AdenoCA (0.03)

Continued on next page

HMF Sample ID	Patient Sex	Rank 1	Rank 2	Rank 3
sample2242	FEMALE	Uterus.AdenoCA (0.58)	Ovary.AdenoCA (0.39)	Prost.AdenoCA (0.01)
sample2324	MALE	Liver.HCC (0.4)	Panc.AdenoCA (0.3)	Kidney.RCC (0.16)
sample2325	MALE	Skin.Melanoma (1)	Bone.Osteosarc (0)	Breast.AdenoCA (0)
sample2327	FEMALE	Panc.AdenoCA (0.21)	Kidney.RCC (0.17)	Kidney.ChRCC (0.12)
sample2337	MALE	Liver.HCC (0.43)	Panc.Endocrine (0.17)	Kidney.RCC (0.12)
sample2361	FEMALE	Panc.AdenoCA (1)	Eso.AdenoCA (0)	Stomach.AdenoCA (0)
sample2390	FEMALE	Head.SCC (0.48)	Lung.SCC (0.37)	Prost.AdenoCA (0.06)
sample2436	FEMALE	Breast.AdenoCA (0.6)	Head.SCC (0.2)	Bone.Osteosarc (0.1)
sample2466	MALE	Eso.AdenoCA (0.46)	Panc.AdenoCA (0.19)	Stomach.AdenoCA (0.17)
sample2478	MALE	Panc.AdenoCA (0.41)	Liver.HCC (0.33)	Lung.AdenoCA (0.1)
sample2493	MALE	Liver.HCC (0.53)	Kidney.RCC (0.34)	Panc.AdenoCA (0.07)
sample2529	MALE	Panc.Endocrine (0.21)	Kidney.ChRCC (0.2)	Lung.AdenoCA (0.2)
sample2641	FEMALE	Lung.AdenoCA (1)	Ovary.AdenoCA (0)	Lung.SCC (0)
sample2826	MALE	Liver.HCC (1)	Panc.Endocrine (0)	ColoRect.AdenoCA (0)
sample2859	MALE	ColoRect.AdenoCA (0.99)	Stomach.AdenoCA (0.01)	Eso.AdenoCA (0)

2.4 Discussion

Cancer of unknown primary site (CUPS) is a heterogeneous set of cancers diagnosed when a patient presents with metastatic disease, but despite extensive imaging, pathological and molecular studies the primary cannot be determined (Greco 2013). CUPS accounts for 3-5% of cancers, making it the seventh to eighth most frequent type of cancer and the fourth most common cause of cancer death (Pavlidis et al. 2003). Even at autopsy, the primary cannot be identified roughly 70% of the time (Ferracin et al. 2011), suggesting regression of the primary in many CUPS cases. CUPS is a clinical dilemma, because therapeutic options are largely driven by tissue of origin, and cancer-specific therapy is more effective than broad-spectrum chemotherapy (Greco 2013). A related diagnostic challenge arises, paradoxically, from the medical community's success in treating cancers and the rising incidence of second primary cancers, now estimated at roughly 16% of incident cancers (Travis 2006). Pathologists are often asked to distinguish a late metastatic recurrence of a previously treated primary from a new unrelated primary. However, histopathology alone may be inaccurate at identifying the site of origin of metastases. In one study (Sheahan 1993), pathologists who were blinded to the patient's clinical history were able to identify the primary site of a metastasis no more than 49% of the time when given a choice among 11 adenocarcinomas. When asked to rank their guesses, the correct diagnosis was among the top 3 choices just 76% of the time.

In this paper, we used the largest collection of uniformly processed primary cancer whole genomes assembled to date to develop a supervised machine learning system capable of accurately distinguishing 24 major tumour types based solely on features that can be derived from DNA sequencing. The accuracy of the system overall when applied in a cross-validation setting was 91%, with 20 of the 24 tumour types achieving an F1 score of 0.83 or higher. When the tumour type predictions were ranked according to their probability scores, the correct prediction was found among the top three rankings 98% of the time. When applied to external validation data sets, the classifier achieved predictive accuracies of 88% and 83% respectively for primary and metastatic tumours. The modestly reduced accuracy in the validation sets is likely due to their differing somatic mutation-calling pipelines, which used different quality-control filters, genome builds and SNV callers from the specimens in the training set.

The regional distribution of somatic passenger mutations across the genome was the single most predictive class of feature, followed by the distribution of mutation types. The regional density of somatic mutations is thought to reflect chromatin accessibility to DNA repair complexes, which in turn relates to the epigenetic state of the cancer's cell of origin. The DNN's predictive accuracy is therefore largely driven by a cell of origin signal, aided to a lesser extent by signatures of exposure. The observation that the classifier was able to identify the site of origin for metastatic and primary tumours with similar accuracy suggests that the cell of origin and exposure signals are already established in the early cancer (or its precursor cell) and are not masked by subsequent mutations that occur during tumour evolution.

Unexpectedly, the distribution of functional mutations across driver genes and pathways were poor predictors of tumour type in all but a few tumour types. This surprising finding may be explained by the observation that there are relatively few driver events per tumour (mean 4.6 events per tumour (Sabarinathan et al. 2017)), and affect a set of common biological pathways related to the hallmarks of cancer (Hanahan and Weinberg 2011). This finding may also explain the observation that automated prediction of tumour type by exome or gene panel sequencing has so far met with mixed success (see below).

There was considerable variability in the classification accuracy among tumour types. In most cases tumour types that were frequently confused with each other had biological similarities such as related tissues or cells of origin. Technical issues that could degrade predictive accuracy include uneven sequencing coverage, low sample purity, inadequate numbers of samples in the training set, and tumour type heterogeneity. Mutational patterns associated with exposure to chemotherapy may also have an impact on overall performance. Many of the samples from the HMF dataset were sequenced following exposure to chemotherapy, which has a marked effect on mutation types observed in a tumour (Pich et al. 2019). Statistical methods to remove mutations associated with chemotherapy and sequencing artefacts may improve model performance. A larger collection of tumours with WGS would allow us to improve the classifier accuracy as well as to train the classifier to recognize clinically-significant subtypes of tumours.

There are other ways of identifying the site of origin of a tumour. In cases in which the tumour type is uncertain pathologists frequently apply a series of antibodies to tissue sections to detect tissue-specific antigens via immunohistochemistry (IHC). The drawback of IHC is that it requires manual interpretation, and the decision tree varies according to the differential diagnosis (D'cruze et al. 2013). Furthermore, IHC is known to be confounded by the loss of antigens in poorly differentiated tumours (Bahrami, Truong, and Ro 2008). In principle, tumour differentiation state should not impact the performance of our classifier because it relies on the distribution of passenger mutations, most of which are already established at the time of tumour initiation. Because of the many different grading systems applied across the PCAWG set a direct test of this notion is difficult, but we are reassured that the independent set of metastases, which frequently represent a higher grade than the primary, performed as well as the external primary tumour validation set.

An alternative to IHC is molecular profiling of tumours using mRNA or miRNA expression, and several commercial systems are now available to identify the tissue of origin using microarray or qRT-PCR assays (Ferracin et al. 2011; Monzon and Koen 2010; Bridgewater et al. 2008). A recent comparative review (Monzon and Koen 2010) of five commercial expression-based kits reported overall accuracies between 76 and 89%; the number of tumour types recognized by each system ranges from six to 47 with accuracy tending to decrease as the number of discriminated types increases.

Patterns of DNA methylation are also strongly correlated with the tissue of origin. A recent report (Capper 2018) demonstrated highly accurate classification of more than 70 central nervous system tumour types using a Random Forest classifier trained on methylation array data. Another recent report (Shen 2018) showed that an immunoprecipitation-based protocol can recover circulating tumour DNA from patient plasma and accurately distinguish among three tumour types (lung, pancreatic and AML) based on methylation patterns.

Previous work in the area of DNA-based tumour type identification has used targeted gene panel (Tothill 2013) and whole exome (Chen et al. 2015; Soh et al. 2017; Marquard 2015) sequencing strategies. The targeted gene-based approach described in Tothill (Tothill 2013) is able to discriminate a handful of tumour types that have distinctive driver gene profiles, and can identify known therapeutic response biomarkers, but does not have broader applicability to the problem of tumour typing. In contrast, the whole exome sequencing approaches reported by Marquard (Marquard 2015), Chen (Chen et al. 2015), and Soh (Soh et al. 2017) and each used machine learning approaches to discriminate among 10, 17, and 28 primary sites respectively, achieving overall accuracies of 69%, 62%, and 78%. Interestingly, all three papers demonstrated that classifiers built on multiple feature categories outperformed those built on a single type of feature, consistent with our findings. We demonstrate here that the addition of whole genome sequencing data substantially improves discriminative ability over exome-based features. It is also worth noting that Soh (Soh et al. 2017) was able to achieve good accuracy using SNVs and CNAs spanning just 50 genes, suggesting that it may be possible to retain high classifier accuracy while using mutation ascertainment across a well-chosen set of whole genomic regions.

In practical terms, whole genome sequencing and analysis of cancers is becoming increasingly cost effective, and there is an accelerating trend to apply genome sequencing to routine cancer care in order to identify actionable mutations and to test for the presence of predictive biomarkers. An example of the trend is the National Health Service of the UK, which recently announced a plan to apply WGS routinely to cancer patients (Sample 2018). Given the increasing likelihood that many or most cancers will eventually have genomic profiling, it is attractive to consider the possibility of simultaneously deriving the cancer type using an automated computational protocol. This would serve as an adjunct to histopathological diagnosis, and could also be used as a quality control check to flag the occasional misdiagnosis or to find genetically unusual tumours. More forward-looking is the prospect of accurately determining the site of origin of circulating cell-free tumour DNA detected in the plasma using so-called liquid biopsies (Chu and Park 2017), possibly in conjunction with methylome analysis(Capper 2018; Shen 2018). As genome sequencing technologies continue to increase in sensitivity and decrease in cost, there are realistic prospects for blood tests to detect early cancers in high risk individuals (Han, Wang, and Sun 2017). The ability to suggest the site and histological type of tumours detected in this way would be invaluable for informing the subsequent diagnostic workup.

In summary, this is the first study to demonstrate the potential of whole genome sequencing to distinguish major cancer types on the basis of somatic mutation patterns alone. Future studies will focus on improving the classifier performance by training with larger numbers of samples, subdividing tumour types into major molecular subtypes, adding new feature types, and adapting the technique to work with clinical specimens such as those from formalin-fixed, paraffin-embedded biopsies and cytologies.

2.5 Materials and methods

PCAWG Training and Testing Data Set

All variant call data were downloaded from the ICGC Portal, and all file names given here are relative to this path. Note that controlled tier access credentials are required from the ICGC and TCGA projects as described in PCAWG-Data. The consensus Somatic SNV and INDEL files (`consensus_snv_indel/final_consensus_snv_indel_passonly_icgc.open.tgz`and) covers 2778 whitelisted samples from 2583 donors. Consensus SV calls from the PCAWG Structural Variation Working Group were downloaded in VCF format (`consensus_sv/final_consensus_sv_vcfs_passonly.icgc.controlled.tgz` and `final_consensus_sv_vcfs_passonly.tcga.controlled.tgz`). Ploidy and purity information are from the PCAWG Evolution and Heterogeneity Working Group and driver events were called by the PCAWG Drivers and Functional Interpretation Group (Rheinbay et al. 2020). Tumour histological classifications were reviewed and assigned by the PCAWG Pathology and Clinical Correlates Working Group (annotation version 9, August 2016; `clinical_and_histology/pcawg_specimen_histology_August2016_v9.xlsx`). For model training, we first removed all samples that had been flagged as exhibiting microsatellite instability (MSI) by the PCAWG Technical Working Group (`msi/MS_analysis.PCAWG_release_v1.RIKEN.xlsx`). In a small number of cases, the same donor contributed both primary and metastatic tumour specimens to the PCAWG data set. In these cases we used only the primary tumor for training and evaluation, except for the case of the small cohort of myeloproliferative neoplasms (Myeloid-MPN; N=55 samples), for which multiple primary samples were available. In this case, we used up to two samples per donor and partitioned the training and testing sets to avoid having the same donor appear more than once in any training/testing set trial.

Independent validation data set: Primary and Metastatic Tumours

To independently validate the neural network-based classifier, we assembled several sets of tumours that had been subject to whole genome sequencing outside of PCAWG (Table 2.5). The primary tumour validation data set consisted of 1236 primary tumours contributed by colleagues participating in the PCAWG Mutational Signatures Working Group and described in (Alexandrov et al. 2020). These represent 12 tumour types overlapping with PCAWG types collected from a variety of published studies, non-PCAWG donors submitted to the ICGC data portal (`http://dcc.icrg.org`), and donors present in the COSMIC database (`http://cancer.sanger.ac.uk/cosmic`). These independent primaries were supplemented using WGS data from 200 advanced primary pancreatic ductal adenocarcinomas (Panc-AdenoCA) derived from the COMPASS Trial (Aung 2018) and used with the gracious permission of Dr. Steven Gallinger. In all, the primary tumour validation set contained 1436 primary tumour samples across 12 tumour types. Only tumour types with 10 or more representatives were used for testing.

The metastatic tumour validation data set was derived from SNV calls on 2028 metastatic tumours across 16 tumour types, provided by the Hartwig Medical Foundation (HMF data set). They are a subset of 2090 total samples provided by Dr. Edwin Cuppen with matched PCAWG histology subtypes and are described in Table 2.5 and Priestley (Priestley et al. 2019). We supplemented this set with 92 metastatic pancreatic ductal adenocarcinomas to the liver from the COMPASS Trial, for a total of 2120 metastatic tumours. As for the primaries, only tumour types with 10 or more representatives were tested. Although the sequencing technologies and genome coverage are comparable among the PCAWG training set and the independent validation data sets, a mixture of different human genome builds, alignment algorithms and SNV calling algorithms were used for the validation data sets. We did not attempt to recall the SNVs, but did lift the genome coordinates of samples that had been aligned to other genome builds to hg19 by CrossMap (Version 0.2.5).

Human Studies Approval

All patients who donated to the PCAWG, COMPASS and HMF data sets consented to international data sharing and secondary analysis of their genomes (Aung 2018; Priestley et al. 2019; “International network of cancer genome projects” 2010). Permission to reanalyze these data was granted by the University of Toronto’s Research Ethics Board.

Somatic Mutation Feature Sets

Mutational type features are based on all point substitutions (single nucleotide variations; SNVs). For each sample, SNVs are categorized across the six possible single nucleotide changes (A >C, A >G, A >T, C >A, C >G, C >T), the 48 possible nucleotide changes plus their 5′ or 3′ flanking base, and the 96 possible nucleotide changes plus both flanking nucleotides. This generates 150 mutational type features in total. The counts in each category are then normalized to the total number of SNVs in the sample, and then represented as Z-scores.

Mutational distribution features are the number of SNVs, small indels, structural variation (SV) breakpoints, and somatic copy number variations (CNVs) in each 1 megabase bins across the genome. The total number of SNV, indel and SV counts in each bin were normalized to the total number of the corresponding mutational events across the genome. When the model being used is a deep neural network, SNV distribution features were represented as the raw SNV counts. In addition, we generated the following features: (1) the total numbers of each type of mutational event per genome; (2) the number of each type of mutational event per chromosome, normalized by chromosome length; (3) sample purity values; and (4) sample ploidy. In total, there are 2897 SNV+indel, 2826 CNV, and 2929 SV features. For the initial selection of feature types, we tested all mutational distribution features. However, the final neural network used SNV features only.

Driver gene and pathway features were derived from the driver event list generated by the PCAWG Drivers and Functional Interpretation Working Group (Rheinbay et al. 2020). This list contains driver events in coding genes, as well as events that affect miRNA and lncRNAs. We generated a boolean matrix from the list in which each row is a tumour sample and each column is a driver event. To mutations to pathways, we selected any non-synonymous SNV affecting a gene in a pathway, regardless of its putative driver status. These SNVs were then assigned to 1,865 pathways from the Reactome resource (`http://www.reactome.org`, version 58) (Croft 2014). A pathway feature was scored as positive if it contained at least one driver gene. Because a gene may be contained within more than one pathway, it is possible for a single driver gene event to generate two or more positive pathway features.

Machine learning procedure - Random Forest

For each of the 24 cancer types selected from the PCAWG sample set, we first used Random Forest (Breiman 2001) model to train classifiers for each cancer type on each of the feature categories described in the above section. The data sets were Z-score normalized across the samples before training. We used nested cross validation to train and test the performance of the classifiers. In the outer loop, the data set was divided into four folds and each fold was later used as an independent testing set. In the inner loop, the training portion of the data set was split into three folds and each fold was used as validation data set to fine-tune the hyperparameters. In the inner loop, we first used a chi-squared test to filter out non-informative (V coefficient equals to 0) features. Then we tuned two hyperparameters for the Random Forest model to achieve the highest cross-validation F1 score. The two hyperparameters were the sample size for positive versus negative classes and the number of trees. We used the default R randomForest package parameter settings to sample the square root of the number of features at each split of the tree. The code was written in R (version 3.3.0). The main packages used were MLR (version 2.11) and randomForest (4.6-12) in training the model.

Machine learning procedure - Neural Network

We ultimately used a fully-connected, feed-forward neural network for the classification of the 24 cancer types based on SNV type and mutational distribution alone. The network had a softmax output, which can be interpreted as a probability distribution of the 24 types. The predicted tumour type was selected by taking the type with the greatest softmax probability.

We used a Bayesian optimization approach to select hyperparameters (Snoek, Larochelle, and Adams 2012). Prior to training, data from PCAWG was split into training, validation and test sets 10 times to create 10 different partitions over the full dataset. For each of the 10 partitions, hyperparameters were selected by optimizing performance on the validation data for that partition. We used the 'gp_minimize' function from the scikit-optimize 0.5.2 python library (Head et al. 2018) to select the following hyperparameters: learning rate for Adam, L2-regularization penalty (otherwise known as weight decay), dropout rate (Srivastava et al. 2014), the number of hidden layers, the number of neurons per hidden layer, and activation function. Each model was trained using Adam (Kingma and Ba 2014) with a batch-size of 32 for 50 epochs. All hyperparameters of Adam other than learning rate were set to the default values specified in the original paper (Kingma and Ba 2014). Bias values were initialized as 0, and all other network weights were initialized using a glorot uniform distribution (Glorot and Bengio 2010). The model was evaluated with 200 hyperparameter combinations (i.e., 200 calls to 'gp_minimize' were made). Briefly, 'gp_minimize' approximates a function of model performance based on the hyperparameters with a Guassian Process. For each function call to 'gp_minimize', the performance on the current set of hyperparameters is evaluated by training the neural network, and assessing accuracy on the validation set. Based upon this accuracy, the Guassian Process is updated, and a new set of hyperparameters is chosen by optimizing an acquisition function. We used expected improvement as the acquisition function. After hyperparameter optimization, model performance was assessed independently on the corresponding test set for that split. Table 2.7 describes the settings for each of the folds for these hyperparameters. This procedure was repeated for each set of mutational features used for performing classification: SNV Distribution SNV Type and SNV Distribution, Driver Genes and Pathways, and all features. A complete description of the optimal hyperparameters and classification accuracy for each feature set on each of the 10 data partitions is described in Table 2.8. For evaluation on independent datasets, an ensemble of each of the 10 neural networks (one for each data partition) trained using SNV Type and SNV Distribution features was used. The ensemble is constructed by taking the mean of the softmax output from each of the 10 neural networks.

Table 2.7: Hyperparameter ranges for Bayesian optimization.

	Range
Hyperparameter	
Learning Rate	1E-4, 1E-2
L2-Penalty	1E-3, 0.50
Dropout Rate	(1E-06, 0.50
Number of Layers	(0,5)
Number of Neurons/Layer	-51,024
Activation Function	(relu, softplus)

In order to compare the accuracy of these models with models trained on different feature sets, the procedure above was repeated using driver genes/pathways as input, and again by appending the driver genes/pathways features to the SNV features used above. The final hyperparameter values and model accuracies for each of the trained models is described in 2.8. Each model was implemented and trained in Tensorflow 1.10.0 (Abadi et al. 2015) and Keras 2.1.5 (Chollet 2015). All code was written in Python 3.6.

Table 2.8: Hyperparameters selected by Bayesian optimization and their test set accuracy for classifiers trained. Optimal hyperparameter setting found using Bayesian optimization for each feature set. Model refers to the data partition the model was trained on. L2 refers to the L2-penalty value used. Dropout refers to the dropout rate used during training. Layers refers to the number of hidden layers. Units refers to the number of units or neurons in each hidden layer. Activation refers to the activation function used. Accuracy is calculated on the held-out dataset corresponding to the data partition used to train the model.

	Model	Learning Rate	L2	Dropout	Layers	Units	Activation	Accuracy
Features								
SNV Type, Distribution	1	0.000195	0.001000	0.000001	4	1024	relu	90.61%
SNV Type, Distribution	2	0.000224	0.001648	0.000003	3	630	relu	88.62%
SNV Type, Distribution	3	0.000167	0.001000	0.000001	3	1024	relu	92.95%
SNV Type, Distribution	4	0.000139	0.010037	0.284099	3	1024	relu	92.31%
SNV Type, Distribution	5	0.000246	0.001000	0.000001	3	1024	softplus	91.29%
SNV Type, Distribution	6	0.000163	0.001000	0.000001	3	826	relu	91.90%
SNV Type, Distribution	7	0.000219	0.001000	0.000079	4	898	relu	90.87%
SNV Type, Distribution	8	0.000100	0.010160	0.500000	1	871	softplus	90.16%
SNV Type, Distribution	9	0.000100	0.001000	0.000001	4	1024	relu	91.21%
SNV Type, Distribution	10	0.000100	0.001000	0.500000	1	630	softplus	91.43%
SNV Distribution	1	0.000216	0.006148	0.004842	3	565	relu	88.57%
SNV Distribution	2	0.000179	0.004249	0.001972	3	716	relu	88.62%
SNV Distribution	3	0.000100	0.001000	0.000017	3	685	relu	90.04%
SNV Distribution	4	0.000320	0.001000	0.052849	3	594	softplus	89.88%
SNV Distribution	5	0.000100	0.003040	0.000001	5	1024	relu	84.23%
SNV Distribution	6	0.000100	0.010714	0.000014	3	832	relu	87.85%
SNV Distribution	7	0.000100	0.001000	0.000001	5	720	softplus	88.38%
SNV Distribution	8	0.000147	0.001126	0.000114	3	852	softplus	88.11%
SNV Distribution	9	0.000329	0.002539	0.003819	3	555	relu	88.70%
SNV Distribution	10	0.000100	0.005926	0.000001	3	1024	softplus	89.39%
All Features	1	0.000425	0.010749	0.000001	2	607	relu	91.02%
All Features	2	0.000100	0.005008	0.000001	3	817	softplus	91.87%
All Features	3	0.000100	0.007431	0.000068	3	1024	relu	90.04%
All Features	4	0.000100	0.001000	0.000001	5	1024	softplus	89.47%
All Features	5	0.000199	0.002103	0.000628	3	795	softplus	90.46%
All Features	6	0.000399	0.031042	0.000001	2	524	relu	89.88%
All Features	7	0.000149	0.013548	0.000001	3	890	relu	92.12%
All Features	8	0.000103	0.001000	0.000001	3	1024	relu	89.34%
All Features	9	0.000100	0.003077	0.000536	5	784	relu	89.96%
All Features	10	0.000219	0.001000	0.019526	4	1024	softplus	90.61%
Genes/Pathways	1	0.000100	0.012803	0.500000	2	530	softplus	38.78%
Genes/Pathways	2	0.000100	0.081871	0.500000	1	1024	softplus	34.96%
Genes/Pathways	3	0.000100	0.001000	0.500000	3	434	softplus	36.93%
Genes/Pathways	4	0.000100	0.120758	0.000049	1	308	softplus	40.49%

Continued on next page

	Model	Learning Rate	L2	Dropout	Layers	Units	Activation	Accuracy
Features								
Genes/Pathways	5	0.000100	0.067909	0.000003	1	371	relu	35.68%
Genes/Pathways	6	0.000100	0.087498	0.500000	1	194	relu	44.94%
Genes/Pathways	7	0.000103	0.002803	0.499240	3	789	relu	41.49%
Genes/Pathways	8	0.000138	0.480229	0.031914	1	67	relu	43.44%
Genes/Pathways	9	0.000100	0.104290	0.000001	2	470	relu	38.91%
Genes/Pathways	10	0.000100	0.018117	0.500000	2	358	softplus	39.59%

Definitions of Accuracy Metrics

To measure the performance of the classifiers, we use the conventional definitions of recall, precision, F1 score and accuracy. In the descriptions below, we use the abbreviations TP (true positive), TN (true negative), FP (false positive), and FN (false negative) to describe correct and incorrect assignments of an unknown tumour to a predicted type:

Recall: The proportion of samples of a particular histopathological type that are correctly assigned to that type:

$$Recall = TP/(TP + FN) \tag{2.1}$$

Precision: The proportion of samples assigned to a particular type that are truly that type:

$$Precision = TP/(TP + FP) \tag{2.2}$$

F1 score: The harmonic mean of recall and precision:

$$F1 = 2(recall * precision)/(recall + precision) \tag{2.3}$$

Accuracy: The proportion of correct assignments.

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \tag{2.4}$$

Chapter 3

Addressing challenges for tumour typing in a clinical setting

I implemented and trained all machine learning methods. Wei Jiao curated the complete dataset of tumour samples and assessed accuracy on the two independent validation sets. I carried out all other experiments and analyses.

3.1 Abstract

The process of tumour typing involves correctly identifying a tumour's organ of origin and histopathology. Despite advances in precision oncology, these two features are the strongest determinants of a tumour's clinical behaviour and are essential for properly understanding the tumour's developmental characteristics and therapeutic sensitivity. Whilst cancer type is typically available at the time of diagnosis, 3-5% of cancer patients present with histologically confirmed metastatic spread but no obvious or identifiable primary tumour. This constitutes a cancer of unknown primary, a heterogeneous set of diseases currently the 8^{th} most common cancer diagnosis. Given that somatic mutations vary significantly across different cancer types, somatic mutations may be used as a feature for identifying primary tumour site. A particularly promising approach is to exploit the association between patterns of somatic passenger mutations and cancer type. To this end, I have previously developed a deep learning classifier for tumour typing based on patterns of somatic passenger mutations. The classifier can accurately discriminate between 24 common cancer-types with an overall F1 score of 0.91. Despite the model's performance, several challenges remain for translating it into a clinical setting. First, there are several uncommon and rare cancer types that the classifier cannot currently identify. Second, the model does not provide any metrics related to the model's predictive uncertainty. To address these challenges, I have explored the use of uncertainty quantification in deep learning to extend the model to a greater number of cancer types. To provide calibrated uncertainty estimates, I have assessed the calibration error of the classifiers and used post-hoc model calibration methods to improve model calibration. Finally, using measures of model uncertainty, I have developed a robust method for automatically detecting rare cancer types that the model was not trained to identify.

3.2 Introduction

The process of tumour typing - identifying the organ of origin and histopathology of a tumour - is typically accomplished through histopathologic assessment, molecular testing and imaging studies. More recently, genetic testing often supplements this process, and in some cases, the use of genome sequencing. The use of genetic testing and genome sequencing has allowed for the presence or absence of mutant genes to guide treatment and inform clinicians of a tumour's biological characteristics. Despite these advances in precision oncology, correctly identifying tumour type still forms the basis for understanding the clinical progression of a tumour. While advances in precision oncology have allowed for treatment and the biological characteristics of cancer growth to be informed by the presence or absence of mutant genes, cell of origin remains the strongest predictor of a tumour's clinical behaviour. Studies suggest that cancer-specific therapy based on a tumour's cell of origin is more effective than broad-spectrum chemotherapy and that cell of origin can impact the efficacy of drugs that target specific cancer-associated mutations (Greco 2013; Hyman et al. 2015). In most cases, primary site is available upon diagnosis, but occasionally, patients present with cancers of unknown primary (CUPS). Correctly identifying the primary tumour site for CUPS forms the basis of a significant diagnostic challenge and is a crucial task for correctly guiding treatment. Patients presenting with multiple primary tumours represent a related diagnostic challenge (Travis 2006). In this diagnostic scenario, pathologists are asked to determine if a newly identified tumour is the result of a late metastatic recurrence of a treated primary tumour or if it results from a new, unrelated primary tumour.

Different tumour types contain distinct patterns of somatic mutations. Studies suggest that the spatial distribution of mutations across the genome is non-uniform and that mutation density is strongly associated with repressive chromatin marks (Schuster-Böckler and Lehner 2012; Supek and Lehner 2015; Polak et al. 2015). This suggests that regional mutation density may provide sufficient information for identifying cell-type. Additionally, the study of mutational signatures - representations or proxies for mutational processes that contributed mutations to a tumour genome - has demonstrated a relationship between specific mutational processes and cancer type (Alexandrov et al. 2020). In Chapter 2, I made use of the intuition that somatic mutations carry a significant amount of tumour-specific information to develop a series of deep learning classifiers that can accurately distinguish between multiple cancer types (Jiao et al. 2020). The classifiers were trained on data from the Pan-cancer analysis of Whole Genomes (PCAWG), which had aggregated WGS of 2658 cancer samples across 34 histologically distinct cancer types. These data were uniformly analyzed with the same computational pipelines. The classifier took as input information about regional mutation density and mutational spectra within the tumour and could distinguish between 24 cancer types with an overall accuracy of 91%. Moreover, the classifier generalized well to additional datasets, showing an overall accuracy of 88% on an additional set of primary tumours, and 83% on a dataset of metastatic tumours.

The classifier presented in Chapter 2 demonstrated the utility of somatic passenger mutations as a feature for accurately identifying cancer type. Despite the impressive performance of this classifier, multiple challenges exist for translating the model into a clinical setting. First, more than 24 cancer types exist, and a model in a clinical setting should identify as many cancer types as possible. Second, in a clinical setting where a classifier's predictions may form the basis for guiding decision making for a patient, a classifier should provide robust estimates of predictive uncertainty. Predictive uncertainty refers to two distinct quantities of importance in a clinical setting. One notion of uncertainty refers to in-distribution uncertainty or confidence calibration. This notion of uncertainty quantifies how well the model's predictions reflect the true uncertainty for classes the model is trained to identify. That is, in-distribution uncertainty quantifies the relationship between the model's prediction for an input sample and the ground truth likelihood of correctness. The second notion of uncertainty pertains to the automatic identification of rare cancer types that the model was not trained to identify. This notion of uncertainty, sometimes referred to as out-of-distribution (OOD) detection, uses the model's predictive uncertainty to determine if an input sample is anomalous or different from the data distribution used during model training.

In this chapter, I address these challenges for translating a tumour type classifier into a clinical setting. I implement and benchmark several algorithmic advancements for extending the classifier to a greater number of cancer types. Using some of the same algorithmic improvements, I assess the reliability of the model for accurately quantifying in-distribution uncertainty, and implement and benchmark several methods for improving in-distribution uncertainty estimation. Finally, I use the uncertainty expressed in the classifier's predictive distribution to accurately identify samples from rare, or OOD, cancer types that the model wasn't trained to identify.

3.3 Results

3.3.1 Description of training data

Using data from the Pan-cancer analysis of Whole Genomes (PCAWG), the Hartwig Medical Foundation (HMF) dataset and an independent collection of primary tumour samples, I built a series of tumour type classifiers using sequence-based features derived from somatic SNVs. The confidence calibration of the best performing classifiers was evaluated and improved using post-hoc calibration methods. Furthermore, the best performing classifiers were evaluated for the task of out-of-distribution (OOD) detection on samples of rare or uncommon cancer types that were not included during model training.

The full PCAWG data set consists of 2778 donors comprising 34 main histopathological tumour types (Campbell et al. 2020). All samples in this dataset are uniformly analysed using the same computational pipeline for quality-control filtering, alignment, and somatic mutation calling. For this work, when a model is trained solely on data from PCAWG, I chose a cut off of at least 15 donors per tumour type. In a small number of cases, the same donor contributed both primary and metastatic tumour specimens to the PCAWG data set. In these cases, I used only the primary tumour for training and evaluation, except for the case of the small cohort of myeloproliferative neoplasms (Myeloid-MPN; N=55 samples), for which multiple primary samples were available. In this case, we used up to two samples per donor and partitioned the training and testing sets to avoid having the same donor appear more than once in any training/testing set trial. The PCAWG-only data set consisted of 2566 samples spanning 29 major cancer types. All data used for training, validation and testing are described in Table 3.1 and Table 3.2.

Table 3.1: Distribution of tumour types in the PCAWG training and test data sets.

	Tumor Type	Samples
Abbreviation		
Liver-HCC	Liver hepatocellular carcinoma	306
Panc-AdenoCA	Pancreatic adenocarcinoma	235
Breast-AdenoCA	Breast adenocarcinoma	198
Prost-AdenoCA	Prostate adenocarcinoma	189
CNS-Medullo	Medulloblastoma	146
Kidney-RCC	Renal cell carcinoma (proximal tubules)	143
Ovary-AdenoCA	Ovarian adenocarcinoma	112
Skin-Melanoma	Skin melanoma	106
Lymph-BNHL	Mature B-cell lymphoma	105
Eso-AdenoCA	Esophageal adenocarcinoma	98
Lymph-CLL	Chronic lymphocytic leukemia	95
CNS-PiloAstro	Pilocytic astrocytoma	89
Panc-Endocrine	Pancreatic neuroendocrine tumor	85
Stomach-AdenoCA	Gastric adenocarcinoma	70
Head-SCC	Head/neck squamous cell carcinoma	57
ColoRect-AdenoCA	Colorectal adenocarcinoma	52
Lung-SCC	Lung squamous cell carcinoma	48
Thy-AdenoCA	Thyroid adenocarcinoma	48
Myeloid-MPN	Myeloproliferative neoplasm	46
Kidney-ChRCC	Renal cell carcinoma (distal tubules)	45
Bone-Osteosarc	Sarcoma, bone	44
CNS-GBM	Diffuse glioma	41
Uterus-AdenoCA	Uterine adenocarcinoma	40
Lung-AdenoCA	Lung adenocarcinoma	38
Biliary-AdenoCA	Cholangiocarcinoma; Papillary cholangioca	35
Bone-Leiomyo	Leiomyosarcoma	34
Bladder-TCC	Transitional cell carcinoma; Papillary TCC	23
CNS-Oligo	Oligodendroglioma	18
Cervix-SCC	Squamous cell carcinoma	18
		2564

Table 3.2: Distribution of tumour types in the complete dataset.

	Tumor Type	Samples
Abbreviation		
Ovary-AdenoCA	Ovarian adenocarcinom	232
CNS-PiloAstro	Pilocytic astrocytoma	164

Continued on next page

Abbreviation	Tumor Type	Samples
Liver-HCC	Liver hepatocellular carcinoma	472
Panc-Endocrine	Pancreatic neuroendocrine tumor	122
Kidney-RCC	Renal cell carcinoma (proximal tubules)	182
Prost-AdenoCA	Prostate adenocarcinoma	644
ColoRect-AdenoCA	Colorectal adenocarcinoma	462
Lymph-BNHL	Mature B-cell lymphoma	137
Uterus-AdenoCA	Uterine adenocarcinoma	87
Breast-AdenoCA	Breast adenocarcinoma	1139
Lung-AdenoCA	Lung adenocarcinoma	63
Panc-AdenoCA	Pancreatic adenocarcinoma	686
Eso-AdenoCA	Esophageal adenocarcinoma	233
Head-SCC	Head/neck squamous cell carcinoma	104
CNS-Medullo	Medulloblastoma	204
CNS-GBM	Diffuse glioma	136
Bone-Leiomyo	Leiomyosarcoma	34
Skin-Melanoma	Skin melanoma	468
Lymph-CLL	Chronic lymphocytic leukemia	180
Thy-AdenoCA	Thyroid adenocarcinoma	61
Kidney-ChRCC	Renal cell carcinoma (distal tubules)	45
Stomach-AdenoCA	Stomach adenocarcinoma	100
Lung-SCC	Lung squamous cell carcinoma	49
Bladder-TCC	Transitional cell carcinoma; Papillary TCC	23
Biliary-AdenoCA	Cholangiocarcinoma; Papillary cholangioca	34
Bone-Osteosarc	Sarcoma, bone	137
Myeloid-MPN	Myeloproliferative neoplasm	64
		6262

To account for various sources of noise or variance that result from non-uniform processing of tumour samples, an additional dataset was created by incorporating data from PCAWG, HMF and an independent set of primary tumours (Jiao et al. 2020; Alexandrov et al. 2020; Priestley et al. 2019). These data had significant differences in alignment algorithms and SNV calling algorithms. Additionally, these samples come from a variety of genome builds. Data were not re-aligned, and SNV calling was not redone. When a sample was not aligned to hg19, variants were lifted over so as to maintain the same genome build for all training samples. I chose a cut off of at least 25 donors per tumour type. In total, this dataset, called the ”complete dataset”, consisted of 6262 samples spanning 27 major cancer types. All data used for training, validation and testing are described in Table 3.2.

3.3.2 Classification using data from PCAWG

To determine if the classifier could be extended to 29 cancer types, I trained a deep neural network on data from PCAWG. The classifier was trained using the mutational distribution and mutational types features.

Two different neural network architectures were compared: Deterministic Uncertainty Quantification (DUQ) and Deep Ensemble (See "Methods" for details) (Lakshminarayanan, Pritzel, and Blundell 2017; Amersfoort et al. 2020). Based on the performance of previous work, I used primarily passenger derived features for training the model. Given the strength of the association between chromatin features from the cell of origin and regional mutation density, I made use of the mutation distribution feature for training the model (Jiao et al. 2020). This feature is captured by dividing the genome into 3000 1-Mbp bins across the autosomes and counting the number of somatic SNVs in each bin. Additionally, there is an association between mutation types and cancer type. For example, lung squamous cell carcinomas have a large amount of exposure to mutation types associated with tobacco smoke (Alexandrov et al. 2020). Similarly, skin cancers have mutation types associated with UV-radiation (Alexandrov et al. 2020). Based on this intuition, I generated an additional set of features that represented the normalized frequency of each nucleotide change in the context of its $5'$ and $3'$ neighbours. Each classifier's input was a vector of mutational distribution concatenated with mutational type features. For both DUQ and the deep ensemble, the core of the neural network architecture is a fully-connected feed-forward neural network. The deep ensemble output was a probability estimate that the specimen belongs to each of the 29 cancer types being considered. An input specimen was assigned to the cancer type with the greatest probability. The output of DUQ was a similarity metric between the input specimen and each of the 29 cancer types being considered (Amersfoort et al. 2020). An input specimen was assigned to the most similar cancer type. Ten different training, validation and test partitions were created to evaluate the classifiers, and each classifier was trained independently on its data partition. In the case of the deep ensemble, this resulted in 10 sets of deep ensembles, one for each data partition. Each deep ensemble consisted of 50 neural network classifiers (See "Methods" for details). To improve model performance, calibration and generalization, the training set was balanced by oversampling underrepresented cancer types, and I included adversarial data during model training (Goodfellow, Shlens, and Szegedy 2015). I report the overall accuracy, recall, precision and F1 score using the average across ten held-out test sets, one for each data partition (See "Methods" for a description of data partitions and definition of terms). For evaluation on independent datasets, an ensemble of 10 deep ensembles (one for each data partition) is used.

There was large variability in overall performances across tumour types and the two neural network architectures evaluated. For the model trained with DUQ, the macro-averaged median F1 score (harmonic mean of recall and precision) was 0.74 and had considerable variability across cancer types (Table 3.3). Average F1 score ranged from 0.05 for Biliary-AdenoCA to 0.97 for ColoRect-AdenoCA (Table 3.4). Figure 3.1 shows a heatmap summarizing the accuracy of the DUQ classifier on held-out tumours. Overall, the accuracy across 29 tumour types was 82%, which is substantially lower than the original version of the model, which classifiers 24 tumour types. Recall (otherwise known as sensitivity) ranged from 0.03 (Biliary-AdenoCA) to 0.98 (Myeloid-MPN).

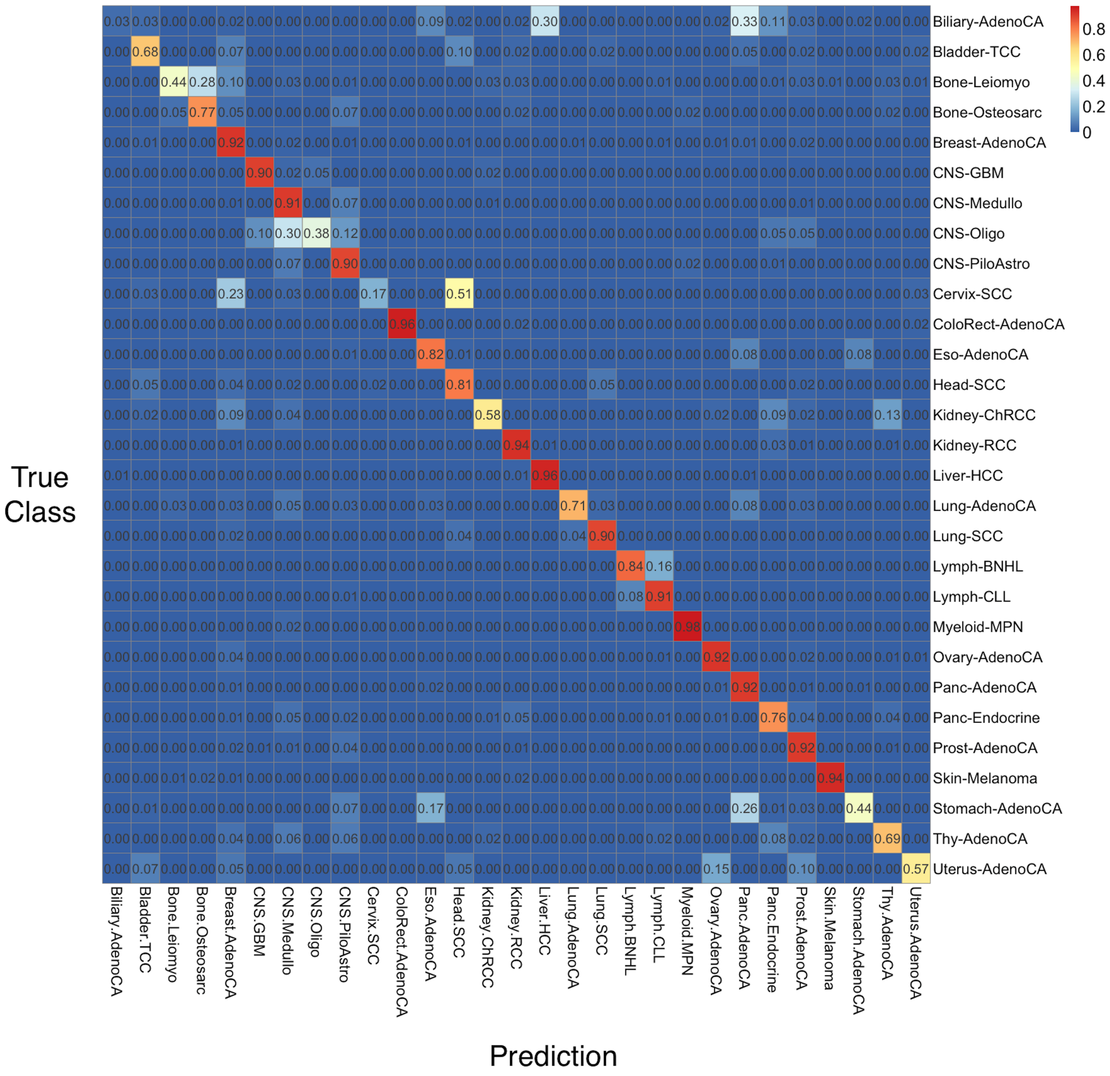


Figure 3.1: Heatmap displaying the accuracy of DUQ classifier using a held-out portion of the PCAWG data set for evaluation. Heatmap displaying the accuracy of the deterministic uncertainty quantification classifier (DUQ) using a held-out portion of the PCAWG data set for evaluation. Each row corresponds to the true tumor type; Columns correspond to the predictions emitted by each of the classifiers. Cells are labeled with the proportion of tumors of a particular type that were called a specific type by the classifier. Due to rounding of values, some rows add up to slightly more or less than 100%.

Table 3.3: Summary of performance metrics for the DUQ classifiers. Recall (specificity), Precision (sensitivity) and F1 score for classification with the DUQ classifier. Performance is averaged over 10 held-out test sets from PCAWG.

	Recall (%)	Precision (%)	F1 score
Cancer Type			
Biliary-AdenoCA	3	50	0.05
Bladder-TCC	68	65	0.67
Bone-Leiomyo	44	90	0.59
Bone-Osteosarc	77	59	0.66
Breast-AdenoCA	92	79	0.86
CNS-GBM	90	88	0.89
CNS-Medullo	91	76	0.83
CNS-Oligo	38	88	0.44
CNS-PiloAstro	90	66	0.77
Cervix-SCC	17	86	0.19
ColoRect-AdenoCA	96	98	0.97
Eso-AdenoCA	82	78	0.80
Head-SCC	81	61	0.69
Kidney-ChRCC	58	81	0.67
Kidney-RCC	94	89	0.92
Liver-HCC	96	93	0.95
Lung-AdenoCA	71	90	0.78
Lung-SCC	90	90	0.90
Lymph-BNHL	84	92	0.88
Lymph-CLL	91	79	0.84
Myeloid-MPN	98	94	0.96
Ovary-AdenoCA	92	88	0.90
Panc-AdenoCA	92	79	0.85
Panc-Endocrine	76	71	0.73
Prost-AdenoCA	92	86	0.88
Skin-Melanoma	94	99	0.97
Stomach-AdenoCA	44	74	0.54
Thy-AdenoCA	69	67	0.66
Uterus-AdenoCA	57	82	0.66

Table 3.4: Summary of performance metrics for the three classifiers compared. Performance metrics for the three classifiers that were trained and compared. Models labeled with "(PCAWG)" are trained on data from PCAWG. The model labeled with "(Complete dataset)" is trained on a collection of data including PCAWG, HMF and independently collected primary tumours.

Model	Macro F1 score	Micro F1 score	Weighted F1 score
Deep Ensemble (PCAWG)	0.83	0.890	0.889
DUQ (PCAWG)	0.74	0.820	0.800
Deep Ensemble (Complete dataset)	0.82	0.897	0.896

Overall, the deep ensemble had a significantly higher F1 score with a macro-averaged F1 score of 0.83 compared to 0.74 (p=2.60e-05, Wilcoxon Signed-Rank test; Table 3.4). The overall accuracy for classifying 29 tumour types was 89% for the deep ensemble method, which is similar to the overall accuracy for classifying 24 cancer types described previously. The discrepancy between macro-averaged F1 score and accuracy is a result of poor overall performance for classifying Biliary-AdenoCA (F1 score = 0.11) and Cervix-SCC (F1 score = 0.27) (Table 3.3). The macro-averaged F1 score after excluding these two cancer types increases to 0.87, comparable to the original version of the classifier. As before, performance varied significantly across cancer types, and F1 score ranged from 0.11 (Biliary-AdenoCA) to 1.00 (Skin-Melanoma). Overall, however, 22 of 29 cancer types had F1 scores of at least 0.80, and 14 cancer types were classified with an F1 score of at least 0.90. Precision (otherwise known as specificity) ranged from 0.64 (Head-SCC) to 1.00 (Skin-Melanoma, Kidney-ChRCC) (Figure 3.2). Recall for the deep ensemble also showed significant variability, ranging from 0.06 (Biliary-AdenoCA) to 1.00 (ColoRect-AdenoCA, Skin-Melanoma) (Figure 3.3). Of the three worst-performing cancer types in the original classifier, CNS-PiloAstro (F1 score = 0.81) and Lung-AdenoCA (F1 score = 0.82) had overall improvements in performance, and only Stomach-AdenoCA saw a decrease in classification performance (F1 score = 0.59 compared to 0.67 previously). A complete summary of the performance for the deep ensemble trained on data from PCAWG is provided in Table 3.4.

Table 3.5: Summary of performance metrics for the deep enesmble classifiers. Recall (specificity), Precision (sensitivity) and F1 score for classification with the deep ensemble classifier. Performance is averaged over 10 held-out test sets from PCAWG.

Cancer Type	Recall (%)	Precision (%)	F1 score
Biliary-AdenoCA	6	67	0.11
Bladder-TCC	86	88	0.87
Bone-Leiomyo	65	86	0.75
Bone-Osteosarc	70	65	0.66
Breast-AdenoCA	99	94	0.96
CNS-GBM	93	78	0.85
CNS-Medullo	97	92	0.94

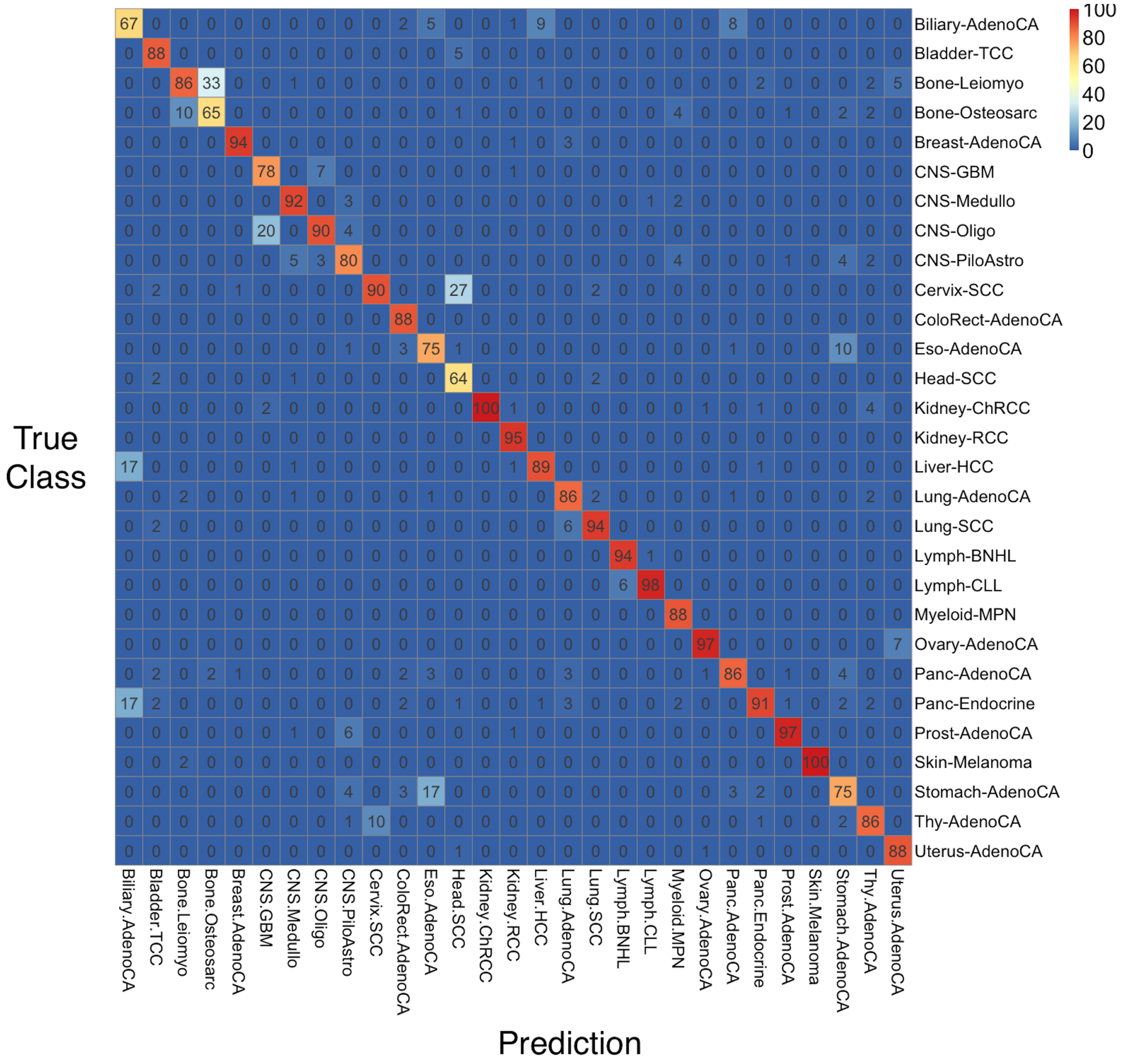
Continued on next page

Cancer Type	Recall (%)	Precision (%)	F1 score
CNS-Oligo	66	90	0.75
CNS-PiloAstro	83	80	0.81
Cervix-SCC	24	90	0.27
ColoRect-AdenoCA	100	88	0.94
Eso-AdenoCA	88	75	0.81
Head-SCC	93	64	0.76
Kidney-ChRCC	84	100	0.89
Kidney-RCC	100	95	0.97
Liver-HCC	98	89	0.94
Lung-AdenoCA	79	86	0.82
Lung-SCC	94	94	0.94
Lymph-BNHL	99	94	0.96
Lymph-CLL	93	98	0.95
Myeloid-MPN	100	88	0.94
Ovary-AdenoCA	97	97	0.97
Panc-AdenoCA	94	86	0.90
Panc-Endocrine	86	91	0.88
Prost-AdenoCA	95	97	0.96
Skin-Melanoma	99	100	1.00
Stomach-AdenoCA	51	75	0.59
Thy-AdenoCA	90	86	0.87
Uterus-AdenoCA	92	88	0.90

3.3.3 Patterns of misclassification

When the deep ensemble made misclassifications, they tended to reflect shared biological characteristics between the tumours that may influence patterns of somatic SNVs in their genomes. Similar to the original classifier, Stomach-AdenoCA is frequently misclassified as oesophageal adenocarcinoma (Eso-AdenoCA, 27% misclassification rate). These organs share a common developmental origin in the embryonic foregut, which may lead to common epigenetic profiles. As the mutation distribution feature aims to represent epigenetic profiles, common cell-of-origin may result in similarities in regional mutation density for these tumour types. Common patterns of mutational signatures, particularly to mutation signature 17, may also provide an explanation for misclassifications (Jiao et al. 2020). Another possible explanation for misclassification for gastric and oesophageal tumours is the difficulty in identifying the origin of tumours that arose at the gastroesophageal junction. As with the original classifier, a small number of misclassifications occur between cancers of the central nervous system (CNS). Specifically, CNS-Oligo is occasionally identified as CNS-GBM (24% misclassification rate), and CNS-PiloAstro is misidentified as CNS-Medullo (9% misclassification rate). CNS-PiloAstro and CNS-Medullo represent the two pediatric tumours in the dataset, suggesting that some of these misclassifications may be due to the relatively low mutation burden for these tumours (Jiao et al. 2020).

Another common source of misclassifications came from Biliary-AdenoCA, which was often identified



	Biliary.AdenoCA	Bladder.TCC	Bone.Leiomyo	Bone.Osteosarc	Breast.AdenoCA	CNS.GBM	CNS.Medullo	CNS.Oligo	CNS.PiloAstro	Cervix.SCC	ColoRect.AdenoCA	Eso.AdenoCA	Head.SCC	Kidney.ChRCC	Kidney.RCC	Liver.HCC	Lung.AdenoCA	Lung.SCC	Lymph.BNHL	Lymph.CLL	Myeloid.MPN	Ovary.AdenoCA	Panc.AdenoCA	Panc.Endocrine	Prost.AdenoCA	Skin.Melanoma	Stomach.AdenoCA	Thy.AdenoCA	Uterus.AdenoCA
Biliary-AdenoCA	67	0	0	0	0	0	0	0	0	0	2	5	0	0	1	9	0	0	0	0	0	0	8	0	0	0	0	0	0
Bladder-TCC	0	88	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bone-Leiomyo	0	0	86	33	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	2	0	0	0	2	5
Bone-Osteosarc	0	0	10	65	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	4	0	0	0	1	0	2	2	0
Breast-AdenoCA	0	0	0	0	94	0	0	0	0	0	0	0	0	0	1	0	3	0	0	0	0	0	0	0	0	0	0	0	0
CNS-GBM	0	0	0	0	0	78	0	7	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CNS-Medullo	0	0	0	0	0	0	92	0	3	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0	0	0
CNS-Oligo	0	0	0	0	0	20	0	90	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CNS-PiloAstro	0	0	0	0	0	0	5	3	80	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	1	0	4	2	0
Cervix-SCC	0	2	0	0	1	0	0	0	0	90	0	0	27	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
ColoRect-AdenoCA	0	0	0	0	0	0	0	0	0	0	88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Eso-AdenoCA	0	0	0	0	0	0	0	0	1	0	3	75	1	0	0	0	0	0	0	0	0	0	1	0	0	0	10	0	0
Head-SCC	0	2	0	0	0	0	1	0	0	0	0	0	64	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
Kidney-ChRCC	0	0	0	0	0	2	0	0	0	0	0	0	0	100	1	0	0	0	0	0	0	1	0	1	0	0	0	4	0
Kidney-RCC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	95	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Liver-HCC	17	0	0	0	0	0	1	0	0	0	0	0	0	0	1	89	0	0	0	0	0	0	0	1	0	0	0	0	0
Lung-AdenoCA	0	0	2	0	0	0	1	0	0	0	0	1	0	0	0	0	86	2	0	0	0	0	1	0	0	0	0	2	0
Lung-SCC	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	94	0	0	0	0	0	0	0	0	0	0	0
Lymph-BNHL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	94	1	0	0	0	0	0	0	0	0	0
Lymph-CLL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	98	0	0	0	0	0	0	0	0	0
Myeloid-MPN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	88	0	0	0	0	0	0	0	0
Ovary-AdenoCA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	97	0	0	0	0	0	0	7
Panc-AdenoCA	0	2	0	2	1	0	0	0	0	0	2	3	0	0	0	0	3	0	0	0	0	1	86	0	1	0	4	0	0
Panc-Endocrine	17	2	0	0	0	0	0	0	0	0	2	0	1	0	0	1	3	0	0	0	2	0	0	91	1	0	2	2	0
Prost-AdenoCA	0	0	0	0	0	0	1	0	6	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	97	0	0	0	0
Skin-Melanoma	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0
Stomach-AdenoCA	0	0	0	0	0	0	0	0	4	0	3	17	0	0	0	0	0	0	0	0	0	0	3	2	0	0	75	0	0
Thy-AdenoCA	0	0	0	0	0	0	0	0	1	10	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	86	0
Uterus-AdenoCA	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	88

Figure 3.2: Heatmap displaying the accuracy of deep ensemble classifier using a held-out portion of the PCAWG data set for evaluation. Heatmap displaying the accuracy of the deep ensemble classifier using a held-out portion of the PCAWG data set for evaluation. Each row corresponds to the true tumor type; Columns correspond to the predictions emitted by each of the classifiers. Cells are labeled with the proportion of tumors of a particular type that were called a specific type by the classifier. The heatmap represents the precision (specificity) of the deep ensemble. Due to rounding of values, some rows add up to slightly more or less than 100%.

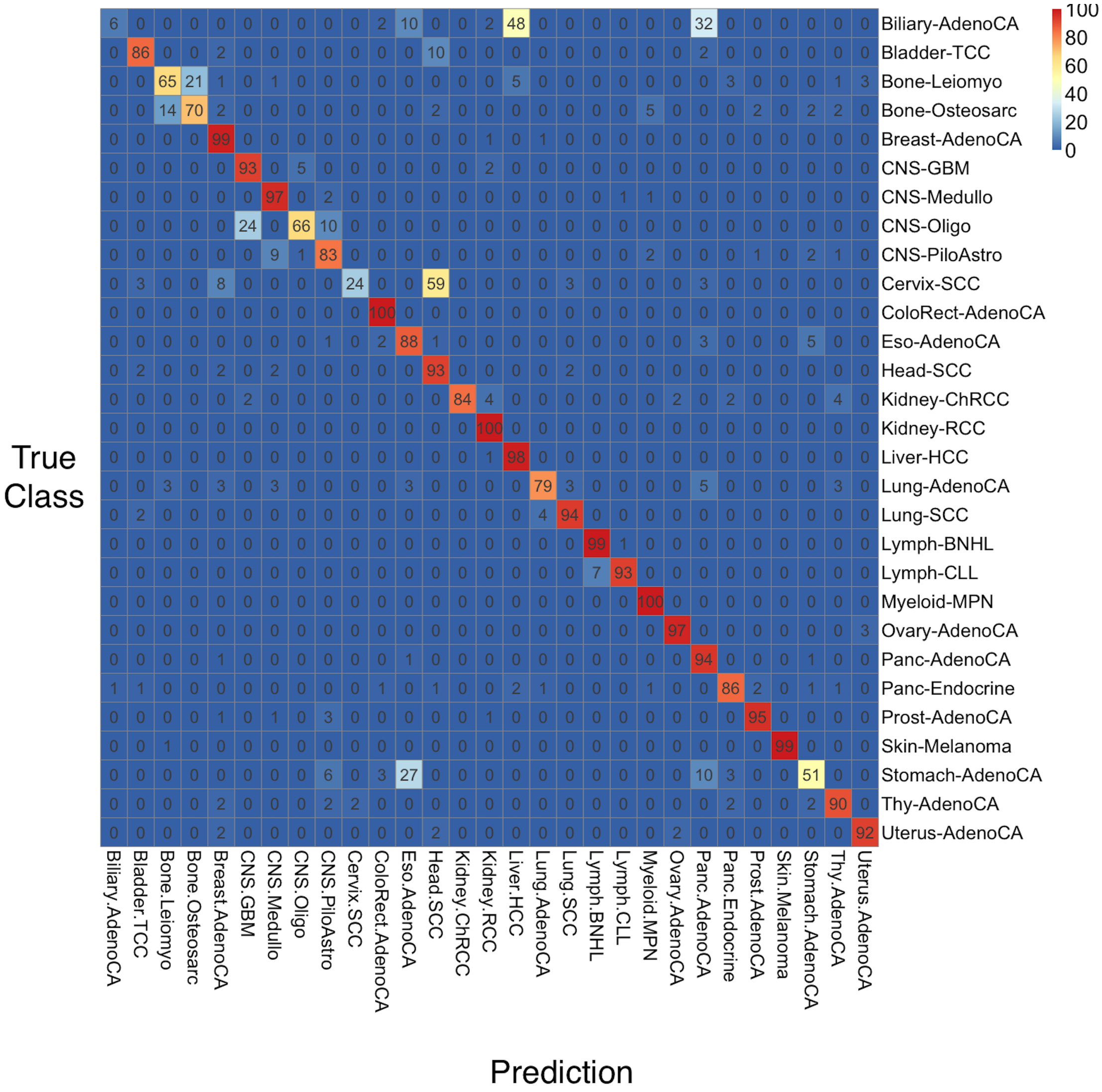


Figure 3.3: Heatmap displaying the accuracy of deep ensemble classifier using a held-out portion of the PCAWG data set for evaluation. Heatmap displaying the accuracy of the deep ensemble classifier using a held-out portion of the PCAWG data set for evaluation. Each row corresponds to the true tumor type; Columns correspond to the predictions emitted by each of the classifiers. Cells are labeled with the proportion of tumors of a particular type that were called a specific type by the classifier. The heatmap represents the recall (sensitivity) of the deep ensemble. Due to rounding of values, some rows add up to slightly more or less than 100%.

as Liver-HCC or Panc-AdenoCA (48% and 32% misclassification rate, respectively). Cholangiocarcinoma (Biliary-AdenoCA) represents a heterogeneous group of cancers that may arise at various points in the biliary tract, which extends from the liver through the pancreas. Cholangiocarcinomas are classified based on their anatomical location, and different subtypes have significant differences in prognosis. One potential explanation for misclassifications comes from similarities in cell-of-origin. Cholangiocarcinoma is the second most common primary cancer of the liver, following hepatocellular carcinoma (Bragazzi et al. 2018). The liver, pancreas and bile duct all share developmental origins in the embryonic foregut, and similar to the stomach and oesophagus, this may result in common mutational distribution features (Faure and De Santa Barbara 2011).

Another large source of misclassifications involves Cervix-SCC and Head-SCC. Cervix-SCC is identified as Head-SCC in the majority of test cases (59% misclassification rate). Multiple biological characteristics are contributing to this result. First, both tumour types are derived from squamous tissue, which may be reflected in shared epigenetic features. Second, most of Cervix-SCC samples in the PCAWG dataset and many of Head-SCC samples in the dataset are HPV-positive tumours (Zapatka et al. 2020). Exposure to HPV has been demonstrated to cause significant alterations in mutational signatures, which may contribute strongly to the model's misclassifications. HPV infection is also associated with alterations in epigenetic features, which, once again, maybe reflected in shared mutational distribution features for these tumours (Karimzadeh et al. 2020).

Despite these difficulties, the deep ensemble can accurately discriminate several tumour types from the same organ. Squamous and adenocarcinoma forms of non-small cell lung cancer (Lung-SCC, Lung-AdenoCA) are readily distinguished (4% misclassification rate). Similarly, Renal cell carcinoma (Kidney-RCC) and chromophobe renal carcinoma (Kidney-ChRCC) are rarely misclassified, with an overall misclassification rate of only 4%. Good performance is seen when distinguishing between chronic lymphocytic leukaemia (Lymph-CLL) and B-cell non-Hodkin's lymphoma (Lymph-BHNL), which are derived from the B-cell lymphocyte lineage. Similarly, cancers of the exocrine and endocrine pancreas are perfectly distinguished by the deep ensemble. Overall, the patterns of misclassifications and the ability to accurately discriminate between cancers that arise in the same organ is similar to the classifier presented in Chapter 2.

3.3.4 Validation on independent set of primary tumours

In real-world settings, such as deployment in a clinical setting, tumour genome samples may be processed using various computational methods. As the deep ensemble model is trained on a uniformly processed dataset, the training set's noise distribution may not reflect the noise distribution of data seen in clinical deployment scenarios. To determine if the deep ensemble could generalize to additional data, I applied the classifier to an independent validation set of 1461 cancer whole genomes assembled from non-PCAWG projects. This dataset spans 14 cancer types from 21 publications or databases. SNV coordinates were lifted from GRCh38 to GRCh37 when necessary, but no other processing of mutation sets was done.

The deep ensemble recall ranged from 0.27 (pediatric gliomas) to 1.00 (Ovary-AdenoCA) (Figure 3.5). Overall accuracy was 84% for classification across the range of cancer types. Generally, cancer types that performed well on data from PCAWG had a similarly strong performance on this validation set. These include Ovary-AdenoCA, Breast-AdenoCA, Panc-AdenoCA, Kidney-RCC, Skin-Melanoma. Interestingly, CNS-PiloAstro samples showed stronger performance on this dataset than on PCAWG. Four cancer types were classified with recall less than 70%: CNS-Medullo, Lymph-CLL, Liver-HCC and Pediatric Gliomas. Compared to the original classifier results, Lymph-CLL samples had many more misclassifications using the deep ensemble. Reassuringly, almost all misclassified Lymph-CLL samples were identified as Lymph-BHNL (46% misclassification rate). Interestingly, Liver-HCC samples were most often misclassified as CNS-Medullo samples with the original classifier, likely due to Liver-HCC samples in this dataset having significantly fewer mutations than Liver-HCC samples from PCAWG (Jiao et al. 2020). However, Liver-HCC samples are most commonly misidentified as Biliary-AdenoCA, likely reflecting shared biological characteristics of these cancer types with the deep ensemble.

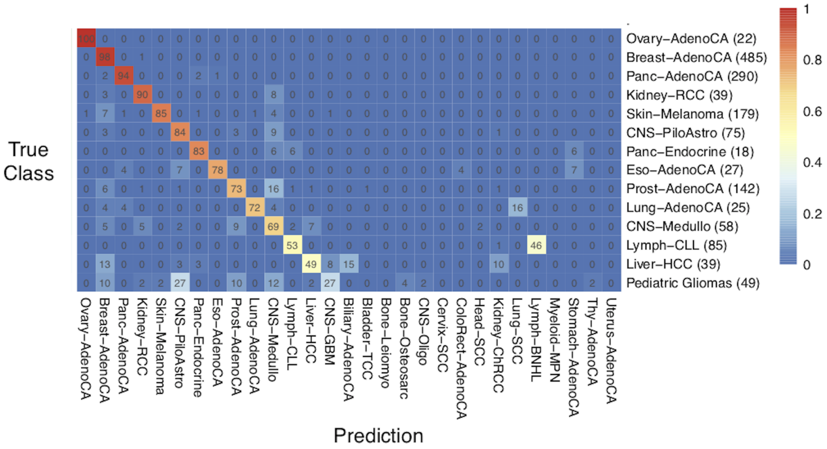


Figure 3.4: Heatmap displaying the accuracy of deep ensemble classifier on an independent set of primary tumours. Heatmap displaying the accuracy of the deep ensemble classifier using a held-out portion of the PCAWG data set for evaluation. Each row corresponds to the true tumor type; Columns correspond to the predictions emitted by each of the classifiers. Cells are labeled with the proportion of tumors of a particular type that were called a specific type by the classifier. The heatmap represents the precision (specificity) of the deep ensemble. Due to rounding of values, some rows add up to slightly more or less than 100%.

3.3.5 Validation on metastatic tumours

To evaluate the deep ensemble model's ability to identify primary tumour site for metastatic tumours correctly, I tested the model on an independent set of metastatic tumours. This data set consisted of 92 metastatic Panc-AdenoCA samples, and an additional 2175 metastatic tumours sequenced by the Hartwig Medical Foundation (HMF), resulting in 2267 metastatic samples from 18 cancer types. All metastatic samples were subjected to paired-end WGS of tumour and normal at a tumour coverage of at least 65x, but did not use the same computational pipelines for alignment, quality-control filtering and SNV calling as those used on data generated by PCAWG. Samples from HMF were obtained using a needle biopsy, limiting spatial heterogeneity in the sequenced sample. Many HMF samples were sequenced following exposure to chemotherapy (Pich et al. 2019). The rules for matching classifier output to the validation set class labels were the same as those previously developed for validating the original version of the classifier (Jiao et al. 2020).

When the deep ensemble was applied to these metastatic samples, the overall accuracy was 83.5% for identifying primary tumour type. Figure 3.6 contains a heatmap summarizing performance on this dataset. This is similar to the accuracy on these samples for the original classifier but is now a classification task with 18 cancer types compared to 16 cancer types before. Compared with the original classifier, the deep ensemble now identifies eight cancer types with recall rates of at least 0.80. This includes: CNS-GBM (0.98), Kidney (0.97), Breast-AdenoCA (0.96), ColoRect-AdenoCA (0.94), Panc-AdenoCA (0.91), Skin-Melanoma (0.87), Lung (0.87), Prost-AdenoCA (0.86). Five tumour types failed to have recall rates of at least 0.50, including Head-SCC (0.43), Uterus-AdenoCA(0.34), Stomach-AdenoCA (0.17), Thy-AdenoCA (0.15) and Biliary-AdenoCA (0.06). Overall, the patterns of misclassification are similar to those seen on data from PCAWG. For example, Stomach-AdenoCA is misclassified as Eso-AdenoCA with a 70% misclassification rate. Similarly, Biliary-AdenoCA is often misclassified as Panc-AdenoCA (45% misclassification rate) and Liver-HCC (23% misclassification rate). The difficulty in correctly identifying Biliary-AdenoCA lowers the overall accuracy of the deep ensemble on this dataset.

While performance on these data was similar to performance on held-out data from PCAWG, the dataset of metastases remains more difficult to classify than primary tumour samples. A potential explanation for this comes from the fact that many of the samples in the HMF dataset were sequenced following exposure to chemotherapy (Pich et al. 2019). Exposure to chemotherapy can result in marked differences in the mutational types seen within a tumour genome, which may reduce performance (Angus et al. 2019; Kucab et al. 2019). A study examining mutational signatures associated with chemotherapy in this dataset suggests that chemotherapy-associated mutations ranged from 1% to 65% in these samples (Pich et al. 2019). This suggests that exposure to chemotherapy can significantly affect the overall tumour mutation burden and mutation types. As mutation types are a feature used for classification, exposure to chemotherapy may be a potential explanation for misclassifications seen in this dataset.

3.3.6 Classification performance using a combined dataset

To determine if incorporating data from other sources could improve performance, I created a complete dataset consisting of samples from PCAWG, the independent set of primary tumour samples, and data from HMF. For this dataset, I restricted myself to 27 cancer types, each with data from at least 25 donors. This left 6262 samples in total. For model training and evaluation, I created a training, validation and test set. Given the strong performance of the deep ensemble model, I trained an additional deep ensemble

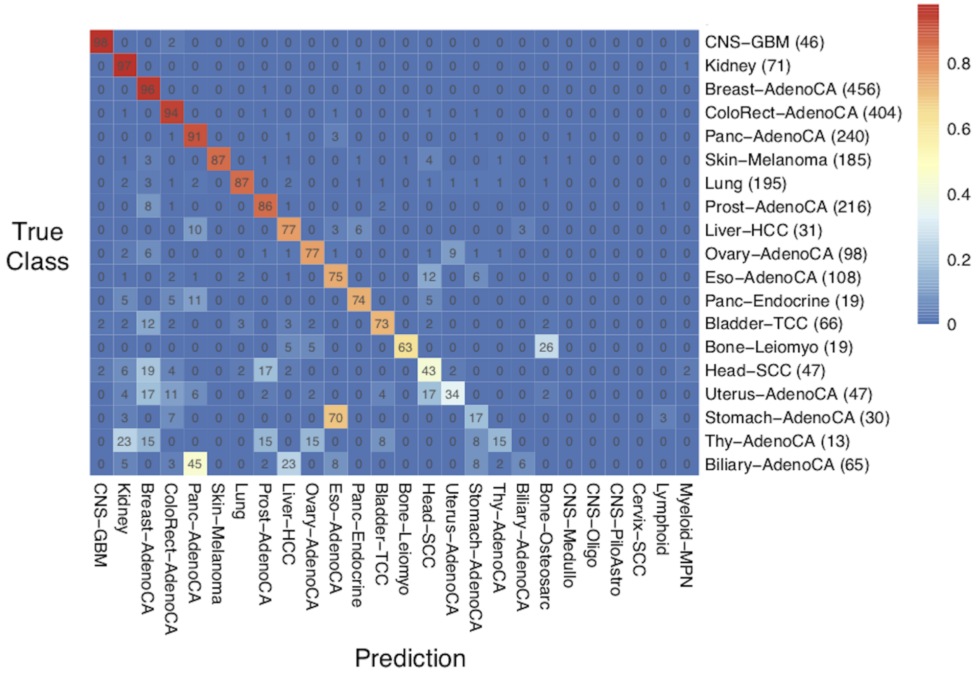


Figure 3.5: Heatmap displaying the accuracy of deep ensemble classifier on an independent set of metastatic tumour samples. Heatmap displaying the accuracy of the deep ensemble classifier using a held-out portion of the PCAWG data set for evaluation. Each row corresponds to the true tumor type; Columns correspond to the predictions emitted by each of the classifiers. Cells are labeled with the proportion of tumors of a particular type that were called a specific type by the classifier. The heatmap represents the precision (specificity) of the deep ensemble. Due to rounding of values, some rows add up to slightly more or less than 100%.

on this dataset.

Figure 3.6 summarizes the performance of this model. Overall, the macro-averaged F1 score was 0.83, and the overall accuracy was 89%. As with the other models tested, there was significant variability in performance across cancer types. Median F1 scores ranged from 0.32 (Stomach-AdenoCA) to 0.99 (Skin-Melanoma, Kidney-RCC) (3.6). 18 of 27 cancer types were classified with an F1 score of at least 0.80, which is fewer cancer types than the original classifier, and the deep ensemble trained solely on data from PCAWG. Figure 3.7 contains a heatmap summarizing the results of the classifier trained on this dataset. Performance on both lung cancer variants (Lung-AdenoCA and Lung-SCC) dropped when incorporating data from different sources. For the deep ensemble trained solely on data from PCAWG, F1 scores for lung cancer were 0.82 and 0.94 for Lung-AdenoCa and Lung-SCC, respectively, compared to 0.73 and 0.75 on the complete dataset. This may be due to chemotherapy-associated mutations and differences in how cancer samples were labelled between the HMF dataset and PCAWG. Overall, patterns of misclassification did not follow those seen in data from PCAWG. For example, Lung-AdenoCA is most often misclassified as Breast-AdenoCA and Kidney-ChRCC (17% misclassification rate for both). Lymph-BHNL is misclassified most often as Bone-Osteosarc (21% misclassification rate). The

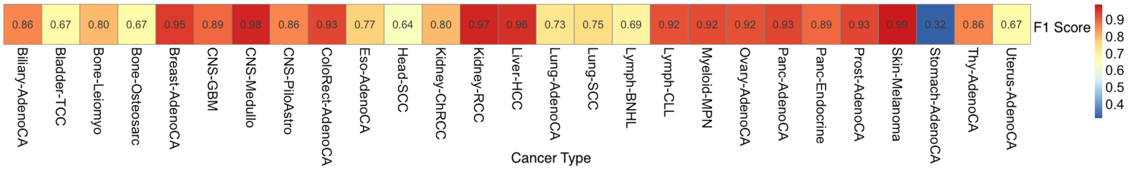


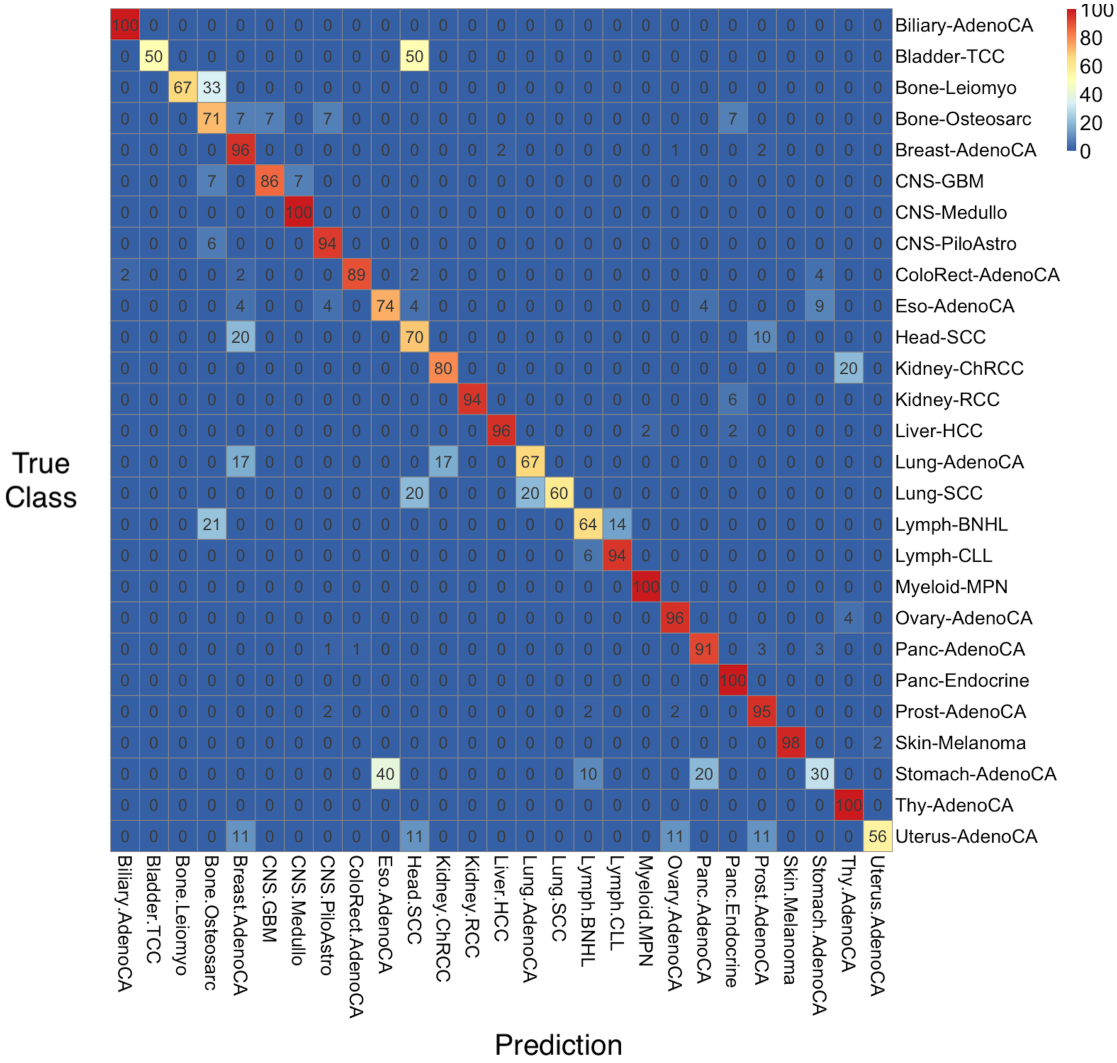
Figure 3.6: Cancer-specific F1 score for the deep ensemble classifier trained on the complete dataset. Heatmap displaying the F1 score for each cancer type the deep ensemble was trained on. The model was trained on a dataset containing samples from PCAWG, HMF and an independent validation set of primary tumours. Cells are labeled with the F1 score on a held-out test set.

lack of apparent biological relatedness found in the misclassifications may result from mutation types caused by chemotherapy exposure. While the overall accuracy of this model is slightly higher than the model trained on data solely from PCAWG, the macro-averaged F1 score drops slightly, and only 18 of 27 cancer types have F1 scores of at least 0.80 compared with 22 of 29 for the model trained on data from PCAWG only. Overall, incorporating data from different sources can provide some modest benefits but tends to reduce accuracy on several cancer types.

3.3.7 Model calibration

In cost-sensitive scenarios, such as clinical decision making for individual patients, a neural network must quantify the uncertainty in the predictions it makes. As the output of the deep ensemble model can be viewed as a probability distribution over cancer types, the neural network output provides an estimate of the probability that an input sample is one of the 29 cancer types the model is trained to identify. Given the importance of high-confidence predictions for clinical decision making, I sought to estimate and improve the confidence calibration and classwise calibration of the model's output probability distribution. Expected calibration error (ECE) and classwise-expected calibration error (classwise-ECE) were assessed for the deep ensemble model trained on data from PCAWG and on the deep ensemble trained with the complete dataset (See ”Methods” for details). To improve calibration performance, I applied four post-hoc calibration methods to both deep ensemble models: Temperature scaling, matrix scaling, vector scaling and Dirichlet scaling (See ”Methods” for details). Overall, the classwise-ECE of the deep ensemble and the temperature scaled model trained on data from PCAWG was lower than the other post-hoc calibration method, with classwise-ECE of 0.012 for the deep ensemble and 0.0081 for the temperature scaled model (Figure 3.9). While temperature scaling reduced the classwise-ECE, it failed to reduce the overall ECE for the deep ensemble model (0.0062 for the deep ensemble compared to 0.0071 for the temperature scaled model) (Table 3.5). This resulted in the deep ensemble model, not the temperature scaled model, with overall ECE consistent with the ECE of a perfectly calibrated classifier (p-value = 0.99, p-value = 0.00, for the deep ensemble and temperature scaled model, respectively; permutation test).

The deep ensemble trained on a complete collection of data showed slightly different results (Figure 3.10). Overall, the uncalibrated deep ensemble's classwise-ECE was lower than that for the temperature scaled model (0.0089 vs 0.0099 for the deep ensemble and temperature scaled model, respectively). Unlike the model trained solely on PCAWG, temperature scaling reduced the overall ECE on this dataset from 0.044 for the uncalibrated model to 0.0032. However, in both cases, ECE values were consistent with



	Biliary.AdenoCA	Bladder.TCC	Bone.Leiomyo	Bone.Osteosarc	Breast.AdenoCA	CNS.GBM	CNS.Medullo	CNS.PiloAstro	ColoRect.AdenoCA	Eso.AdenoCA	Head.SCC	Kidney.ChRCC	Kidney.RCC	Liver.HCC	Lung.AdenoCA	Lung.SCC	Lymph.BNHL	Lymph.CLL	Myeloid.MPN	Ovary.AdenoCA	Panc.AdenoCA	Panc.Endocrine	Prost.AdenoCA	Skin.Melanoma	Stomach.AdenoCA	Thy.AdenoCA	Uterus.AdenoCA
Biliary-AdenoCA	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bladder-TCC	0	50	0	0	0	0	0	0	0	0	50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bone-Leiomyo	0	0	67	33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bone-Osteosarc	0	0	0	71	7	7	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0
Breast-AdenoCA	0	0	0	0	96	0	0	0	0	0	0	0	0	2	0	0	0	0	0	1	0	0	2	0	0	0	0
CNS-GBM	0	0	0	7	0	86	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CNS-Medullo	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CNS-PiloAstro	0	0	0	6	0	0	0	94	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ColoRect-AdenoCA	2	0	0	0	2	0	0	0	89	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0
Eso-AdenoCA	0	0	0	0	4	0	0	4	0	74	4	0	0	0	0	0	0	0	0	0	4	0	0	0	9	0	0
Head-SCC	0	0	0	0	20	0	0	0	0	0	70	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0
Kidney-ChRCC	0	0	0	0	0	0	0	0	0	0	0	80	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0
Kidney-RCC	0	0	0	0	0	0	0	0	0	0	0	0	94	0	0	0	0	0	0	0	0	6	0	0	0	0	0
Liver-HCC	0	0	0	0	0	0	0	0	0	0	0	0	0	96	0	0	0	0	2	0	0	2	0	0	0	0	0
Lung-AdenoCA	0	0	0	0	17	0	0	0	0	0	0	17	0	0	67	0	0	0	0	0	0	0	0	0	0	0	0
Lung-SCC	0	0	0	0	0	0	0	0	0	0	20	0	0	0	20	60	0	0	0	0	0	0	0	0	0	0	0
Lymph-BNHL	0	0	0	21	0	0	0	0	0	0	0	0	0	0	0	0	64	14	0	0	0	0	0	0	0	0	0
Lymph-CLL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	94	0	0	0	0	0	0	0	0	0
Myeloid-MPN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0
Ovary-AdenoCA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	96	0	0	0	0	0	4	0
Panc-AdenoCA	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	91	0	3	0	3	0	0
Panc-Endocrine	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0
Prost-AdenoCA	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	2	0	0	2	0	0	95	0	0	0	0
Skin-Melanoma	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	98	0	0	2
Stomach-AdenoCA	0	0	0	0	0	0	0	0	0	40	0	0	0	0	0	0	10	0	0	0	20	0	0	0	30	0	0
Thy-AdenoCA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0
Uterus-AdenoCA	0	0	0	0	11	0	0	0	0	0	11	0	0	0	0	0	0	0	0	11	0	0	11	0	0	0	56

Figure 3.7: Heatmap displaying the accuracy of deep ensemble classifier trained on the complete dataset. Heatmaps displaying the accuracy of the deep ensemble classifier that was trained on a combination of data from PCAWG, HMF and an independent set of primary tumours. Results are based on held-out data from this complete dataset. Each row corresponds to the true tumor type; Columns correspond to the predictions emitted by each of the classifiers. Cells are labeled with the proportion of tumors of a particular type that were called a specific type by the classifier. The heatmap represents the recall (sensitivity) of the deep ensemble. Due to rounding of values, some rows add up to slightly more or less than 100%.

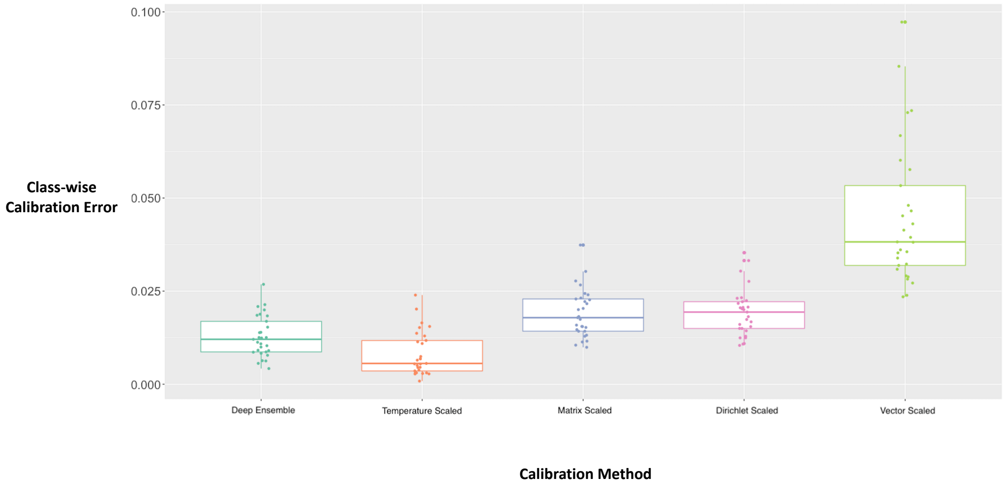


Figure 3.8: Boxplot displaying the classwise calibration error for the deep ensemble model trained on data from PCAWG. Classwise calibration error for the deep ensemble model and the deep ensemble model after post-hoc calibration with: Temperature scaling, matrix scaling, Dirichlet scaling and vector scaling. Points represent the expected calibration error for each of the 29 cancer types. The centre line in the boxplot is the median classwise expected calibration error. The lower and upper bounds of the box represent the first and third quartile. The whiskers extend to 1.5 IQR plus the third quartile or minus the first quantile.

those expected from a perfectly calibrated classifier (p=0.123, p-value = 1.0; permutation test). Overall, these results suggest that both variants of the uncalibrated deep ensemble model can provide highly well-calibrated predictions.

Interestingly, matrix scaling, Dirichlet scaling, and vector scaling increased both classwise-ECE and ECE. While these methods have favourable properties for improving classwise-ECE, they run the risk of overfitting to a smaller validation set. Together, these results suggest that more complex calibration maps may not be suitable for reducing the deep ensemble model's calibration error. A summary of classwise-ECE for both deep ensemble models can be found in Table 3.5 and Table 3.6.

Notably, the PCAWG-trained deep ensemble tended to produce high confidence predictions, with confidence values tending to be higher than 0.80 when making predictions. In contrast, since temperature scaling can increase the entropy of the softmax distribution, the confidence ranges for the temperature scaled models showed considerable variation. So, while this model provides confidence-calibrated predictions, the predictions tend not to be high-confident. These results suggest that the deep ensemble method produces both highly confident and highly reliable predictions (Figure 3.10).

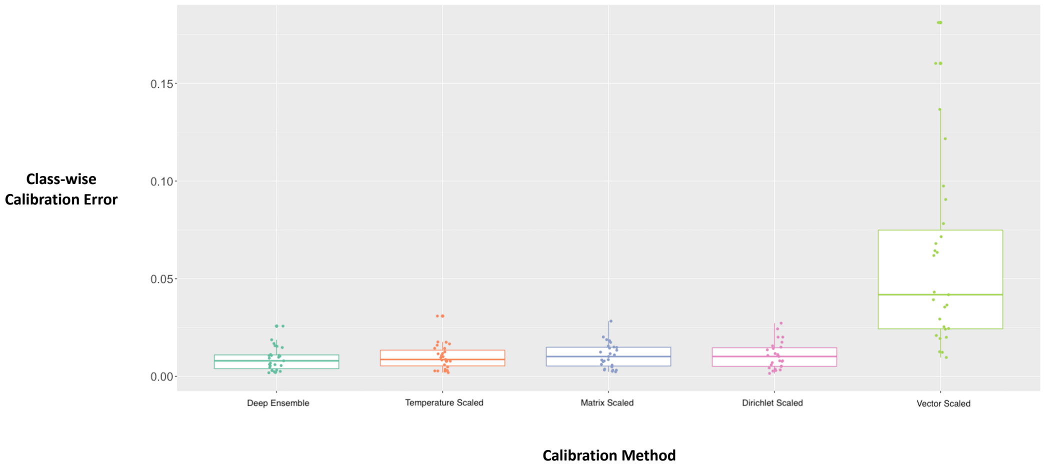


Figure 3.9: Boxplot displaying the classwise calibration error for the deep ensemble model trained on the complete dataset. Classwise calibration error for the deep ensemble model and the deep ensemble model after post-hoc calibration with: Temperature scaling, matrix scaling, Dirichlet scaling and vector scaling. Points represent the expected calibration error for each of the 27 cancer types. The centre line in the boxplot is the median classwise expected calibration error. The lower and upper bounds of the box represent the first and third quartile. The whiskers extend to 1.5 IQR plus the third quartile or minus the first quantile.

Table 3.6: Summary of classwise-ECE for the deep ensemble trained on PCAWG Summary of the classwise expected calibration error for the deep ensemble classifier trained on data from PCAWG. Results are averaged across 10 held-out test sets from PCAWG.

	Deep Ensemble	Temperature Scaled	Matrix Scaled	Dirichlet Scaled	Vector Scaled
Cancer					
Biliary-AdenoCA	0.021387	0.023951	0.021634	0.022465	0.032275
Bladder-TCC	0.007779	0.004484	0.013175	0.016762	0.035581
Bone-Leiomyo	0.016895	0.013684	0.020358	0.019385	0.031922
Bone-Osteosarc	0.012424	0.011404	0.017431	0.017391	0.028241
Breast-AdenoCA	0.020873	0.006472	0.030302	0.033213	0.066776
CNS-GBM	0.009135	0.005590	0.015249	0.010796	0.045224
CNS-Medullo	0.015332	0.006817	0.026667	0.027627	0.072960
CNS-Oligo	0.010350	0.006856	0.014248	0.014984	0.023481
CNS-PiloAstro	0.018814	0.012964	0.022900	0.023089	0.060157
Cervix-SCC	0.012068	0.010940	0.015885	0.015474	0.023888
ColoRect-AdenoCA	0.006332	0.002895	0.011575	0.012472	0.038230
Eso-AdenoCA	0.018523	0.015226	0.023150	0.021718	0.048038
Head-SCC	0.013863	0.011753	0.020041	0.020038	0.029111
Kidney-ChRCC	0.008617	0.003055	0.014238	0.014995	0.041391
Kidney-RCC	0.010861	0.003092	0.024381	0.022163	0.057656
Liver-HCC	0.018361	0.015522	0.027742	0.030391	0.097264
Lung-AdenoCA	0.008304	0.005453	0.011352	0.010930	0.027183
Lung-SCC	0.005606	0.003076	0.009921	0.012881	0.030882
Lymph-BNHL	0.008685	0.003809	0.015526	0.014354	0.046555
Lymph-CLL	0.012547	0.004556	0.022211	0.020236	0.053352
Myeloid-MPN	0.006262	0.002776	0.012888	0.012459	0.043077
Ovary-AdenoCA	0.009987	0.002789	0.018175	0.020639	0.039449
Panc-AdenoCA	0.026831	0.020192	0.037390	0.035331	0.085378
Panc-Endocrine	0.013948	0.007450	0.017884	0.020723	0.035281
Prost-AdenoCA	0.012547	0.005032	0.022671	0.020538	0.073491
Skin-Melanoma	0.004217	0.000853	0.010505	0.010436	0.038144
Stomach-AdenoCA	0.019961	0.016485	0.024027	0.023248	0.036106
Thy-AdenoCA	0.011279	0.005485	0.014669	0.016064	0.028843
Uterus-AdenoCA	0.009032	0.003581	0.015491	0.018148	0.033890

Table 3.7: Summary of classwise-ECE for the deep ensemble trained on the complete dataset Summary of the classwise expected calibration error for the deep ensemble classifier trained on a combination of data from PCAWG, HMF and an independent set of primary tumours.

	Deep Ensemble	Temperature Scaled	Matrix Scaled	Dirichlet Scaled	Vector Scaled
Cancer					
Biliary-AdenoCA	0.008406	0.011909	0.009317	0.009158	0.024688
Bladder-TCC	0.002620	0.004369	0.001487	0.002371	0.008213
Bone-Leiomyo	0.002421	0.003946	0.003297	0.004458	0.011860
Bone-Osteosarc	0.015622	0.014237	0.017237	0.017919	0.023735
Breast-AdenoCA	0.030853	0.050186	0.030439	0.029928	0.165787
CNS-GBM	0.007820	0.010542	0.012452	0.011394	0.067186
CNS-Medullo	0.005804	0.008678	0.006146	0.005841	0.321075
CNS-PiloAstro	0.010146	0.012419	0.015398	0.016397	0.024579
ColoRect-AdenoCA	0.009091	0.006793	0.006800	0.008004	0.071743

Continued on next page

	Deep Ensemble	Temperature Scaled	Matrix Scaled	Dirichlet Scaled	Vector Scaled
Cancer					
Eso-AdenoCA	0.017504	0.016151	0.015909	0.016907	0.030165
Head-SCC	0.013331	0.016028	0.013975	0.014140	0.054933
Kidney-ChRCC	0.003724	0.005966	0.004120	0.005114	0.016503
Kidney-RCC	0.003351	0.005369	0.004920	0.004403	0.030379
Liver-HCC	0.008463	0.012283	0.016161	0.016631	0.095551
Lung-AdenoCA	0.006702	0.009608	0.006517	0.005846	0.042343
Lung-SCC	0.003234	0.002943	0.004045	0.003599	0.063663
Lymph-BNHL	0.010924	0.011575	0.010860	0.012341	0.051836
Lymph-CLL	0.005292	0.007235	0.007136	0.008713	0.016747
Myeloid-MPN	0.002549	0.003462	0.003827	0.003866	0.018071
Ovary-AdenoCA	0.010970	0.020493	0.015510	0.015468	0.030691
Panc-AdenoCA	0.018397	0.026838	0.026576	0.023977	0.102180
Panc-Endocrine	0.009190	0.012510	0.012998	0.013203	0.038142
Prost-AdenoCA	0.015128	0.020950	0.021558	0.020083	0.082948
Skin-Melanoma	0.002078	0.003820	0.010856	0.010798	0.091594
Stomach-AdenoCA	0.015740	0.017547	0.014496	0.013527	0.015407
Thy-AdenoCA	0.007066	0.010001	0.008131	0.008086	0.020170
Uterus-AdenoCA	0.010155	0.009417	0.010539	0.011197	0.030965

3.3.8 Automatic detection of rare cancer samples

Another notion of uncertainty that is important when the predictions of a deep neural network may be used as the basis for guiding clinical decision-making is automatically detecting input data that are anomalous or significantly different from those used to train the model. It is possible to receive cancer samples from cancer types outside of the 29 cancer types the model is trained to identify in a clinical setting. The task of identifying these samples can be referred to as rare-cancer detection. To automatically determine if input samples come from rare cancer samples that the model isn't trained to classify, I made use of the deep ensemble's predictive entropy. Predictive entropy was calculated on all validation set samples independently for each data partition, and a partition-specific threshold value was set at the 95th percentile of predictive entropy (See ”Methods” for details).

Figure 3.11 shows a summary of the predictive entropy for all samples in PCAWG. Test-set samples from in-distribution cancer types (cancer types the model is trained to identify) tend to have predictive entropy values lower than the 95th percentile cutoff used for identifying rare cancer samples. In some cases, the median entropy of a cancer type is almost zero. This includes the following cancer types: CNS-GBM, ColoRect-AdenoCA, Lung-SCC, Lymph-BHNL, Skin-Melanoma. Interestingly, the F1 score for classification performance for this set of cancers tended to be very high, ranging from 0.85 (CNS-GBM) to 1.00 (Skin-Melanoma). This result is consistent with the confidence calibration results, which suggested that the deep ensemble model tended to make high-confidence predictions with comparably high accuracy.

The predictive entropy of most out-of-distribution (OOD) cancer types was similarly low. Of the OOD cancer types, Breast-LobularCA, Breast-DCIS, Lymph-NOS, Myeloid-AML and Myeloid-MDS tended to have entropy values below the threshold entropy. These cancer types tend to have similar biological characteristics to some in-distribution cancer types. Similarly, Lymph-NOS is similar to the other B-cell malignancies in the dataset (Lymph-CLL and Lymph-BHNL), and the two OOD myeloid lineage cancers are similar to Myeloid-MPN. OOD cancer types that are similar to in-distribution cancer types

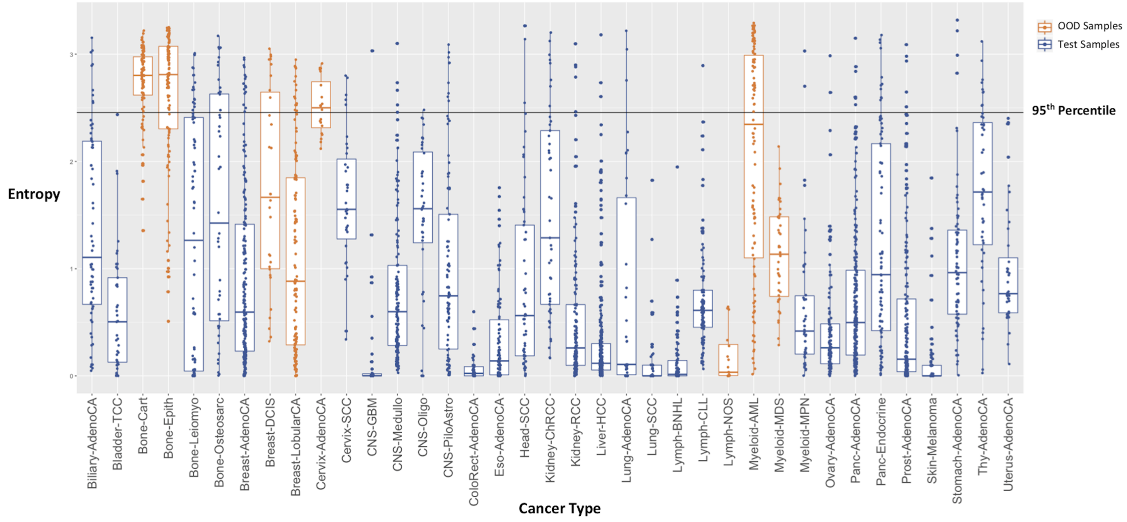


Figure 3.10: Boxplot displaying the predictive entropy of the deep ensemble model trained on PCAWG. Predictive entropy (nats) for each cancer sample from the held-out test sets used to evaluate the deep ensemble trained on PCAWG, and samples from cancer types the model did not see during training. Points represent the entropy of each samples in one of the 10 deep ensemble models. The horizontal black line represents the 95^{th} percentile of predictive entropy, calculated on all held-out validation samples from PCAWG. Samples labeled in blue belong to cancer types the model saw during training. Samples labeled in orange belong to cancer types the model did not see during training. The centre line in the boxplot is the median classwise expected calibration error. The lower and upper bounds of the box represent the first and third quartile. The whiskers extend to 1.5 IQR plus the third quartile or minus the first quantile.

having lower entropy is consistent with the classifier often making misclassifications due to biological relatedness.

Three cancer types in the dataset are relatively uncommon or rare cancers - cervical adenocarcinomas (Cervix-AdenoCA), chordoma (Bone-Epith) and chondroblastoma (Bone-Cart). All three of these cancer types have relatively high entropy compared to the dataset as a whole. Using the threshold value, I assigned cancer samples as either in-distribution (if they belonged to either the test set or were highly related to test set samples) or OOD (cervix-AdenoCA, Bone-Cart or Bone-Epith samples) and assessed classification performance. Overall, the accuracy for classification was 93%. Given that the in-distribution samples represented approximately 92% of the dataset, the class in-balance might have been driving OOD detection accuracy. To investigate this, I calculated the Matthew's Correlation Coefficient (MCC) for each deep ensemble partition. The mean MCC was 0.62, which suggests that entropy is strongly correlated with the correct label (in-distribution or OOD) (Table 3.7). This result suggests that the deep ensemble's predictive entropy can accurately identify cancer samples that differ significantly from the training data distribution. Table 3.8 contains a summary of the OOD performance for each data partition.

Table 3.8: Summary of out-of-distribution detection using the deep ensemble model. Summary of out-of-distribution (OOD) detection results using the deep ensemble model trained on PCAWG, evaluated on data from PCAWG, the independent validation set of primary tumours and metastatic samples from HMF. Results for PCAWG are averaged across 10 models. Proportion refers to percentage of samples that are OOD. MCC refers to the Matthews correlation coefficient.

	Accuracy (%)	Macro F1 score	MCC	Proportion (%)
Dataset				
PCAWG	93	0.80	0.62	8
Independent Primary Tumours	74	0.62	0.40	11
Hartwig Medical Foundation	90	0.69	0.38	10

Table 3.9: Summary of out-of-distribution detection using the deep ensemble model evaluated on PCAWG. Summary of out-of-distribution (OOD) detection results using the deep ensemble model trained on PCAWG, evaluated on held-out test data from PCAWG and OOD samples from PCAWG. Results are shown for each of the 10 models corresponding to each data partition. MCC refers to the Matthew's correlation coefficient.

	Accuracy (%)	MCC	Macro F1 Score
Model			
1	93	0.61	0.80
2	93	0.62	0.80
3	94	0.61	0.80
4	92	0.58	0.77
5	94	0.64	0.82
6	92	0.61	0.80
7	93	0.62	0.80
8	91	0.57	0.77
9	94	0.66	0.82
10	96	0.70	0.85

To determine if predictive entropy generalizes to the independent dataset of primary tumours, I calculated entropy on all samples from that dataset. Overall, the predictive entropy of many OOD cancers was higher than the threshold value (Figure 3.11). Interestingly, all oesophagal squamous cell carcinoma samples (ESCC) had entropy greater than the cut-off value, suggesting that these tumours are highly dissimilar to both Eso-AdenoCA and Stomach-AdenoCA. A set of pediatric brain cancers (PBCA) also had entropy values higher than the threshold, suggesting that it may contain brain cancers that differ significantly from those found in PCAWG. Interestingly, some in-distribution cancer samples had relatively high entropy. CNS-Medullo, CNS-PiloAstro, Myeloid-MPN and Panc-Endocrine had entropy values higher than the threshold despite being present in PCAWG. Using the thresholding procedure above, I assigned samples to in-distribution or OOD based on entropy. Overall, the accuracy was 74%, and the MCC for this task was 0.40 (Table 3.7). This dataset contained 72 ESCC, all of which were correctly identified as OOD samples, although oesophagal adenocarcinoma was included in the training data. This result is consistent with studies suggesting that ESCC is genomically more similar to other

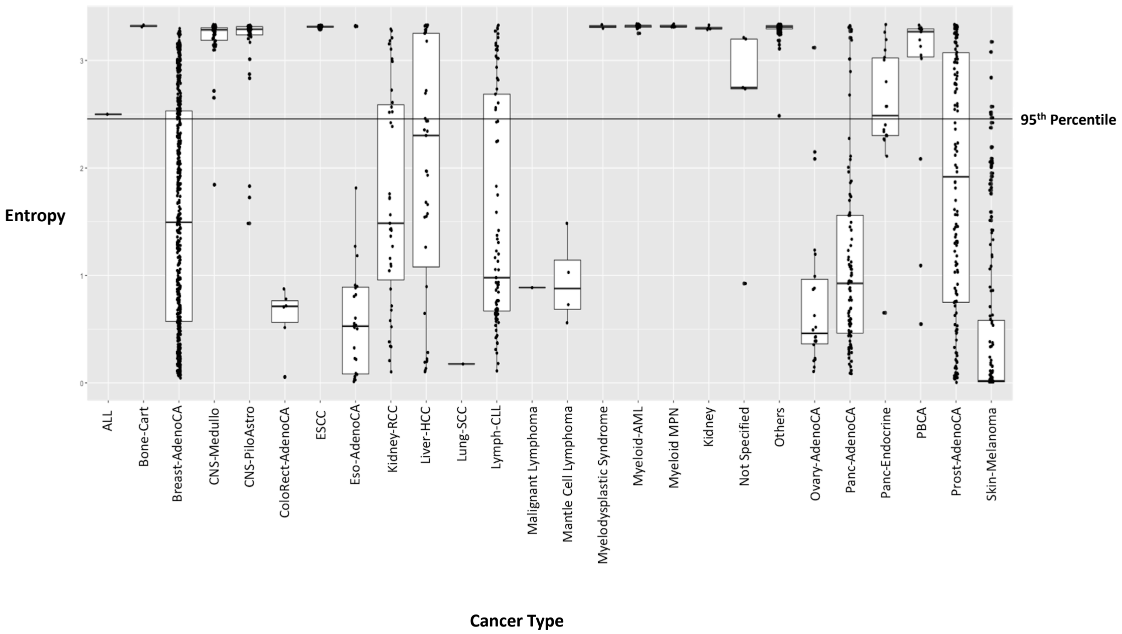


Figure 3.11: Boxplot displaying the predictive entropy of the deep ensemble model trained on PCAWG on an independent set of primary tumours. Predictive entropy (nats) for each cancer sample from an independent validation set of primary tumours. Entropy for each sample is averaged across the 10 deep ensemble classifiers trained on data from PCAWG. The horizontal black line represents the 95^{th} percentile of predictive entropy, calculated on all held-out validation samples from PCAWG. The centre line in the boxplot is the median classwise expected calibration error. The lower and upper bounds of the box represent the first and third quartile. The whiskers extend to 1.5 IQR plus the third quartile or minus the first quantile.

squamous cell carcinomas than with oesophagal adenocarcinomas (Kim et al. 2017).

To determine if the predictive entropy generalizes to the HMF samples, I calculated predictive entropy for all samples in the HMF dataset, including those not used for classification (Figure 3.12). In general, most cancer samples tended to have relatively low predictive entropy, but there were some outlier cases. About half of the Bone/Soft Tissue cancers had predictive entropy greater than the threshold value. As these cancer samples lack finer-grained categorization, some of these samples may be rare cancer samples, while a portion of them may be highly similar to other sarcomas in the PCAWG dataset. Of the highly dissimilar cancer types, most mesothelioma samples had entropy values above the threshold. All other OOD cancer samples tended to have a relatively low entropy, suggesting that predictive entropy may need to be re-calibrated for this dataset. Using the same thresholding procedure described above for identifying in-distribution vs OOD samples, the overall accuracy for detecting OOD samples was 89%. MCC for this classification task was 0.38, indicating a modest correlation between entropy and OOD status (Table 3.7). While this dataset's performance is sufficiently good, re-calibration of the predictive

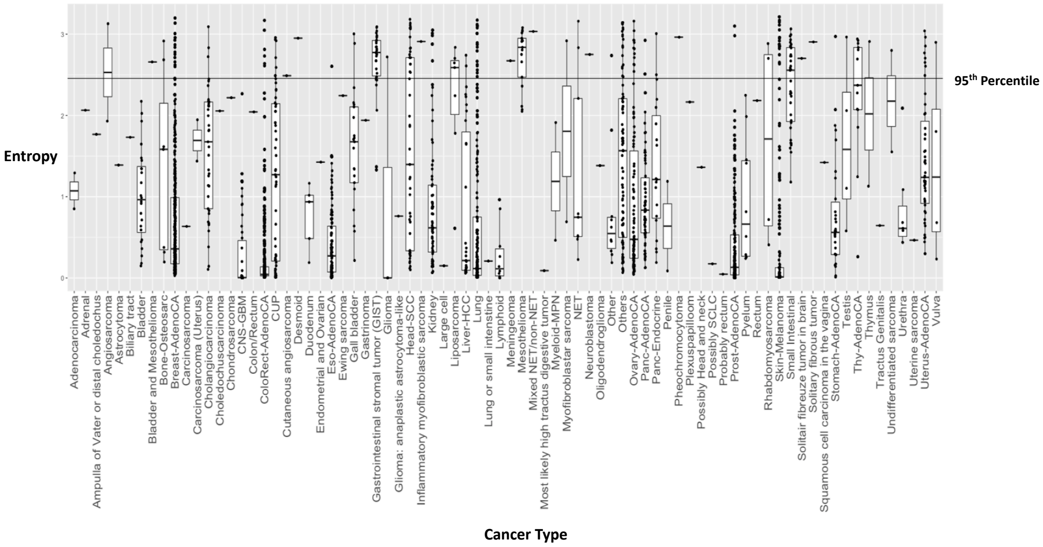


Figure 3.12: Boxplot displaying the predictive entropy of the deep ensemble model trained on PCAWG on an independent set metastatic tumours. Predictive entropy (nats) for each cancer sample from an independent set of metastatic tumours from the HMF dataset. Entropy for each sample is averaged across the 10 deep ensemble classifiers trained on data from PCAWG. The horizontal black line represents the 95^{th} percentile of predictive entropy, calculated on all held-out validation samples from PCAWG. The centre line in the boxplot is the median classwise expected calibration error. The lower and upper bounds of the box represent the first and third quartile. The whiskers extend to 1.5 IQR plus the third quartile or minus the first quantile.

entropy for this dataset may improve overall OOD detection performance.

Interestingly, nearly every CUPS sample had predictive entropy below the threshold value. Ground truth labels for CUPS in this dataset are unknown, so it is impossible to determine if these cancer samples come from in-distribution cancer types. Despite this fact, having a lower predictive entropy suggests that they have a reasonable likelihood of belonging to in-distribution cancer types. Overall, this result suggests that CUPS would pass the OOD detection criteria in a clinical setting and would receive a prediction from the classifier.

3.3.9 Entropy improves accuracy

In a clinical setting where the model's predictive entropy is used to rule out potential OOD samples, overall model performance is best judged on samples that pass the entropy threshold. To determine if lower entropy samples were easier to classify, I calculated entropy on all samples passing the threshold value and all higher entropy samples that would be ruled OOD (Figure 3.13). Overall, the accuracy on low entropy samples was 90% which is higher than the accuracy on all test samples. High entropy samples had a much lower overall accuracy, with an average classification accuracy of 62% across the ten data partitions, which is significantly lower than the accuracy on low entropy samples (p <1e-3, Wilcoxon Signed-Rank test). This result is consistent with the result that the deep ensemble is highly well-calibrated.

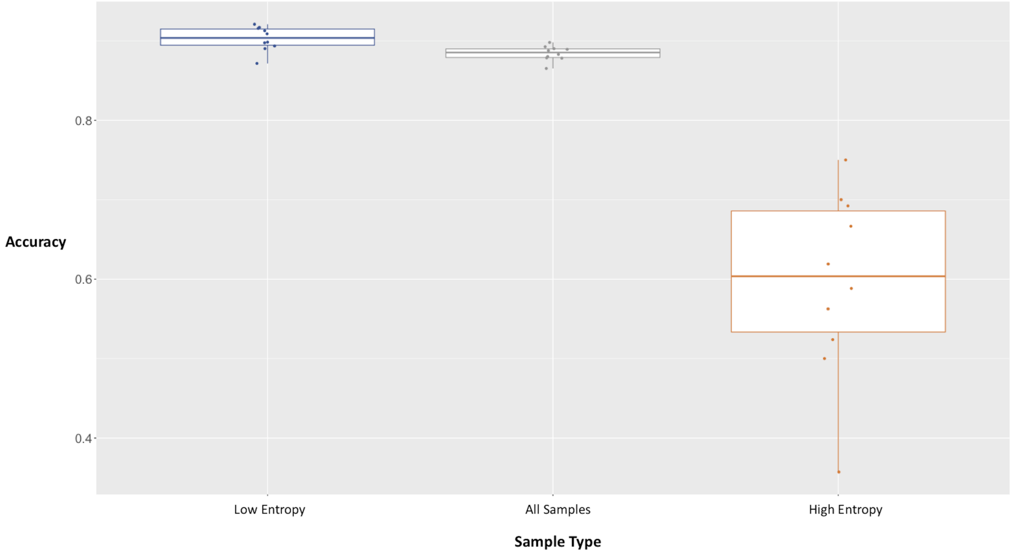


Figure 3.13: Boxplot displaying the accuracy of held-out data from PCAWG based on entropy. Accuracy of held-out samples from PCAWG that were evaluated with the deep ensemble method. Samples were assigned to the low-entropy group if predictive entropy was lower than the 95^{th} percentile of entropy on held-out validation data from PCAWG, and high entropy otherwise. Points represent performance on each of the 10 held-out test sets from PCAWG. The centre line in the boxplot is the median classwise expected calibration error. The lower and upper bounds of the box represent the first and third quartile. The whiskers extend to 1.5 IQR plus the third quartile or minus the first quantile.

To assess the accuracy of low entropy samples from the independent validation sets, I repeated the same procedure described above for the dataset of independent primary tumours and the dataset of metastatic samples. The independent dataset of primary tumour samples contained 1064 low entropy tumour samples (1533 samples total in this dataset). Accuracy on the low entropy samples was higher than that seen on the entire independent validation set, and improved from 84.4% for all samples to 90.4% for low entropy samples (Figure 3.14). The dataset of metastatic samples contained 1945 low entropy samples (2267 samples in total for this dataset). Similar to the dataset of primary tumours, low entropy samples had a higher overall accuracy. Accuracy for low entropy samples from the dataset of metastatic samples improved from 83.5% to 86% for low entropy samples (Figure 3.15).

3.4 Discussion

Cancers of unknown primary site (CUPS) occur when a patient presents with identifiable metastatic cancer, but despite examination, the primary tumour site cannot be determined. This syndrome is the fourth most common cause of cancer related mortality (Greco 2013; Pavlidis et al. 2003). As current

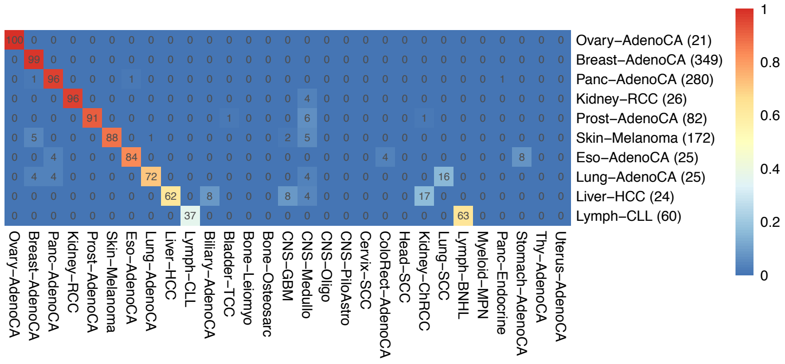


Figure 3.14: Heatmaps displaying the accuracy of low entropy samples on the two independent validation sets. Accuracy of low entropy samples from the two independent validation sets were evaluated with the deep ensemble method. Samples were assigned to the low-entropy group if predictive entropy was lower than the 95^{th} percentile of entropy on held-out validation data from PCAWG. Heatmap summarizing the performance of low entropy samples from the independent dataset of primary tumours.Each row corresponds to the true tumor type; Columns correspond to the predictions emitted by each of the classifiers. Cells are labeled with the proportion of tumors of a particular type that were called a specific type by the classifier. The heatmap represents the recall (sensitivity) of the deep ensemble. Due to rounding of values, some rows add up to slightly more or less than 100%.

therapeutic approaches are heavily guided by a tumour's cell of origin, identifying the primary tumour site for CUPS is an important clinical task. Outside of the clinical dilemma presented by CUPS patients,

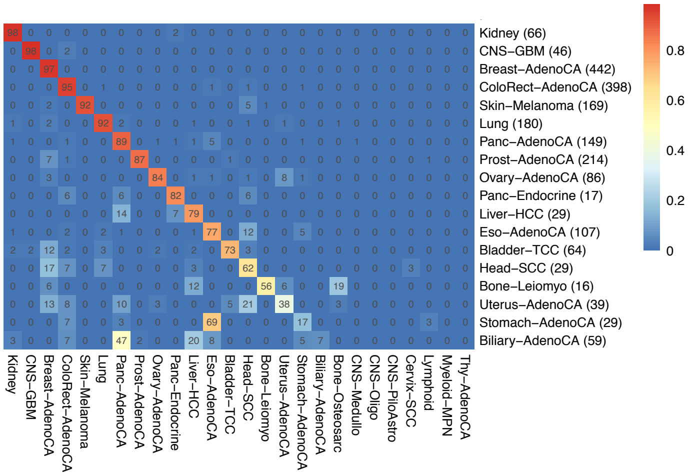


Figure 3.15: Heatmaps displaying the accuracy of low entropy samples on the two independent validation sets. Accuracy of low entropy samples from the two independent validation sets were evaluated with the deep ensemble method. Samples were assigned to the low-entropy group if predictive entropy was lower than the 95^{th} percentile of entropy on held-out validation data from PCAWG. Heatmap summarizing the performance of low entropy samples from the dataset of metastases. Each row corresponds to the true tumor type; Columns correspond to the predictions emitted by each of the classifiers. Cells are labeled with the proportion of tumors of a particular type that were called a specific type by the classifier. The heatmap represents the recall (sensitivity) of the deep ensemble. Due to rounding of values, some rows add up to slightly more or less than 100%.

the inability to correctly identify the primary tumour site severely limits researchers' ability to investigate the biological characteristics of this disease. Correctly identifying the primary tumour site can allow for the evolutionary mechanisms and biological processes involved in the development of CUPS to be correctly characterized. A similarly difficult diagnostic challenge arises when patients present with a history of multiple primary tumours. For these patients, correctly discriminating a late metastatic recurrence from a new primary tumour and correctly identifying which primary tumour was responsible for a metastatic lesion is critical for making clinical decisions (Vogt et al. 2017). In Chapter 2, I aimed to address the need for automatic tumour typing by developing a deep learning system that could accurately discriminate between 24 common cancer types (Jiao et al. 2020). While this model showed impressive performance for identifying primary tumour site, multiple challenges existed for translating this system into a clinical setting. First, the original classifier could only identify 24 cancer types. Second, the original classifier does not provide estimates of its predictive uncertainty, which are important when clinical decisions may be guided based on a neural network's predictions.

In this chapter, I used a large collection of both primary and metastatic cancer whole genomes to develop a series of deep learning models that can accurately distinguish between 29 cancer types whilst providing well-calibrated estimates of in-distribution and out-of-distribution uncertainty. The system had an overall accuracy of 88%, with 22 of 29 cancer types achieving an F1 score of at least 0.80. When applied to external validation data sets, the classifier achieved predictive accuracies of 84% and 84%, respectively, for primary and metastatic tumours. The reduction in accuracy on metastatic tumours is driven by multiple features of that data and based on the inclusion of Biliary-AdenoCA, which is often misclassified as Liver-HCC. Including data from sources outside of PCAWG tended not to improve performance. Overall, when a classifier was trained using a complete dataset including PCAWG, independently collected primary tumours and metastases, the overall accuracy of the model was 89% for distinguishing 27 cancer types, but with the same F1 score of 0.83 as the model trained only on data from PCAWG. The model trained on this complete dataset also had different patterns of misclassification. A possible explanation for this may be that many of these samples were sequenced following exposure to chemotherapy. If a chemotherapeutic agent causes specific mutation types in a tumour genome, two unrelated cancer types treated with the same agent may have very similar mutation types, leading to misclassification. As many of the treated samples were exposed to platinum-based therapy and fluorouracil, chemotherapy-associated mutations may be contributed to misclassification. While exposure to chemotherapy provides a possible explanation for misclassifications, this result is still a limitation of the current model. This limitation may be addressed by systematically removing mutations associated with chemotherapy, but doing so requires both a complete understanding of the mutational consequences of cancer therapy and a suitable statistical model for removing the contribution of chemotherapy-associated mutations.

When the predictions of a neural network may guide clinical decision making, properly calibrated uncertainty estimates are essential. In this setting, for example, a highly confident prediction may form the basis for cancer-specific therapy. In contrast, a high uncertainty prediction may signal that broad-spectrum chemotherapy is more advisable. The deep ensemble utilized for classification, combined with adversarial training, resulted in a confidence-calibrated classifier capable of providing good estimates of predictive uncertainty. Temperature scaling tended to reduce the classwise expected calibration error but fails to reduce the overall calibration error for the model trained on data from PCAWG. In this case, the uncalibrated deep ensemble model trained solely on data from PCAWG had an expected calibration error that was consistent with that seen from a perfectly calibrated classifier. More complex calibration maps such as matrix scaling and Dirichlet scaling tended to increase calibration error, possibly due to overfitting on a relatively small validation set. Overall, this result suggests that the deep ensemble model can provide well-calibrated estimates of in-distribution uncertainty. A second use of predictive uncertainty is to automatically identify rare cancer samples that the model wasn't trained to classify. Using the model's predictive uncertainty, rare cancer samples could automatically be detected with an overall accuracy of 93% and a Matthew's correlation coefficient of 0.62 for data from PCAWG. When extending this approach to the independent set of primary tumours, predictive entropy continued to perform well with an overall accuracy of 74% and a Matthew's correlation coefficient of 0.40. Interestingly, this approach perfectly identified squamous-cell carcinoma of the oesophagus as an out-of-distribution cancer type, which is consistent with the two histological variants of oesophagal cancer differing significantly at the genomic level (Kim et al. 2017). Overall performance on the HMF samples was lower than that seen on data from PCAWG and independent primary cancers but was still reasonably good. A potential method for improving this dataset's performance would be to remove chemotherapy-associated mutations that may constitute a relatively large proportion of all mutations in samples sequenced after treatment. Interestingly, when the predictive entropy of CUPS samples was calculated, they tended to have a relatively low entropy, suggesting that the model can make high-confidence predictions for these cancer samples.

In summary, this study further demonstrates the potential of whole-genome sequencing to distinguish major cancer types based on patterns of somatic mutations. The work presented here addresses significant challenges for translating tumour-typing algorithms into clinical settings and is one of the few applications of deep learning in genomics that assesses model calibration and uncertainty quantification. In the future, it would be desirable to make algorithmic improvements that allow tumour types to be subdivided into molecular subtypes, incorporate additional feature types, and extend the model to work with data generated from liquid biopsies.

3.5 Materials and methods

3.5.1 PCAWG training and test data

All variant calls data were downloaded from the ICGC Portal, and all file names given here are relative to this path. Note that controlled tier access credentials are required from the ICGC and TCGA projects. The consensus Somatic SNV file covers 2778 whitelisted samples from 2583 donors. Tumour histological classifications were reviewed and assigned by the PCAWG Pathology and Clinical Correlates Working Group. All samples flagged as exhibiting microsatellite instability (MS) by the PCAWG Technical Working Group were removed for model training. In a small number of cases, the same donor contributed both primary and metastatic specimens to the PCAWG data set. In these cases, I only used the primary tumour for training and evaluation, except for the case of a small cohort of myeloproliferative neoplasms (Myeloid-MPN; N = 55 samples), for which multiple primary samples were available. In this case, I used up to two samples per donor and partitioned the training and testing sets to avoid having the same donor appear more than once in any training/test set trial. A complete characterization of data from PCAWG is provided in Jiao et al., 2020 (Jiao et al. 2020).

3.5.2 Independent validation data set: primary and metastatic tumours

To independently validate the neural network classifier, I assembled several sets of tumours that had been subject to whole-genome sequencing outside of PCAWG.

The primary tumour validation data set consisted of 1333 primary tumours contributed by colleagues participating in the PCAWG Mutational Signatures Working Group (Alexandrov et al. 2020). These represent 14 tumour types overlapping with PCAWG types. These independent primaries were supplemented using WGS from 200 advanced primary pancreatic ductal adenocarcinomas (Panc-AdenoCA) derived from the COMPASS Trial (Aung 2018). In all, the primary tumour validation set contained 1533 primary tumour samples across 14 tumour types. Only tumour types with 10 or more representatives were used for testing. Only tumour types with ten or more representatives were tested.

The metastatic tumour validation data was derived from SNV calls on 2175 metastatic tumours across 18 tumour types, provided by the Hartwig Medical Foundation. These data were supplemented with 92 metastatic pancreatic ductal adenocarcinomas to the liver from the COMPASS Trial for a total of 2267 metastatic tumours. Only tumour types with ten or more representatives were tested.

Although the sequencing technologies and genome coverage are comparable among the PCAWG training set and the independent validation data sets, a mixture of different human genome builds, alignment algorithms and SNV calling algorithms were used for the validation data sets. We did not attempt to recall the SNVs but did lift the genome coordinates of the samples that had been aligned to the other genome builds to hg19 using CrossMap (Version 0.2.5). Complete descriptions of the validation set of primary tumours is provided in Table 2.5, and a complete description of the metastatic tumours can be found in Aung (Aung 2018) and Preistley (Priestley et al. 2019).

3.5.3 Somatic mutation feature sets

Mutational-type features are based on all SNVs. For each sample, SNVs are categorized across the possible single-nucleotide changes (A >C, A >G, A >T, C >A, C >G, C >T), the 48 possible nucleotide changes plus their 5′ or 3′ flanking base and the 96 possible nucleotide changes plus both flanking nucleotides. This generates 150 mutational-type features in total. The counts in each category are either represented as is or represented as normalized Z-scores. Z-scores were used for training models that made use of data only from PCAWG. The model trained on the complete dataset used the raw counts of the mutational-type features. The mutational distribution feature consists of the number of somatic SNVs in each 1-megabase bin across the genome. For the construction of mutational distribution features, sex chromosomes are excluded.

3.5.4 Deep learning procedure

In all cases, the deep neural networks were trained on the mutational type and mutational distribution features described above. The neural networks were trained to classify either 27 or 29 cancer types. The details for each data set are as follows:

Training of Deep Ensemble on PCAWG

For classification, I used a fully-connected feed-forward neural network. Prior to training, data from PCAWG were split into training, validation and test sets ten times to create ten different partitions over the full data set. Data from any cancer type with at least 15 donors in PCAWG was included for a total of 2566 samples across 29 cancer types. In order to balance the number of samples per cancer type during training, the cancer samples in the training set were oversampled using SMOTE with default settings (Chawla et al. 2002). After oversampling, each cancer type had the same number of examples in the training set. Hyperparameter optimization (described in detail below) was done independently for each data partition. Models were trained with a minibatch size of 32 for 100 epochs using Adam.

After hyperparameter optimization was completed, a deep ensemble was created for each of the 10 data partitions. The oversampled training set was used for training the networks. The deep ensemble was created by initializing 50 neural networks with the same hyperparameters but with random weight initialization (Lakshminarayanan, Pritzel, and Blundell 2017). Each neural network was trained for 100 epochs using a minibatch size of 32 using Adam (Kingma and Ba 2014). As with hyperparameter optimization, adversarial data were included during training using the fast gradient sign method (Goodfellow, Shlens, and Szegedy 2015). After training was completed for all ten data partitions, deep ensembles were created in two ways. First, a deep ensemble was created for each data partition by averaging the logit vectors produced by each neural network trained on the respective data partition. These networks were used for partition-specific performance evaluation. Another deep ensemble was created by creating a deep ensemble where each member of the ensemble consisted of a partition-specific deep ensemble. This network was used for evaluation on non-PCAWG datasets. In both cases, the class probability vector was derived by applying the softmax function to the average of the logit outputs from each neural network in the deep ensemble. The predicted tumour type in these cases is selected to be the greatest softmax probability.

Training of deterministic uncertainty estimation classifier on PCAWG

A fully connected feed-forward neural network was used as the feature extractor network. During training, the neural network learns a feature embedding of each input sample and learns class-specific centroids. Classification is done by first computing the feature embedding of an input sample and then computing the kernel distance between the feature embedding and all class-specific centroids (Amersfoort et al. 2020). An input sample is then assigned to the class with the closest centroid. Kernel distance was computed using a Radial Basis Function (RBF):

$$RBF(f_\theta(x), e_k) = \exp\left[-\frac{\frac{1}{n}\left\|W_k f_t heta(x) - e_c\right\|_2^2}{2\sigma^2}\right] \tag{3.1}$$

Data from PCAWG were split into training, validation, and test sets ten times to create ten different partitions over the full data set. Data partitions were the same as those used for training the deep ensemble model described above. Data from any cancer type with at least 15 donors in PCAWG was included for a total of 2566 samples across 29 cancer types. In order to balance the number of samples per cancer type during training, the cancer samples in the training set were oversampled using SMOTE with default settings. After oversampling, each cancer type had the same number of examples in the training set. Hyperparameter optimization (described in detail below) was done independently for each data partition. Models were trained with a minibatch size of 32 for 100 epochs using Adam. The learning rate was varied using a learning rate schedule to remain consistent with the original paper describing this model. After hyperparameter optimization, a single network was trained for each partition using the associated hyperparameters. The oversampled training set was used for training the networks. As with hyperparameter optimization, adversarial data were included during training using the fast-gradient sign method. Training was done for 100 epochs using a minibatch size of 32 with Adam.

Training of Deep Ensemble on the complete data set

A fully connected, feed-forward neural network was used for classification. A complete dataset was created by incorporating data from PCAWG and both independent validation sets. Data from a total of 27 cancer types were used for classification. Overall, this resulted in 6262 samples. Prior to training, data were split into training, validation or test samples. To balance cancer samples in the training set, the training set was oversampled using SMOTE with default settings. Hyperparameter optimization (described in detail below) was done to select hyperparameters for the model. Models were trained with a minibatch size of 32 for 150 epochs.

After hyperparameter optimization, a deep ensemble was created for this dataset. As with above, the deep ensemble was created by initializing 50 neural networks with the same hyperparameters but random weight initialization. Each ensemble was trained using the oversampled training set. Training was done using a minibatch size of 32 and for 150 epochs using Adam. As with hyperparameter optimization, adversarial data were included during training using the fast-gradient sign method.

Hyperparameter Optimization

In all cases, I used a Bayesian optimization approach to select hyperparameters (Snoek, Larochelle, and Adams 2012). Bayesian optimization was done for each dataset (described above) and for each classification approach. In each case, hyperparameters were selected by optimizing accuracy on the associated validation data for that data partition. I used the ’$gp_minimize$’ function from the scikit-optimise 0.5.2 Python library to select hyperparameters. In the case of deep ensemble-based classifiers, the following hyperparameters were optimized: learning rate for Adam, L2-regularisation penalty, dropout rate, the number of hidden layers, and the number of neurons per hidden layer. For deterministic uncertainty quantification models, two additional hyperparameters: gradient penalty and length scale, were also optimized. For deterministic uncertainty quantification models, the learning rate was not optimized, as a learning rate schedule was used, as per the original paper describing this model. Each model was trained using Adam with a batch size of 32. All hyperparameters of Adam other than the learning rate were set to the default values specified in the original paper. Weights and bias values were initialized with a Kaiming uniform distribution (He et al. 2015b). During training, adversarial data is included in training by first generating adversarial examples using the fast gradient sign method, with ϵ set to 1, and then augmenting the minibatch to include the adversarial examples. The model was evaluated with 400 hyperparameter combinations (i.e., 400 calls to ’*gp_minimize*’ were made). Briefly, ’*gp_minimize*’ approximates a function of model performance based on the hyperparameters with a Gaussian Process. For each call to ’*gp_minimize*’, the performance on the current set of hyperparameters is evaluated by training the neural network and assessing accuracy on a validation set. Based on this accuracy, the Gaussian Process posterior distribution is updated, and a new set of hyperparameters is chosen by optimizing an acquisition function. I used the ’*gp_hedge*’ acquisition function (Brochu, Hoffman, and De Freitas 2011). A complete description of hyperparameter settings found by Bayesian optimization is provided in Table 3.9 and Table 3.10.

Table 3.10: Hyperparameter settings found by Bayesian optimization for deep ensemble models. Description of the optimal hyperparameters found using Bayesian optimization for the deep ensemble classifiers. Dataset refers to the dataset used for training the model. L2 refers to the L2-penalization value. Dropout refers to the dropout rate. Layers refers to the number of hidden layers. Units refers to the number of units or neurons per hidden layer. Model refers to the data partition the model was trained on.

	Learning Rate	L2	Dropout	Layers	Units	Model	Accuracy(%)
Dataset							
PCAWG	0.000100	0.007058	0.007244	1	1024	1	88
PCAWG	0.000100	0.004336	0.500000	2	540	2	88
PCAWG	0.000100	0.005592	0.000001	1	650	3	87
PCAWG	0.000100	0.001000	0.500000	2	792	4	92
PCAWG	0.000226	0.001000	0.500000	1	335	5	88
PCAWG	0.000100	0.001282	0.000035	1	507	6	89
PCAWG	0.000100	0.001000	0.000092	2	1024	7	89
PCAWG	0.000241	0.001000	0.500000	1	280	8	89
PCAWG	0.000100	0.005271	0.000001	1	715	9	90
PCAWG	0.000100	0.007937	0.500000	2	328	10	90
Complete dataset	0.000100	0.001425	0.500000	2	379	1	89

Table 3.11: Hyperparameter settings found by Bayesian optimization for deterministic uncertainty quantification classifiers. Description of the optimal hyperparameters found using Bayesian optimization for the deterministic uncertainty quantification classifiers trained on data from PCAWG. Dataset refers to the dataset used for training the model. Penalty refers to the gradient penalty used during training. Scale refers to the length parameter of a Radial Basis Function. L2 refers to the L2-penalization value. Layers refers to the number of hidden layers. Units refers to the number of units or neurons per hidden layer of the feature extractor network. Embedding refers to the number of units in the embedding layer of the feature extractor. Model refers to the data partition the model was trained on.

	Penalty	Scale	Embedding	Layers	Units	Model	Accuracy (%)
Dataset							
PCAWG	0.000100	0.029153	6	5	1024	1	83
PCAWG	0.001801	0.055506	6	4	1024	2	83
PCAWG	0.000788	0.027730	6	3	1024	3	83
PCAWG	0.038864	0.023241	6	3	990	4	80
PCAWG	0.000100	0.012137	6	4	892	5	80
PCAWG	0.002971	0.020739	6	3	845	6	85
PCAWG	0.000100	0.030852	6	5	1024	7	84
PCAWG	0.003690	0.002336	33	5	1024	8	78
PCAWG	0.000100	0.023252	6	4	846	9	82
PCAWG	0.000100	0.020533	6	3	665	10	82

In all cases, models were implemented and trained with PyTorch 1.4.0 (Paszke et al. 2019). All code was written in Python 3.7.3 (Van Rossum and Drake 2009).

Definitions of accuracy metrics

To measure the performance of the classifiers, I use the conventional definitions of recall, precision, F1 score and accuracy. In the descriptions below, I use the abbreviations TP (true positive), TN (true negative), FP (false positive) and FN (false negative) to describe correct and incorrect assignments of an unknown tumour to a predicted type:

Recall: The proportion of samples of a particular histopathological type that are correctly assigned to that type:

$$Recall = TP/(TP + FN) \tag{3.2}$$

Precision: The proportion of samples assigned to a particular type that are truly that type:

$$Precision = TP/(TP + FP) \tag{3.3}$$

F1 score: The harmonic mean of recall and precision:

$$F1 = 2(recall * precision)/(recall + precision) \tag{3.4}$$

Accuracy: The proportion of correct assignments:

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \tag{3.5}$$

3.5.5 Model calibration

A probabilistic classifier such as a deep neural network is well calibrated if the predicted class distribution is approximately equal to the true class distribution (Kull et al. 2019). Typically, deep neural networks produce poorly calibrated output probabilities, yielding overly confident predictions.

The strictest notion of calibration is multiclass calibration. A multiclass-calibrated classifier is perfectly calibrated for every single class that the model is trained to classify (Kull et al. 2019). Consider a neural network classifier $\hat{f} : X \rightarrow \Delta_k$ that outputs probabilities for k classes. For any input $x \in X$, the classifier outputs a class probability vector $\hat{f}(x) = (\hat{f}_1(x), \hat{f}_2(x), ..., \hat{f}_k(x))$ belonging to $\Delta_k = \{(q_1, q_2, ..., q_k) \in [0, 1]^k | \sum_{i=1}^{k} q_i = 1\}$ which is the $(k-1)$-dimensional probability simplex over k classes.

Definition 4 *A probabilistic classifier $\hat{f} : X \rightarrow \Delta_k$ is multiclass-calibrated if for any prediction vector $q = (q_1, q_2, ..., q_k) \in \Delta_k$, the proportions of classes among all possible $x \in X$ getting the same predictions $\hat{f}(x) = q$ are equal to the prediction vector q:*

$$P(Y = i|\hat{p}(x) = q) = q_i \; for \; i = 1, ...k. \tag{3.6}$$

A necessary condition for obtaining a multiclass-calibrated classifier is for the classifier to be calibrated for all individual classes (Kull et al. 2019). That is, for any given class, the classifier is perfectly calibrated. Formally, a classwise-calibrated classifier is as follows:

Definition 5 *A probabilistic classifier $\hat{f} : X \to \Delta_k$ is classwise-calibrated if for any class i and any predicted probability q_i:*

$$P(Y = i|\hat{f}(x) = q_i) = q_i \tag{3.7}$$

The notion of calibration that is typically of concern is confidence-calibration. When neural networks make predictions, an input x is assigned to the class with the largest element in the output class probability vector. This value is referred to as the model's confidence. A classifier is confidence-calibrated if, for all instances where the confidence is predicted to be c, the expected accuracy of the classifier is c (Kull et al. 2019). Formally, a confidence-calibrated classifier is defined as follows:

Definition 6 *A probabilistic classifier $\hat{f} : X \to \Delta_k$ is confidence-calibrated, if for any $c \in [0, 1]$:*

$$P(Y = argmax(\hat{f}(x))|max(\hat{f}(x)) = c) = c \tag{3.8}$$

Post-hoc calibration

Given an already trained classifier, a number of post-hoc calibration methods exist. These methods are essentially post-processing steps that can produce better calibrated probabilities. These methods have a varying number of parameters or hyperparameters which are tuned by minimizing a negative log-likelihood on a held-out validation set. Four commonly used post-hoc calibration methods are temperature scaling, matrix scaling, vector scaling and Dirichlet scaling.

Temperature scaling is one of the simplest post-hoc calibration method. Recall that the output of a probabilistic neural network is a softmax function. Given the model's logit vector z_x for an input sample x, the confidence prediction is as follows:

$$c = max(\sigma_{sm}(z_x)) \tag{3.9}$$

Where σ_{sm} is the $softmax$ function.

In temperature scaling, instead of working directly with z_x, the logit vector is scaled by a single tempering parameter, $T > 0$ for each class. Consequently, the scaled prediction is given by:

$$c = max(\sigma_{sm}(z_x/T)) \tag{3.10}$$

When the tempering parameter greater than zero, T will raise the class probability vector's entropy. As $T \to 0$, the confidence value will go up. Since each element of the logit vector is scaled by the same positive number T, the relative ordering of class probabilities on the softmax output does not change.

Similar to temperature Scaling, matrix and vector scaling work by performing a transformation on the logit vector. Matrix scaling works by learning a linear transformation on the logits such that the scaled confidence value is:

$$c = max(\sigma_{sm}(Wz_x + b)) \tag{3.11}$$

This approach can be viewed as learning a multiclass logistic regression model using the model logit vector as the input features. Vector scaling is a specific case of Matrix scaling where W is constrained to a diagonal matrix.

Dirichlet scaling is similar to matrix scaling in that it learns a multiclass logistic regression model on some output from the original classifier. The key difference is that Dirichlet scaling uses the class probability vector as a feature. In its linear parameterization, Dirichlet scaling is as follows:

$$c = max(\sigma_{sm}(W \ln{(q)} + b)) \tag{3.12}$$

Where q represents the class probability vector of the original neural network.

Post-hoc calibration methods were trained and applied to the Deep Ensemble trained on PCAWG data, and the Deep Ensemble trained on the complete dataset. When applied to PCAWG, post-hoc calibration was trained and applied independently for each of the 10 data partitions. In all cases, the post-hoc calibration method was trained on the held-out validation set associated with each dataset. To match the class distribution used during training, samples in the validation set were oversampled so that the class abundances matched those found during training. For temperature scaling, the model was trained using LFBGS for 20 epochs. Vector scaling was trained using Adam for 20 epochs. Both Dirichlet scaling and matrix scaling were optimized using LFBGS, and internal cross-validation was used to select an L2-norm penalty. After training, all post-hoc calibrated models and the original, uncalibrated model were evaluated on the associated held-out test set. Evaluation was done using the following metrics:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)| \tag{3.13}$$

Where n is the number of samples, B_m is the set of indices of samples whose predictive confidence falls into bin m, $acc(B_m)$ is the accuracy in bin m and $conf(B_m)$ is the confidence for samples in bin m.

$$ECE_{classwise} = \frac{1}{k} \sum_{j=1}^{k} \sum_{m=1}^{M} \frac{|B_{m,j}|}{n} |y_j(B_{m,j}) - \hat{p}_j(B_{m,j})| \tag{3.14}$$

Where k, m, n are the numbers of classes, bins and instances, respectively, $B_{m,j}$ refers to bin m for instances of class j, $\hat{p}_j(B_{m,j})$ is the average probability of class j, and $y_j(B_{m,j})$ is the true proportion of class j in bin $B_{m,j}$. For both ECE and $ECE_{classwise}$, 50 bins were used for calculating error.

In addition to the above calibration metrics, a statistical test proposed in (Vaicenavicius et al. 2019) was used to determine if a classifier deviated significantly from a perfectly confidence-calibrated classifier. Reliability diagrams were constructed as described in (Guo et al. 2017). For models trained on PCAWG data only, test samples were pooled across the 10 deep ensemble models prior to constructing the reliability diagram.

3.5.6 Out-of Distribution Detection

Samples were predicted to be out-of-distribution (OOD) based on either their predictive entropy from the deep ensemble or the kernel distance using DUQ. Predictive entropy from the deep ensemble was calculated as follows:

$$H(p(y|X)) = -\sum_{m=1}^{M} p_{\theta m}(y|x, \theta_m) \log p_{\theta m}(y|x, \theta_m) \tag{3.15}$$

For neural networks trained only on PCAWG, OOD detection performance was assessed independently using in-distribution samples from the test set associated with each data partition. For the deep ensemble trained on the complete dataset, the same OOD samples described above were used for evaluation, and in-distribution samples were taken as the associated test samples for that dataset.

Additional validation was done using OOD samples from the two independent validation sets. Central nervous system malignancies were removed for evaluation as there was some discrepancy in their true origin (Jiao et al. 2020). This comprised 1213 in-distribution and 143 OOD samples of primary tumours. All cancer samples in the HMF dataset, with the exception of 62 cancers of unknown primary, were used for evaluation. Only those cancer samples matching cancer types from the 29 cancer types the model was trained to identify were included. This totalled 2179 in-distribution and 231 OOD samples of metastatic tumours from the HMF dataset.

Performance was evaluated using the following metrics: Matthew's Correlation Coefficient (MCC), accuracy score and F1 score. In the descriptions below, I use the abbreviations TP (true positive), TN (true negative), FP (false positive) and FN (false negative) to describe correct and incorrect assignments OOD or in-distribution tumours:

Recall: The proportion of samples of a particular category (OOD or in-distribution) that are identified as the correct category:

$$Recall = TP/(TP + FN) \tag{3.16}$$

Precision: The proportion of samples assigned to a particular category that are truly from that category:

$$Precision = TP/(TP + FP) \tag{3.17}$$

F1 Score: The harmonic mean of recall and precision:

$$F1 = 2(recall * precision)/(recall + precision) \tag{3.18}$$

Accuracy: The proportion of correct assignments.

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \tag{3.19}$$

Matthews Correlation Coefficient: The Pearson correlation coefficient between the predicted class and the true class:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.20}$$

Chapter 4

Discussion and Summary

4.1 Summary

Cancers are most commonly categorized by their cell-of-origin, which is shaped by both the organ of origin and histology of the tumour. While advances in precision medicine have allowed for tumours to be treated based on molecular alterations specific to the tumour, identifying a tumour's cell of origin is still a critical task in a clinical setting. While cell-of-origin is often easy to identify for most tumours, challenging situations arise when a patient presents with multiple primary tumours or CUPS. For these patients, traditional diagnostic approaches are unable to identify cancer type (Greco 2013). As identifying primary tumour site is essential for these cases, many approaches to using a tumour's genomic characteristics have been proposed. In this thesis, I set out to develop a series of deep learning models that can accurately identify cancer type based on patterns of somatic mutations, and then address challenges for translating this model into a clinical setting, by developing statistical methods for quantifying predictive uncertainty in deep neural networks.

Despite a plethora of tumour typing algorithms using alterations to cancer-associated genes, computational tumour typing methods either have relatively low accuracy or are restricted to classifying a small number of cancer types (Ma et al. 2006; Bender and Erlander 2009; Penson et al. 2019; Grewal et al. 2019; Yuan et al. 2016; Salvadores, Mas-Ponte, and Supek 2019). Recent studies that find strong associations between somatic passenger mutations and chromatin state provide a potential avenue for improving upon these methods (Schuster-Böckler and Lehner 2012; Supek and Lehner 2015; Polak et al. 2015). The work presented in Chapter 2 outlines the comparison of multiple mutation-derived feature sets for identifying cancer type and results in the development of a deep learning system that can accurately discriminate between 24 common cancer types. Using random forest classifiers, we show that passenger mutation derived features, corresponding to regional-mutation density of somatic SNVs and mutation types outperform both driver mutation derived features and features associated with large-scale mutations such as copy-number variants and structural variants. To improve the classifier's performance, we develop a series of deep neural networks trained on the best performing passenger mutation derived features. This model's performance was significantly higher than the random forest model, emphasizing the ability for deep neural networks to learn highly accurate models from complex data. Interestingly, the work in this chapter demonstrated that adding explicit information about alterations to cancer-associated genes doesn't significantly improve model performance. When a deep neural network was trained using the passenger derived features with the addition of explicit information of which cancer-associated alterations occurred in the tumours, classification performance failed to significantly improve. This result suggests that passenger mutations are sufficient for accurately identifying cancer type. While previous work has demonstrated a relationship between regional mutation density and cell-of-origin, most previous work has been done by aggregating mutations over multiple tumour samples (Polak et al. 2015). Therefore, this result is one of the first studies to demonstrate that a non-linear model (a deep neural network) can learn a strong relationship between mutation rate derived from a single cancer sample and cell-of-origin. The strength of this relationship is further emphasized by the patterns of misclassifications produced by the model. When the deep neural network made misclassifications, they tended to reflect shared biological characteristics. In particular, patterns of misclassifications potentially reflect common chromatin features that arise from tumours that share developmental origins. Despite these misclassifications, the classifier could typically discriminate between different cancers that arise in the same organ. This provides further evidence for the relationship between passenger mutations and cell-type.

This model, which was trained on a large collection of uniformly processed data, accurately generalized to data collected from additional sources, suggesting that the model has applicability in real-world scenarios. Furthermore, the model accurately identified the primary tumour site for a large collection of metastatic tumours coming from a diverse set of cancer types. While the classifier had lower accuracy on this dataset compared to data from primary tumours, many misclassifications may be the result of mutations caused by exposure to chemotherapeutic agents such as cisplatin and fluorouracil (Pich et al. 2019). The model provides state-of-the-art performance for discriminating between multiple cancer types. In comparison to classifiers trained on features traditionally thought to be more indicative of cell-of-origin, such as gene expression profiles, our classifier provides the similar or better performance when accounting for the relatively large number of cancer types the model is trained to identify. This classifier's overall high performance, coupled with the stability and robustness of DNA, means that this classifier has immediate clinical applicability in identifying primary tumour site for metastatic tumours of unknown origin.

Despite the relatively strong performance of this model, challenges exist for translating it into a clinical setting. First, the model is only trained to identify 24 cancer types. Second, in a clinical setting where a neural network's predictions may form the basis for making patient-specific decisions, a neural network needs to provide well-calibrated estimates of model uncertainty. In Chapter 3, I address these challenges by developing and benchmarking multiple deep learning and statistical methods for extending the classifier to a greater number of cancer types and quantifying the predictive uncertainty of a deep neural network. Using a deep ensemble architecture, I was able to extend the classifier to discriminate between 29 cancer types accurately. Despite increasing the number of cancer types the model identifies, overall accuracy and performance are comparable to that for the classifier presented in Chapter 2. To assess in-distribution uncertainty - the ability for the classifier to quantify how reliable the predictions it makes are for cancer samples that come from cancer types the model is trained to identify - I assessed confidence-calibration, and then implemented and compared several post-hoc calibration methods. Overall, the original deep ensemble model provided highly well-calibrated predictions, an essential task in a clinical setting. To use the model's predictive uncertainty to automatically identify cancer samples from rare cancer types that the model was not trained to classify, I made use of the entropy of the predictive distribution. Using the entropy of the predictive distribution of the deep ensemble, I accurately identified cancer samples that come from cancer types the model is not trained to identify, allowing for the model to automatically rule out cancer samples that it cannot make high-confidence predictions on.

My thesis research demonstrates the utility of a DNA-based tumour classifier trained primarily on information from somatic passenger mutations. Second, my thesis research is one of the first demonstrations of the utility of confidence-calibration and uncertainty quantification in deep learning applied to cancer genomics. The classifiers I have developed now form the basis of work not discussed in this thesis. They are being used as the basis for an online tumour typing application and will be tested on a large set of clinical tumour samples. In the rest of this chapter, I outline a number of future directions stemming from the work I've presented in this thesis.

4.2 Correcting therapy-induced mutations

Cancer therapy has the potential for introducing a large number of mutations in a tumour. Work studying the mutational footprints of chemotherapy suggests that chemotherapy can contribute more than 50% of the mutations found in a tumour sample (Pich et al. 2019). Additional work has uncovered mutational spectra associated with radiation therapy (Behjati et al. 2016). As seen in both Chapter 2 and Chapter 3, the classifier, despite doing well on treated samples, had a small drop-off in accuracy compared to performance on untreated tumours. As samples in clinical scenarios, particularly those corresponding to patients with multiple primary tumours, may have exposure to cancer therapies, correcting for the effects of these variants may improve performance in real-world settings.

The characterization of mutational spectra associated with cancer therapy falls in the general framework of mutational signature analysis. At its core, mutational signature analysis works by taking observed counts of mutation types in a tumour and learning a latent representation of the mutational spectra, which can be grouped into mutational signatures. More specifically, the frameworks used to perform this analysis can be seen as variants of topic modelling, which aims to summarize topics (signatures) based on the counts of words (mutation types) found in documents (tumour genomes) (Alexandrov et al. 2020; Blei 2003). To correct for mutations associated with chemotherapy, mutational spectra can be embedded in the latent signature space, the contribution of therapy associated mutational signatures can then be removed, and the corrected mutational signatures can be used to reconstruct mutation types that no longer have contributions from mutational signatures associated with cancer therapy. While the traditional statistical framework for topic modelling makes this difficult, recent advances in amortized variational inference allow for topic models to be learned using neural networks (Figurnov, Mohamed, and Mnih 2018). By learning a topic model using a variational autoencoder, it becomes possible to fit mutational signatures to input samples, and then work directly in the mutational signature space to remove the contribution of mutational signatures associated with exposure to therapy. The altered mutational signature representation can then be used to reconstruct a mutation types vector that no longer contains mutations associated with cancer therapy. Using these approaches, it would be interesting to study the effects of removing mutation types associated with cancer therapy on overall classification performance.

An alternative approach would be to focus on the VAF of the mutations in a tumour sample. Mutations contributed by cancer therapy may be present at lower VAF than other mutations within the tumour (Pich et al. 2019). Using the VAF of mutations, it may be possible to reduce the contribution of therapy-induced mutations by simply removing low VAF mutations. This approach has two potential pitfalls. First, clonal expansions following exposure to therapy can increase the VAF of therapy-associated mutations, making them difficult to distinguish from other mutations. Second, the activity of mutational signatures varies throughout cancer evolution, with some mutational signatures becoming active in the later stages of cancer development (Gerstung et al. 2017). The mutations caused by mutational signatures active in the later stages of cancer development would also be present at lower VAF than other mutations and may be removed alongside mutations potentially associated with cancer therapy.

4.3 Incorporation of additional data modalities

Several different modalities have been used as features for developing tumour typing classifiers. Some success has been seen using gene expression data, cancer-associated mutations, and methylation assays (Ma et al. 2006; Bender and Erlander 2009; Penson et al. 2019; Grewal et al. 2019; Yuan et al. 2016; Locke et al. 2019). While these models have shown some success, they have been limited either in overall accuracy or in the number of cancer types they can classify. Data from modalities such as RNA-sequencing has the benefit of providing information about the tumour's current phenotype, which may be beneficial as the understanding of precision medicine advances.

In future studies, it would be interesting to incorporate data from additional modalities such as RNA-sequencing or methylation assays to the passenger mutation based classifier. Using passenger mutations will allow for ancestral information about a primary tumour to be represented in the classifier, and would offset any potential uncertainty introduced by using gene expression from highly undifferentiated tumours. The addition of gene expression data and methylation data may provide the classifier with the ability to differentiate highly related tumour types that have subtle differences not present at the level of regional mutation density, and would provide a readout of the tumour's current phenotype, which may be beneficial for targeted therapy.

4.4 Molecular subtyping

Tumours can be categorized at different levels of resolution. The work in this thesis has focused on developing methods to categorize tumours at the level of cell-of-origin, which provides vital information for clinical decision making. As molecular characterization of tumours has advanced, tumours have started to become categorized at the level of molecular subtypes. Molecular subtyping of tumours allows tumours of a single type, such as pancreatic adenocarcinomas, to be further subdivided based on molecular characteristics such as alterations to cancer-associated genes, gene expression signatures and chromatin modifications (Moffitt et al. 2015; Lomberk et al. 2018; Hayward et al. 2017). The subtype of a tumour can be used as the basis for guiding therapy. For example, triple-negative breast cancer is unresponsive to endocrine therapy or HER2 inhibition (Yin et al. 2020). As molecular subtypes are often encoded in gene expression programs or chromatin state, its possible that distinct molecular subtypes of a tumour have differences in the regional mutation density, which may be used to identify tumour subtype.

Currently, the classifier described in this thesis does not attempt to discriminate between multiple subtypes of a tumour category. In the future, it will be of interest to explore the use of passenger mutation derived features for automatically assigning a tumour sample to a distinct molecular subtype. This would then allow the classifier described in this thesis to be extended such that it is able to first assign an input sample to a tumour category and then able to assign the cancer sample to a molecular subtype of the tumour category.

An alternative approach would be to forego assignment to tumour type and use unsupervised or self-supervised learning methods to learn new groupings of tumours based on patterns of somatic passenger mutations. This approach would learn to group cancer samples based on patterns of mutations and then make use of gene expression programs from cancer samples to define the new tumour groupings' molecular characteristics. Using this approach, tumour samples would be assigned to passenger mutation-based subtypes, and treatment would be guided based on approaches that specifically target gene expression programs associated with the new subtypes.

4.5 Application to liquid biopsies and cell-free DNA

The presence of cell-free DNA (cfDNA) molecules has been known for over 50 years (Mandel and Metais 1948). Subsequent work has demonstrated that cancer patients have more cfDNA than healthy individuals, likely due to cancer cells shedding fragments of DNA (Fleischhacker and Schmidt 2007). As these DNA fragments contain mutations and other alterations present within the tumour that shed them, they have potential utility for early cancer detection and monitoring of tumour progression. Unfortunately, cfDNA tends to be more sparse and contain noisier data than the DNA uncovered from bulk WGS (Zviran et al. 2020).

While the classifiers presented in this thesis make use of mutation counts from bulk WGS, this thesis's work also demonstrates the utility of using patterns of mutations to identify cancer type accurately. As cfDNA presents noisier data than bulk WGS, it may not be possible to rely on alterations to specific, predefined cancer-associated genes. Instead, methods that allow for high-sensitivity mutation calling at the level of individual reads, instead of at specific loci as is common for bulk WGS, may provide estimates of regional mutation density across the entire genome (Zviran et al. 2020). The work presented in Chapter 2 demonstrates that a classifier trained using only information from regional mutation density has the ability to accurately identify cancer type. As this classifier does not explicitly require locus-specific mutation calls, mutation calling at the read level from cfDNA is potentially applicable to the model I have developed. Using cfDNA, it may be possible to adapt the classifiers I have developed using transfer learning methods (*Transfer Learning*) so that the classifier can be applied for early-cancer detection using non-invasive liquid biopsies. Currently, however, the sparsity of variant calls from cfDNA render this approach difficult (Zviran et al. 2020), but future technological advances may allow for deep, genome-wide variant calling from cfDNA.

4.6 Inferring chromatin state from patterns of somatic mutations

Chromatin state is a defining characteristic of a cell-type. For tumours, the chromatin state of the initiating cells in a tumour can provide clues to the developmental origins of a tumour. For example, the chromatin state of tumour initiating cells may be used to determine if tumour initiating cells have characteristics of stem-like cells. As cancer cells have a large amount of plasticity, chromatin state tends to change throughout tumour evolution, with some cells diverging significantly from the cells that initiated the tumour (Lan et al. 2017). While single-cell approaches have allowed for the heterogeneity in chromatin accessibility to be characterized, these approaches only provide information about extant cells. They cannot provide information about ancestral chromatin states that were once present in a human tumour. (Pritykin et al. 2020).

Ancestral information about real human tumours can reliably be uncovered through the evolutionary reconstruction of mutations in a human tumour (Gerstung et al. 2017). Evolutionary reconstruction of mutations allows for mutations to be ordered throughout tumour development, and assigned to distinct epochs of evolution such as the early-clonal stage or the subclonal stage. This thesis's work has made use of the relationship between chromatin state and mutation rate to develop classifiers that can identify cancer type from regional mutation density (Schuster-Böckler and Lehner 2012; Supek and Lehner 2015; Polak et al. 2015). A future avenue of research would be to forgo identifying cell-type directly, and instead, use sequence-based neural networks such as transformer networks (Vaswani et al. 2017) to develop a model that uses regional mutation density to infer chromatin accessibility. With this model, it may be possible to use the evolutionary reconstruction of mutations in tumour evolution to describe how chromatin accessibility changes throughout cancer evolution. Using multi-region sequencing studies, spatial chromatin heterogeneity may also be analyzed using this approach (Jamal-Hanjani et al. 2017).

4.7 Closing Remarks

Challenging scenarios arise when determining cancer type for metastatic lesions. In extreme cases, patients present with metastatic tumours that cannot be identified using imaging studies, examination by a pathologist, and immunohistochemistry. Large-scale genome sequencing studies have uncovered strong relationships between somatic mutations and cancer type, opening an avenue for using genomics as a tool for identifying cancer type. The work presented in this thesis demonstrates the utility of using somatic passenger mutations and deep learning as a tool for identifying cancer type and contains one of the first applications of confidence-calibration in deep learning applied to genomics.

As sequencing of clinical tumour samples becomes more accessible, the whole-genomes of tumour samples will routinely be sequenced in clinical scenarios. It is hoped that the methods described in this thesis will serve as an adjunct to traditional diagnostic approaches, allowing for the accuracy and reliability of cancer diagnosis to be improved.

Bibliography

Abadi, Martín et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.* Software available from tensorflow.org. URL: http://tensorflow.org/.

Al-Abbadi, Mousa A. (2011). "Basics of cytology". In: *Avicenna Journal of Medicine* 1.1, pp. 18–28. ISSN: 2231-0770. DOI: 10.4103/2231-0770.83719. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3507055/ (visited on 10/30/2020).

Al-Farsi, Khalil (Jan. 2013). "Multiple Myeloma: An Update". In: *Oman Medical Journal* 28.1, pp. 3–11. ISSN: 1999-768X. DOI: 10.5001/omj.2013.02. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3562980/ (visited on 09/19/2020).

Alexandrov, L. B. et al. (2013). "Deciphering signatures of mutational processes operative in human cancer". In: *Cell Rep.* 3. DOI: 10.1016/j.celrep.2012.12.008. URL: https://doi.org/10.1016/j.celrep.2012.12.008.

Alexandrov, Ludmil B. et al. (Feb. 2020). "The repertoire of mutational signatures in human cancer". In: *Nature* 578.7793, pp. 94–101. ISSN: 1476-4687. DOI: 10.1038/s41586-020-1943-3. URL: https://www.nature.com/articles/s41586-020-1943-3 (visited on 09/12/2020).

Alshareeda, Alaa T. et al. (Apr. 6, 2020). "Cancer of Unknown Primary Site: Real Entity or Misdiagnosed Disease?" In: *Journal of Cancer* 11.13, pp. 3919–3931. ISSN: 1837-9664. DOI: 10.7150/jca.42880. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7171483/ (visited on 09/18/2020).

Amer, Magid H. (2014). "Multiple neoplasms, single primaries, and patient survival". In: *Cancer Management and Research* 6, pp. 119–134. ISSN: 1179-1322. DOI: 10.2147/CMAR.S57378.

Amersfoort, Joost van et al. (2020). "Uncertainty Estimation Using a Single Deep Deterministic Neural Network". In:

Amouroux, Rachel et al. (May 2010). "Oxidative stress triggers the preferential assembly of base excision repair complexes on open chromatin regions". In: *Nucleic Acids Research* 38.9, pp. 2878–2890. ISSN: 0305-1048. DOI: 10.1093/nar/gkp1247. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2875005/ (visited on 11/05/2020).

Anastasiadou, Eleni, Leni S. Jacob, and Frank J. Slack (Jan. 2018). "Non-coding RNA networks in cancer". In: *Nature Reviews Cancer* 18.1. Number: 1 Publisher: Nature Publishing Group, pp. 5–18. ISSN: 1474-1768. DOI: 10.1038/nrc.2017.99. URL: http://www.nature.com/articles/nrc.2017.99 (visited on 11/03/2020).

Angus, Lindsay et al. (Oct. 2019). "The genomic landscape of metastatic breast cancer highlights changes in mutation and signature frequencies". In: *Nature Genetics* 51.10, pp. 1450–1458. ISSN: 1546-1718. DOI: 10.1038/s41588-019-0507-7. URL: https://www.nature.com/articles/s41588-019-0507-7 (visited on 10/31/2020).

Aung, K. L. (2018). "Genomics-driven precision medicine for advanced pancreatic cancer: early results from the COMPASS Trial". In: *Clin. Cancer Res.* 24. DOI: 10.1158/1078-0432.CCR-17-2994. URL: https://doi.org/10.1158/1078-0432.CCR-17-2994.

Bahrami, A., L. D. Truong, and J. Y. Ro (2008). "Undifferentiated tumor: true identity by immunohistochemistry". In: *Arch. Pathol. Lab. Med.* 132.

Bakhoum, Samuel F. et al. (Jan. 17, 2018). "Chromosomal instability drives metastasis through a cytosolic DNA response". In: *Nature* 553.7689. Publisher: Nature Publishing Group, pp. 467–472. ISSN: 0028-0836. DOI: 10.1038/nature25432. URL: http://www.nature.com/doifinder/10.1038/nature25432 (visited on 03/21/2018).

Behjati, Sam et al. (Sept. 12, 2016). "Mutational signatures of ionizing radiation in second malignancies". In: *Nature Communications* 7.1. Number: 1 Publisher: Nature Publishing Group, p. 12605. ISSN: 2041-1723. DOI: 10.1038/ncomms12605. URL: http://www.nature.com/articles/ncomms12605 (visited on 11/04/2020).

Bender, Richard A. and Mark G. Erlander (Feb. 2009). "Molecular classification of unknown primary cancer". In: *Seminars in Oncology* 36.1, pp. 38–43. ISSN: 0093-7754. DOI: 10.1053/j.seminoncol.2008.10.002.

Biankin, A. V. (2012). "Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes". In: *Nature* 491. DOI: 10.1038/nature11547. URL: https://doi.org/10.1038/nature11547.

Blei, David M (Mar. 1, 2003). "Latent Dirichlet Allocation". In: p. 30.

Blundell, Charles et al. (May 21, 2015). "Weight Uncertainty in Neural Networks". In: *arXiv:1505.05424 [cs, stat]*. arXiv: 1505.05424. URL: http://arxiv.org/abs/1505.05424 (visited on 11/05/2020).

Boekhout, Annelies H., Jos H. Beijnen, and Jan H.M. Schellens (June 2011). "Trastuzumab". In: *The Oncologist* 16.6, pp. 800–810. ISSN: 1083-7159. DOI: 10.1634/theoncologist.2010-0035. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3228213/ (visited on 11/19/2020).

Bragazzi, Maria Consiglia et al. (2018). "New insights into cholangiocarcinoma: multiple stems and related cell lineages of origin". In: *Annals of Gastroenterology* 31.1, pp. 42–55. ISSN: 1108-7471. DOI: 10.20524/aog.2017.0209. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5759612/ (visited on 10/17/2020).

Breiman, Leo (Oct. 1, 2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. URL: https://doi.org/10.1023/A:1010933404324 (visited on 11/05/2020).

Bridgewater, J. et al. (2008). "Gene expression profiling may improve diagnosis in patients with carcinoma of unknown primary". In: *Br. J. Cancer* 98. DOI: 10.1038/sj.bjc.6604315. URL: https://doi.org/10.1038/sj.bjc.6604315.

Brochu, Eric, Matthew Hoffman, and Nando De Freitas (2011). *Portfolio Allocation for Bayesian Optimization.* arXiv: 1009.5419v2. (Visited on 09/26/2018).

Brunner, Simon F. et al. (Oct. 2019). "Somatic mutations and clonal dynamics in healthy and cirrhotic human liver". In: *Nature* 574.7779, pp. 538–542. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1670-9. URL: https://www.nature.com/articles/s41586-019-1670-9 (visited on 10/24/2019).

Campbell, Peter J et al. (Feb. 2020). "Pan-cancer analysis of whole genomes". In: *Nature* 578.7793, pp. 82–93. ISSN: 1476-4687. DOI: 10.1038/s41586-020-1969-6. URL: https://www.nature.com/articles/s41586-020-1969-6 (visited on 02/13/2020).

Capper, D. (2018). “DNA methylation-based classification of central nervous system tumours”. In: *Nature* 555. DOI: `10.1038/nature26000`. URL: `https://doi.org/10.1038/nature26000`.

Carlsson, Steven K, Shaun P Brothers, and Claes Wahlestedt (Oct. 13, 2014). “Emerging treatment strategies for glioblastoma multiforme.” In: *EMBO molecular medicine* 6.11. Publisher: Wiley-Blackwell, pp. 1359–70. ISSN: 1757-4684. DOI: `10.15252/emmm.201302627`. URL: `http://www.ncbi.nlm.nih.gov/pubmed/25312641` (visited on 01/25/2017).

Ceccarelli, M. (2016). “Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma”. In: *Cell* 164. DOI: `10.1016/j.cell.2015.12.028`. URL: `https://doi.org/10.1016/j.cell.2015.12.028`.

Ceyssens, Sarah and Sigrid Stroobants (2011). “Sarcoma”. In: *Methods in Molecular Biology (Clifton, N.J.)* 727, pp. 191–203. ISSN: 1940-6029. DOI: `10.1007/978-1-61779-062-1_11`.

Chaffer, Christine L. and Robert A. Weinberg (Jan. 2015). “How does multistep tumorigenesis really proceed?” In: *Cancer discovery* 5.1, pp. 22–24. ISSN: 2159-8274. DOI: `10.1158/2159-8290.CD-14-0788`. URL: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4295623/` (visited on 11/03/2020).

Chawla, N. V. et al. (June 1, 2002). “SMOTE: Synthetic Minority Over-sampling Technique”. In: *Journal of Artificial Intelligence Research* 16, pp. 321–357. ISSN: 1076-9757. DOI: `10.1613/jair.953`. arXiv: `1106.1813`. URL: `http://arxiv.org/abs/1106.1813` (visited on 11/06/2020).

Chen, Y. et al. (2015). “Classification of cancer primary sites using machine learning and somatic mutations”. In: *Biomed. Res. Int* 2015.

Chiang, Serena P. H., Ramon M. Cabrera, and Jeffrey E. Segall (July 1, 2016). “Tumor cell intravasation”. In: *American Journal of Physiology - Cell Physiology* 311.1, pp. C1–C14. ISSN: 0363-6143. DOI: `10.1152/ajpcell.00238.2015`. URL: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4967137/` (visited on 11/03/2020).

Chollet, Francois (2015). *Keras*. https:/keras.io. URL: `https://keras.io/` (visited on 06/01/2018).

Chu, D. and B. H. Park (2017). “Liquid biopsy: unlocking the potentials of cell-free DNA”. In: *Virchows Arch.* 471. DOI: `10.1007/s00428-017-2137-8`. URL: `https://doi.org/10.1007/s00428-017-2137-8`.

Cibulskis, Kristian et al. (Feb. 10, 2013). “Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples”. In: *Nature Biotechnology* 31.3, pp. 213–219. ISSN: 1087-0156. DOI: `10.1038/nbt.2514`. URL: `http://www.nature.com/doifinder/10.1038/nbt.2514` (visited on 01/24/2017).

Ciriello, G. (2013). “Emerging landscape of oncogenic signatures across human cancers”. In: *Nat. Genet* 45. DOI: `10.1038/ng.2762`. URL: `https://doi.org/10.1038/ng.2762`.

Croft, D. (2014). “The Reactome pathway knowledgebase”. In: *Nucleic Acids Res.* 42. DOI: `10.1093/nar/gkt1102`. URL: `https://doi.org/10.1093/nar/gkt1102`.

Davies, Helen et al. (Apr. 13, 2017). “HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures”. In: *Nature Medicine* 23.4. Publisher: Nature Publishing Group, pp. 517–525. ISSN: 1078-8956. DOI: `10.1038/nm.4292`. URL: `http://www.nature.com/articles/nm.4292` (visited on 02/07/2018).

Devlin, Jacob et al. (May 24, 2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *arXiv:1810.04805 [cs]*. arXiv: `1810.04805`. URL: `http://arxiv.org/abs/1810.04805` (visited on 11/27/2020).

Druker, Brian J. (2004). “Imatinib as a paradigm of targeted therapies”. In: *Advances in Cancer Research* 91, pp. 1–30. ISSN: 0065-230X. DOI: 10.1016/S0065-230X(04)91001-9.

Duncan, Bruce K. and Jeffrey H. Miller (Oct. 1980). “Mutagenic deamination of cytosine residues in DNA”. In: *Nature* 287.5782. Number: 5782 Publisher: Nature Publishing Group, pp. 560–561. ISSN: 1476-4687. DOI: 10.1038/287560a0. URL: http://www.nature.com/articles/287560a0 (visited on 11/04/2020).

Duraiyan, Jeyapradha et al. (Aug. 2012). “Applications of immunohistochemistry”. In: *Journal of Pharmacy & Bioallied Sciences* 4 (Suppl 2), S307–S309. ISSN: 0976-4879. DOI: 10.4103/0975-7406.100281. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3467869/ (visited on 09/06/2020).

D’cruze, Lawrence et al. (July 2013). “The Role of Immunohistochemistry in the Analysis of the Spectrum of Small Round Cell Tumours at a Tertiary Care Centre”. In: *Journal of Clinical and Diagnostic Research : JCDR* 7.7, pp. 1377–1382. ISSN: 2249-782X. DOI: 10.7860/JCDR/2013/5127.3132. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3749639/ (visited on 11/05/2020).

Esteva, Andre et al. (2017). “Dermatologist-level classification of skin cancer with deep neural networks”. In: *Nature*. Publisher: Nature Research. DOI: 10.1038/nature21056. (Visited on 01/25/2017).

Falk, Gary W (Feb. 2015). “Barrett’s oesophagus: frequency and prediction of dysplasia and cancer.” In: *Best practice & research. Clinical gastroenterology* 29.1. Publisher: NIH Public Access, pp. 125–38. ISSN: 1532-1916. DOI: 10.1016/j.bpg.2015.01.001. URL: http://www.ncbi.nlm.nih.gov/pubmed/25743461 (visited on 02/22/2017).

Fares, Jawad et al. (Mar. 12, 2020). “Molecular principles of metastasis: a hallmark of cancer revisited”. In: *Signal Transduction and Targeted Therapy* 5.1. Number: 1 Publisher: Nature Publishing Group, pp. 1–17. ISSN: 2059-3635. DOI: 10.1038/s41392-020-0134-x. URL: http://www.nature.com/articles/s41392-020-0134-x (visited on 11/03/2020).

Faure, Sandrine and Pascal De Santa Barbara (May 2011). “Molecular embryology of the foregut”. In: *Journal of Pediatric Gastroenterology and Nutrition* 52 (Suppl 1), S2–3. ISSN: 0277-2116. DOI: 10.1097/MPG.0b013e3182105a1a. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3374328/ (visited on 11/06/2020).

Ferracin, Manuela et al. (Sept. 2011). “MicroRNA profiling for the identification of cancers with unknown primary tissue-of-origin.” In: *The Journal of pathology* 225.1. Publisher: NIH Public Access, pp. 43–53. ISSN: 1096-9896. DOI: 10.1002/path.2915. URL: http://www.ncbi.nlm.nih.gov/pubmed/21630269 (visited on 09/10/2017).

Figurnov, Mikhail, Shakir Mohamed, and Andriy Mnih (2018). “Implicit Reparameterization Gradients”. In: *Advances in Neural Information Processing Systems* 31, pp. 441–452. URL: https://papers.nips.cc/paper/2018/hash/92c8c96e4c37100777c7190b76d28233-Abstract.html (visited on 11/09/2020).

Fleischhacker, M. and B. Schmidt (Jan. 1, 2007). “Circulating nucleic acids (CNAs) and cancer—A survey”. In: *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* 1775.1, pp. 181–232. ISSN: 0304-419X. DOI: 10.1016/j.bbcan.2006.10.001. URL: http://www.sciencedirect.com/science/article/pii/S0304419X0600059X (visited on 11/08/2020).

Forbes, S. A. (2017). “COSMIC: somatic cancer genetics at high-resolution”. In: *Nucleic Acids Res* 45. DOI: 10.1093/nar/gkw1121. URL: https://doi.org/10.1093/nar/gkw1121.

Fort, Stanislav, Huiyi Hu, and Balaji Lakshminarayanan (June 24, 2020). "Deep Ensembles: A Loss Landscape Perspective". In: *arXiv:1912.02757 [cs, stat]*. arXiv: 1912.02757. URL: http://arxiv.org/abs/1912.02757 (visited on 09/30/2020).

Fousteri, Maria and Leon HF Mullenders (Jan. 2008). "Transcription-coupled nucleotide excision repair in mammalian cells: molecular mechanisms and biological effects". In: *Cell Research* 18.1. Number: 1 Publisher: Nature Publishing Group, pp. 73–84. ISSN: 1748-7838. DOI: 10.1038/cr.2008.6. URL: http://www.nature.com/articles/cr20086 (visited on 11/04/2020).

Fox, Edward J et al. (2014). "Accuracy of Next Generation Sequencing Platforms". In: *Next generation, sequencing & applications* 1. DOI: 10.4172/jngsa.1000106. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4331009/ (visited on 11/27/2020).

Fu, Yu et al. (July 27, 2020). "Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis". In: *Nature Cancer*, pp. 1–11. ISSN: 2662-1347. DOI: 10.1038/s43018-020-0085-8. URL: https://www.nature.com/articles/s43018-020-0085-8 (visited on 07/27/2020).

Fueyo, Raquel et al. (Apr. 20, 2018). "Lineage specific transcription factors and epigenetic regulators mediate TGF-dependent enhancer activation". In: *Nucleic Acids Research* 46.7. Publisher: Oxford Academic, pp. 3351–3365. ISSN: 0305-1048. DOI: 10.1093/nar/gky093. URL: https://academic.oup.com/nar/article/46/7/3351/4846346 (visited on 11/19/2020).

Gal, Yarin and Zoubin Ghahramani (Oct. 4, 2016). "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning". In: *arXiv:1506.02142 [cs, stat]*. arXiv: 1506.02142. URL: http://arxiv.org/abs/1506.02142 (visited on 11/05/2020).

Gerstung, Moritz et al. (Aug. 30, 2017). "The evolutionary history of 2,658 cancers". In: *bioRxiv*. Publisher: Cold Spring Harbor Laboratory, p. 161562. DOI: 10.1101/161562. URL: https://www.biorxiv.org/content/early/2017/08/30/161562 (visited on 03/22/2018).

Glorot, Xavier and Yoshua Bengio (Mar. 31, 2010). "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. ISSN: 1938-7228. JMLR Workshop and Conference Proceedings, pp. 249–256. URL: http://proceedings.mlr.press/v9/glorot10a.html (visited on 11/05/2020).

Gomes, Ana P. et al. (Oct. 14, 2019). "Dynamic Incorporation of Histone H3 Variants into Chromatin Is Essential for Acquisition of Aggressive Traits and Metastatic Colonization". In: *Cancer Cell* 36.4. Publisher: Elsevier, 402–417.e13. ISSN: 1535-6108, 1878-3686. DOI: 10.1016/j.ccell.2019.08.006. URL: https://www.cell.com/cancer-cell/abstract/S1535-6108(19)30374-5 (visited on 11/05/2020).

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. http://www.deeplearningbook.org. MIT Press.

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy (Mar. 20, 2015). "Explaining and Harnessing Adversarial Examples". In: *arXiv:1412.6572 [cs, stat]*. arXiv: 1412.6572. URL: http://arxiv.org/abs/1412.6572 (visited on 03/25/2020).

Goodfellow, Ian J. et al. (June 10, 2014). "Generative Adversarial Networks". In: *arXiv:1406.2661 [cs, stat]*. arXiv: 1406.2661. URL: http://arxiv.org/abs/1406.2661 (visited on 11/05/2020).

Govindan, Ramaswamy et al. (Sept. 14, 2012). "Genomic landscape of non-small cell lung cancer in smokers and never-smokers". In: *Cell* 150.6, pp. 1121–1134. ISSN: 1097-4172. DOI: 10.1016/j.cell.2012.08.024.

Govindarajan, Rajeshwar et al. (Aug. 2012). “Microarray and its applications”. In: *Journal of Pharmacy & Bioallied Sciences* 4 (Suppl 2), S310–S312. ISSN: 0976-4879. DOI: 10.4103/0975-7406.100283. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3467903/ (visited on 11/03/2020).

Greco, F. Anthony (Dec. 1, 2013). “Molecular Diagnosis of the Tissue of Origin in Cancer of Unknown Primary Site: Useful in Patient Management”. In: *Current Treatment Options in Oncology* 14.4, pp. 634–642. ISSN: 1534-6277. DOI: 10.1007/s11864-013-0257-1. URL: https://doi.org/10.1007/s11864-013-0257-1 (visited on 09/02/2020).

Greco, F. Anthony et al. (2010). “Molecular profiling in unknown primary cancer: accuracy of tissue of origin prediction”. In: *The Oncologist* 15.5, pp. 500–506. ISSN: 1549-490X. DOI: 10.1634/theoncologist.2009-0328.

Greco, F. Anthony et al. (Feb. 16, 2012). “Carcinoma of Unknown Primary Site: Outcomes in Patients with a Colorectal Molecular Profile Treated with Site Specific Chemotherapy”. In: *Journal of Cancer Therapy* 3.1. Number: 1 Publisher: Scientific Research Publishing, pp. 37–43. DOI: 10.4236/jct.2012.31005. URL: http://www.scirp.org/Journal/Paperabs.aspx?paperid=17214 (visited on 11/04/2020).

Grewal, Jasleen K. et al. (Apr. 26, 2019). “Application of a Neural Network Whole Transcriptome–Based Pan-Cancer Method for Diagnosis of Primary and Metastatic Cancers”. In: *JAMA Network Open* 2.4, e192597. ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2019.2597. URL: http://jamanetworkopen.jamanetwork.com/article.aspx?doi=10.1001/jamanetworkopen.2019.2597 (visited on 11/05/2020).

Guo, Chuan et al. (Aug. 3, 2017). “On Calibration of Modern Neural Networks”. In: *arXiv:1706.04599 [cs]*. arXiv: 1706.04599. URL: http://arxiv.org/abs/1706.04599 (visited on 03/25/2020).

Haas, Brian J. et al. (Dec. 2019). “Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods”. In: *Genome Biology* 20.1. Number: 1 Publisher: BioMed Central, pp. 1–16. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1842-9. URL: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1842-9 (visited on 11/03/2020).

Hainsworth, John D. et al. (June 2012). “A retrospective study of treatment outcomes in patients with carcinoma of unknown primary site and a colorectal cancer molecular profile”. In: *Clinical Colorectal Cancer* 11.2, pp. 112–118. ISSN: 1938-0674. DOI: 10.1016/j.clcc.2011.08.001.

Hainsworth, John D. et al. (Jan. 10, 2013). “Molecular gene expression profiling to predict the tissue of origin and direct site-specific therapy in patients with carcinoma of unknown primary site: a prospective trial of the Sarah Cannon research institute”. In: *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 31.2, pp. 217–223. ISSN: 1527-7755. DOI: 10.1200/JCO.2012.43.3755.

Hall, Bradley R. et al. (Apr. 10, 2018). “Advanced pancreatic cancer: a meta-analysis of clinical trials over thirty years”. In: *Oncotarget* 9.27, pp. 19396–19405. ISSN: 1949-2553. DOI: 10.18632/oncotarget.25036. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5922405/ (visited on 11/04/2020).

Han, X., J. Wang, and Y. Sun (2017). “Circulating tumor DNA as biomarkers for cancer detection”. In: *Genomics Proteom. Bioinforma.* 15. DOI: 10.1016/j.gpb.2016.12.004. URL: https://doi.org/10.1016/j.gpb.2016.12.004.

Hanahan, Douglas and Robert A. Weinberg (Mar. 2011). “Hallmarks of Cancer: The Next Generation”. In: *Cell* 144.5, pp. 646–674. ISSN: 00928674. DOI: `10.1016/j.cell.2011.02.013`. URL: `http://linkinghub.elsevier.com/retrieve/pii/S0092867411001279` (visited on 01/24/2017).

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition.* 2nd ed. Springer Series in Statistics. New York: Springer-Verlag. Chap. Model Assessment and Selection. ISBN: 978-0-387-84857-0. DOI: `10.1007/978-0-387-84858-7`. URL: `https://www.springer.com/gp/book/9780387848570` (visited on 11/22/2020).

Hayat, Matthew J. et al. (Jan. 2007). “Cancer statistics, trends, and multiple primary cancer analyses from the Surveillance, Epidemiology, and End Results (SEER) Program”. In: *The Oncologist* 12.1, pp. 20–37. ISSN: 1083-7159. DOI: `10.1634/theoncologist.12-1-20`.

Hayward, Nicholas K. et al. (2017). “Whole-genome landscapes of major melanoma subtypes”. In: *Nature* 545.7653. ISBN: 0028-0836. ISSN: 14764687. DOI: `10.1038/nature22071`.

He, Kaiming et al. (Dec. 10, 2015a). “Deep Residual Learning for Image Recognition”. In: *arXiv:1512.03385 [cs]*. arXiv: `1512.03385`. URL: `http://arxiv.org/abs/1512.03385` (visited on 11/27/2020).

— (Feb. 6, 2015b). “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification”. In: *arXiv:1502.01852 [cs]*. arXiv: `1502.01852`. URL: `http://arxiv.org/abs/1502.01852` (visited on 11/06/2020).

Head, Tim et al. (Mar. 25, 2018). “scikit-optimize/scikit-optimize: v0.5.2”. In: DOI: `10.5281/ZENODO.1207017`. URL: `https://zenodo.org/record/1207017` (visited on 11/15/2018).

Hedley, D. W., J. A. Leary, and F. Kirsten (Feb. 1985). “Metastatic adenocarcinoma of unknown primary site: abnormalities of cellular DNA content and survival”. In: *European Journal of Cancer & Clinical Oncology* 21.2, pp. 185–189. ISSN: 0277-5379. DOI: `10.1016/0277-5379(85)90171-3`.

Heyer, Erin E. et al. (Mar. 27, 2019). “Diagnosis of fusion genes using targeted RNA sequencing”. In: *Nature Communications* 10.1, p. 1388. ISSN: 2041-1723. DOI: `10.1038/s41467-019-09374-9`. URL: `https://www.nature.com/articles/s41467-019-09374-9` (visited on 10/30/2020).

Hodgkinson, Alan, Ying Chen, and Adam Eyre-Walker (2012). “The large-scale distribution of somatic mutations in cancer genomes”. In: *Human Mutation* 33.1, pp. 136–143. ISSN: 1098-1004. DOI: `10.1002/humu.21616`. URL: `https://onlinelibrary.wiley.com/doi/abs/10.1002/humu.21616` (visited on 09/14/2020).

Hollstein, M. et al. (July 5, 1991). “p53 mutations in human cancers”. In: *Science (New York, N.Y.)* 253.5015, pp. 49–53. ISSN: 0036-8075. DOI: `10.1126/science.1905840`.

Hollstein, M. et al. (Dec. 17, 1999). “New approaches to understanding p53 gene tumor mutation spectra”. In: *Mutation Research* 431.2, pp. 199–209. ISSN: 0027-5107. DOI: `10.1016/s0027-5107(99)00162-1`.

Hornick, Jason L. (Jan. 2014). “Novel uses of immunohistochemistry in the diagnosis and classification of soft tissue tumors”. In: *Modern Pathology* 27.1, S47–S63. ISSN: 1530-0285. DOI: `10.1038/modpathol.2013.177`. URL: `https://www.nature.com/articles/modpathol2013177` (visited on 09/06/2020).

Hu, Zheng and Christina Curtis (June 25, 2020). “Looking backward in time to define the chronology of metastasis”. In: *Nature Communications* 11.1. Number: 1 Publisher: Nature Publishing Group, p. 3213. ISSN: 2041-1723. DOI: `10.1038/s41467-020-16995-y`. URL: `http://www.nature.com/articles/s41467-020-16995-y` (visited on 11/03/2020).

Hunter, Chris et al. (Apr. 15, 2006). “A Hypermutation Phenotype and Somatic MSH6 Mutations in Recurrent Human Malignant Gliomas after Alkylator Chemotherapy”. In: *Cancer Research* 66.8. Publisher: American Association for Cancer Research Section: Priority Reports, pp. 3987–3991. ISSN: 0008-5472, 1538-7445. DOI: 10.1158/0008-5472.CAN-06-0127. URL: http://cancerres.aacrjournals.org/content/66/8/3987 (visited on 11/04/2020).

Hyman, David M. et al. (Aug. 20, 2015). “Vemurafenib in Multiple Nonmelanoma Cancers with BRAF V600 Mutations”. In: *New England Journal of Medicine* 373.8, pp. 726–736. ISSN: 0028-4793. DOI: 10.1056/NEJMoa1502309. URL: https://doi.org/10.1056/NEJMoa1502309 (visited on 09/06/2020).

Inamura, Kentaro (Mar. 14, 2018). “Update on Immunohistochemistry for the Diagnosis of Lung Cancer”. In: *Cancers* 10.3. ISSN: 2072-6694. DOI: 10.3390/cancers10030072. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5876647/ (visited on 09/20/2020).

“International network of cancer genome projects” (2010). In: *Nature* 464. DOI: 10.1038/nature08987. URL: https://doi.org/10.1038/nature08987.

Jamal-Hanjani, Mariam et al. (June 1, 2017). “Tracking the Evolution of Non–Small-Cell Lung Cancer”. In: *New England Journal of Medicine* 376.22. Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMoa1616288, pp. 2109–2121. ISSN: 0028-4793. DOI: 10.1056/NEJMoa1616288. URL: https://doi.org/10.1056/NEJMoa1616288 (visited on 11/03/2020).

Jiao, Wei et al. (Feb. 5, 2020). “A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns”. In: *Nature Communications* 11.1, pp. 1–12. ISSN: 2041-1723. DOI: 10.1038/s41467-019-13825-8. URL: https://www.nature.com/articles/s41467-019-13825-8 (visited on 03/25/2020).

Kandoth, C. (2013). “Mutational landscape and significance across 12 major cancer types”. In: *Nature* 502. DOI: 10.1038/nature12634. URL: https://doi.org/10.1038/nature12634.

Karimzadeh, Mehran et al. (Feb. 13, 2020). “Viral integration transforms chromatin to drive oncogenesis”. In: *bioRxiv*, p. 2020.02.12.942755. DOI: 10.1101/2020.02.12.942755. URL: https://www.biorxiv.org/content/10.1101/2020.02.12.942755v1 (visited on 08/20/2020).

Kei, Si and Oyedele A. Adeyi (May 1, 2020). “Practical Application of Lineage-Specific Immunohistochemistry Markers: Transcription Factors (Sometimes) Behaving Badly”. In: *Archives of Pathology & Laboratory Medicine* 144.5, pp. 626–643. ISSN: 0003-9985. DOI: 10.5858/arpa.2019-0226-RA. URL: https://meridian.allenpress.com/aplm/article/144/5/626/427467/Practical-Application-of-Lineage-Specific (visited on 09/08/2020).

Keller, Florian et al. (Nov. 2011). “Carcinoma of unknown primary in the head and neck: comparison between positron emission tomography (PET) and PET/CT”. In: *Head & Neck* 33.11, pp. 1569–1575. ISSN: 1097-0347. DOI: 10.1002/hed.21635.

Kim, C.S. et al. (Oct. 2018a). “Survival outcome differences based on treatments used and knowledge of the primary tumour site for patients with cancer of unknown and known primary in Ontario”. In: *Current Oncology* 25.5, pp. 307–316. ISSN: 1198-0052. DOI: 10.3747/co.25.4003. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6209558/ (visited on 11/04/2020).

Kim, Jihun et al. (Jan. 2017). “Integrated genomic characterization of oesophageal carcinoma”. In: *Nature* 541.7636, pp. 169–175. ISSN: 1476-4687. DOI: 10.1038/nature20805. URL: https://www.nature.com/articles/nature20805 (visited on 10/30/2020).

Kim, Sangtae et al. (Aug. 2018b). “Strelka2: fast and accurate calling of germline and somatic variants”. In: *Nature Methods* 15.8. Number: 8 Publisher: Nature Publishing Group, pp. 591–594. ISSN: 1548-

7105. DOI: 10.1038/s41592-018-0051-x. URL: http://www.nature.com/articles/s41592-018-0051-x (visited on 11/25/2020).

Kingma, Diederik P. and Jimmy Ba (Dec. 22, 2014). "Adam: A Method for Stochastic Optimization". In: arXiv: 1412.6980. URL: http://arxiv.org/abs/1412.6980 (visited on 11/15/2018).

Kingma, Diederik P. and Max Welling (May 1, 2014). "Auto-Encoding Variational Bayes". In: *arXiv:1312.6114 [cs, stat]*. arXiv: 1312.6114. URL: http://arxiv.org/abs/1312.6114 (visited on 03/25/2020).

Koboldt, Daniel C. et al. (Mar. 2012). "VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing". In: *Genome Research* 22.3, pp. 568–576. ISSN: 1088-9051. DOI: 10.1101/gr.129684.111. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3290792/ (visited on 11/25/2020).

Kucab, Jill E. et al. (Apr. 2019). "A Compendium of Mutational Signatures of Environmental Agents". In: *Cell* 0.0. Publisher: Elsevier. ISSN: 00928674. DOI: 10.1016/j.cell.2019.03.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S0092867419302636 (visited on 04/12/2019).

Kull, Meelis et al. (Oct. 28, 2019). "Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration". In: *arXiv:1910.12656 [cs, stat]*. arXiv: 1910.12656. URL: http://arxiv.org/abs/1910.12656 (visited on 01/15/2020).

Kurahashi, Issei et al. (May 9, 2013). "A Microarray-Based Gene Expression Analysis to Identify Diagnostic Biomarkers for Unknown Primary Cancer". In: *PLOS ONE* 8.5. Publisher: Public Library of Science, e63249. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0063249. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0063249 (visited on 11/03/2020).

Kurzrock, Razelle et al. (May 20, 2003). "Philadelphia ChromosomePositive Leukemias: From Basic Mechanisms to Molecular Therapeutics". In: *Annals of Internal Medicine* 138.10. Publisher: American College of Physicians, pp. 819–830. ISSN: 0003-4819. DOI: 10.7326/0003-4819-138-10-200305200-00010. URL: https://www.acpjournals.org/doi/10.7326/0003-4819-138-10-200305200-00010 (visited on 11/06/2020).

Kübler, Kirsten et al. (Jan. 11, 2019). "Tumor mutational landscape is a record of the pre-malignant state". In: *bioRxiv*. Publisher: Cold Spring Harbor Laboratory, p. 517565. DOI: 10.1101/517565. URL: https://www.biorxiv.org/content/10.1101/517565v1 (visited on 05/16/2019).

Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell (Nov. 3, 2017). "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles". In: *arXiv:1612.01474 [cs, stat]*. arXiv: 1612.01474. URL: http://arxiv.org/abs/1612.01474 (visited on 03/25/2020).

Lan, Xiaoyang et al. (2017). "Fate mapping of human glioblastoma reveals an invariant stem cell hierarchy". In: *Nature* 549.7671, pp. 227–232. ISSN: 1476-4687. DOI: 10.1038/nature23666.

Lander, Eric S. et al. (Feb. 2001). "Initial sequencing and analysis of the human genome". In: *Nature* 409.6822. Number: 6822 Publisher: Nature Publishing Group, pp. 860–921. ISSN: 1476-4687. DOI: 10.1038/35057062. URL: http://www.nature.com/articles/35057062 (visited on 11/02/2020).

Laplante, Jean-François and Moulay A. Akhloufi (July 2020). "Predicting Cancer Types From miRNA Stem-loops Using Deep Learning*". In: *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*. 2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC). ISSN: 1558-4615, pp. 5312–5315. DOI: 10.1109/EMBC44109.2020.9176345.

Lawrence, M. S. (2013). "Mutational heterogeneity in cancer and the search for new cancer-associated genes". In: *Nature* 499. DOI: 10.1038/nature12213. URL: https://doi.org/10.1038/nature12213.

Lee, Christian A., Diala Abd-Rabbo, and Jüri Reimand (July 30, 2020). "Functional and genetic determinants of mutation rate variability in regulatory elements of cancer genomes". In: *bioRxiv*, p. 2020.07.29.226373. DOI: 10.1101/2020.07.29.226373. URL: https://www.biorxiv.org/content/10.1101/2020.07.29.226373v1 (visited on 07/30/2020).

Lee-Six, Henry et al. (Sept. 13, 2018). "The landscape of somatic mutation in normal colorectal epithelial cells". In: *bioRxiv*. Publisher: Cold Spring Harbor Laboratory, p. 416800. DOI: 10.1101/416800. URL: https://www.biorxiv.org/content/10.1101/416800v1 (visited on 05/15/2019).

Li, Yilong et al. (Feb. 2020). "Patterns of somatic structural variation in human cancer genomes". In: *Nature* 578.7793. Number: 7793 Publisher: Nature Publishing Group, pp. 112–121. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1913-9. URL: http://www.nature.com/articles/s41586-019-1913-9 (visited on 11/05/2020).

Lin, Jessica J. et al. (Aug. 3, 2020). "Small cell transformation of ROS1 fusion-positive lung cancer resistant to ROS1 inhibition". In: *npj Precision Oncology* 4.1. Number: 1 Publisher: Nature Publishing Group, pp. 1–8. ISSN: 2397-768X. DOI: 10.1038/s41698-020-0127-9. URL: http://www.nature.com/articles/s41698-020-0127-9 (visited on 11/02/2020).

Locke, Warwick J. et al. (Nov. 14, 2019). "DNA Methylation Cancer Biomarkers: Translation to the Clinic". In: *Frontiers in Genetics* 10. ISSN: 1664-8021. DOI: 10.3389/fgene.2019.01150. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6870840/ (visited on 11/09/2020).

Lodato, Michael A et al. (Feb. 2, 2018). "Aging and neurodegeneration are associated with increased mutations in single human neurons." In: *Science (New York, N.Y.)* 359.6375. Publisher: American Association for the Advancement of Science, pp. 555–559. ISSN: 1095-9203. DOI: 10.1126/science.aao4426. URL: http://www.ncbi.nlm.nih.gov/pubmed/29217584 (visited on 07/30/2019).

Lomberk, Gwen et al. (2018). "Distinct epigenetic landscapes underlie the pathobiology of pancreatic cancer subtypes." In: *Nature communications* 9.1. Publisher: Nature Publishing Group, p. 1978. ISSN: 2041-1723. DOI: 10.1038/s41467-018-04383-6. URL: http://www.ncbi.nlm.nih.gov/pubmed/29773832 (visited on 05/15/2019).

Ma, Xiao-Jun et al. (Apr. 2006). "Molecular classification of human cancers using a 92-gene real-time quantitative polymerase chain reaction assay". In: *Archives of Pathology & Laboratory Medicine* 130.4, pp. 465–473. ISSN: 1543-2165. DOI: 10.1043/1543-2165(2006)130[465:MCOHCU]2.0.CO;2.

MacKay, David J. C. (1992). "Bayesian methods for adaptive models". PhD thesis. California Institute of Technology. DOI: 10.7907/H3A1-WM07. URL: https://resolver.caltech.edu/CaltechETD:etd-01042007-131447 (visited on 11/05/2020).

"Chapter 20 - Hematopoietic Cancers" (Jan. 1, 2014). In: *Primer to the Immune Response (Second Edition)*. Ed. by Tak W. Mak, Mary E. Saunders, and Bradley D. Jett. Boston: Academic Cell, pp. 553–585. ISBN: 978-0-12-385245-8. DOI: 10.1016/B978-0-12-385245-8.00020-0. URL: http://www.sciencedirect.com/science/article/pii/B9780123852458000200 (visited on 11/02/2020).

Mandel, P. and P. Metais (Feb. 1948). "Nuclear Acids In Human Blood Plasma". In: *Comptes Rendus Des Seances De La Societe De Biologie Et De Ses Filiales* 142.3, pp. 241–243. ISSN: 0037-9026.

Marquard, A. M. (2015). "TumorTracer: a method to identify the tissue of origin from the somatic mutations of a tumor specimen". In: *BMC Med. Genomics* 8. DOI: 10.1186/s12920-015-0130-0. URL: https://doi.org/10.1186/s12920-015-0130-0.

Martincorena, Iñigo et al. (2015). "High burden and pervasive positive selection of somatic mutations in normal human skin". In: *Science* 348.6237. (Visited on 01/21/2017).

Martincorena, Iñigo et al. (Oct. 18, 2017). "Universal Patterns of Selection in Cancer and Somatic Tissues." In: *Cell* 171.5. Publisher: Elsevier, 1029–1041.e21. ISSN: 1097-4172. DOI: 10.1016/j.cell.2017.09.042. URL: http://www.ncbi.nlm.nih.gov/pubmed/29056346 (visited on 11/17/2017).

Matos, Leandro Luongo de et al. (Feb. 9, 2010). "Immunohistochemistry as an Important Tool in Biomarkers Detection and Clinical Practice". In: *Biomarker Insights* 5, pp. 9–20. ISSN: 1177-2719. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2832341/ (visited on 09/08/2020).

McKenna, Aaron et al. (Sept. 2010). "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data". In: *Genome Research* 20.9, pp. 1297–1303. ISSN: 1549-5469. DOI: 10.1101/gr.107524.110.

Mocellin, Simone et al. (May 1, 2003). "Quantitative real-time PCR: a powerful ally in cancer research". In: *Trends in Molecular Medicine* 9.5, pp. 189–195. ISSN: 1471-4914. DOI: 10.1016/S1471-4914(03)00047-9. URL: http://www.sciencedirect.com/science/article/pii/S1471491403000479 (visited on 09/20/2020).

Moffitt, Richard A et al. (Sept. 7, 2015). "Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma". In: *Nature Genetics* 47.10. Publisher: Nature Research, pp. 1168–1178. ISSN: 1061-4036. DOI: 10.1038/ng.3398. URL: http://www.nature.com/doifinder/10.1038/ng.3398 (visited on 10/12/2017).

Monzon, Federico A. and Tracie J. Koen (Feb. 2010). "Diagnosis of metastatic neoplasms: molecular approaches for identification of tissue of origin". In: *Archives of Pathology & Laboratory Medicine* 134.2, pp. 216–224. ISSN: 1543-2165. DOI: 10.1043/1543-2165-134.2.216.

Moran, C. A. et al. (May 15, 1994). "Benign and malignant salivary gland-type mixed tumors of the lung. Clinicopathologic and immunohistochemical study of eight cases". In: *Cancer* 73.10, pp. 2481–2490. ISSN: 0008-543X. DOI: 10.1002/1097-0142(19940515)73:10<2481::aid-cncr2820731006>3.0.co;2-a.

Morgan, Graeme, Robyn Ward, and Michael Barton (Dec. 1, 2004). "The contribution of cytotoxic chemotherapy to 5-year survival in adult malignancies". In: *Clinical Oncology* 16.8, pp. 549–560. ISSN: 0936-6555. DOI: 10.1016/j.clon.2004.06.007. URL: http://www.sciencedirect.com/science/article/pii/S0936655504002225 (visited on 11/21/2020).

Neal, Radford M. (1996). *Bayesian Learning for Neural Networks*. Red. by P. Bickel et al. Vol. 118. Lecture Notes in Statistics. New York, NY: Springer New York. ISBN: 978-0-387-94724-2 978-1-4612-0745-0. DOI: 10.1007/978-1-4612-0745-0. URL: http://link.springer.com/10.1007/978-1-4612-0745-0 (visited on 11/05/2020).

Nik-Zainal, Serena et al. (May 25, 2012a). "Mutational Processes Molding the Genomes of 21 Breast Cancers". In: *Cell* 149.5, pp. 979–993. ISSN: 0092-8674. DOI: 10.1016/j.cell.2012.04.024. URL: http://www.sciencedirect.com/science/article/pii/S0092867412005284 (visited on 11/04/2020).

Nik-Zainal, Serena et al. (May 25, 2012b). "The life history of 21 breast cancers." In: *Cell* 149.5. Publisher: Elsevier, pp. 994–1007. ISSN: 1097-4172. DOI: 10.1016/j.cell.2012.04.023. URL: http://www.ncbi.nlm.nih.gov/pubmed/22608083 (visited on 01/03/2017).

Orakpoghenor, Ochuko et al. (2018). "A Short Review of Immunochemistry". In: 3.1, p. 6.

Paemel, Ruben Van et al. (July 14, 2020). "Minimally invasive classification of paediatric solid tumours using reduced representation bisulphite sequencing of cell-free DNA: a proof-of-principle study". In:

Epigenetics 0.0, pp. 1–13. ISSN: 1559-2294. DOI: 10.1080/15592294.2020.1790950. URL: https://doi.org/10.1080/15592294.2020.1790950 (visited on 07/18/2020).

Painter, J. T., N. P. Clayton, and R. A. Herbert (Jan. 2010). “Useful immunohistochemical markers of tumor differentiation”. In: *Toxicologic Pathology* 38.1, pp. 131–141. ISSN: 1533-1601. DOI: 10.1177/0192623309356449.

Parsons, D. Williams et al. (Sept. 26, 2008). “An Integrated Genomic Analysis of Human Glioblastoma Multiforme”. In: *Science* 321.5897. Publisher: American Association for the Advancement of Science Section: Research Article, pp. 1807–1812. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1164382. URL: http://science.sciencemag.org/content/321/5897/1807 (visited on 11/04/2020).

Paszke, Adam et al. (2019). “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Patch, A. . M. (2015). “Whole-genome characterization of chemoresistant ovarian cancer”. In: *Nature* 521. DOI: 10.1038/nature14410. URL: https://doi.org/10.1038/nature14410.

Pavlidis, N., H. Khaled, and R. Gaafar (2015). “A mini review on cancer of unknown primary site: a clinical puzzle for the oncologists”. In: *J. Advert. Res* 6. DOI: 10.1016/j.jare.2014.11.007. URL: https://doi.org/10.1016/j.jare.2014.11.007.

Pavlidis, N. et al. (Sept. 2003). “Diagnostic and therapeutic management of cancer of an unknown primary”. In: *European Journal of Cancer (Oxford, England: 1990)* 39.14, pp. 1990–2005. ISSN: 0959-8049. DOI: 10.1016/s0959-8049(03)00547-1.

Pavlidis, Nicholas and George Pentheroudakis (Apr. 14, 2012). “Cancer of unknown primary site”. In: *The Lancet* 379.9824, pp. 1428–1435. ISSN: 0140-6736, 1474-547X. DOI: 10.1016/S0140-6736(11)61178-1. URL: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(11)61178-1/abstract (visited on 09/23/2020).

Pedersen, Magnus Erik Hvass (Nov. 7, 2020). *Hvass-Labs/TensorFlow-Tutorials*. GitHub. URL: https://github.com/Hvass-Labs/TensorFlow-Tutorials (visited on 11/07/2020).

Penson, Alexander et al. (Nov. 14, 2019). “Development of Genome-Derived Tumor Type Prediction to Inform Clinical Cancer Care”. In: *JAMA oncology*. ISSN: 2374-2445. DOI: 10.1001/jamaoncol.2019.3985.

Pfeifer, G. P. (2006). *Mutagenesis at Methylated CpG Sequences*. Ed. by Walter Doerfler and Petra Böhm. Current Topics in Microbiology and Immunology. Berlin, Heidelberg: Springer, pp. 259–281. ISBN: 978-3-540-31390-8. DOI: 10.1007/3-540-31390-7_10. URL: https://doi.org/10.1007/3-540-31390-7_10 (visited on 11/27/2020).

Pich, Oriol et al. (Nov. 18, 2019). “The mutational footprints of cancer therapies”. In: *Nature Genetics*, pp. 1–9. ISSN: 1546-1718. DOI: 10.1038/s41588-019-0525-5. URL: https://www.nature.com/articles/s41588-019-0525-5 (visited on 11/25/2019).

Polak, Paz et al. (Jan. 2014). “Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair”. In: *Nature Biotechnology* 32.1, pp. 71–75. ISSN: 1546-1696. DOI: 10.1038/nbt.2778. URL: https://www.nature.com/articles/nbt.2778 (visited on 09/14/2020).

Polak, Paz et al. (Feb. 18, 2015). “Cell-of-origin chromatin organization shapes the mutational landscape of cancer”. In: *Nature* 518.7539. Publisher: Nature Research, pp. 360–364. ISSN: 0028-0836. DOI:

10.1038/nature14221. URL: http://www.nature.com/doifinder/10.1038/nature14221 (visited on 01/03/2017).

Polak, Paz et al. (Oct. 2017). “A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer”. In: *Nature genetics* 49.10, pp. 1476–1486. ISSN: 1061-4036. DOI: 10.1038/ng.3934. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7376751/ (visited on 11/04/2020).

Pollock, P. M. and P. S. Meltzer (2002). “A genome-based strategy uncovers frequent BRAF mutations in melanoma”. In: *Cancer Cell* 2. DOI: 10.1016/S1535-6108(02)00089-2. URL: https://doi.org/10.1016/S1535-6108(02)00089-2.

Prentice, Leah M. et al. (Apr. 26, 2018). “Formalin fixation increases deamination mutation signature but should not lead to false positive mutations in clinical practice”. In: *PLoS ONE* 13.4. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0196434. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5919577/ (visited on 11/03/2020).

Priestley, Peter et al. (Jan. 16, 2019). “Pan-cancer whole genome analyses of metastatic solid tumors”. In: *bioRxiv*. Publisher: Cold Spring Harbor Laboratory, p. 415133. DOI: 10.1101/415133. URL: https://www.biorxiv.org/content/10.1101/415133v2 (visited on 05/16/2019).

Pritykin, Yuri et al. (July 26, 2020). “A unified atlas of CD8 T cell dysfunctional states in cancer and infection”. In: *bioRxiv*, p. 2020.07.25.220673. DOI: 10.1101/2020.07.25.220673. URL: https://www.biorxiv.org/content/10.1101/2020.07.25.220673v1 (visited on 07/26/2020).

Puente, Xose S. et al. (July 2011). “Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia”. In: *Nature* 475.7354. Number: 7354 Publisher: Nature Publishing Group, pp. 101–105. ISSN: 1476-4687. DOI: 10.1038/nature10113. URL: http://www.nature.com/articles/nature10113 (visited on 11/04/2020).

Rassy, Elie, Tarek Assi, and Nicholas Pavlidis (Apr. 2020). “Exploring the biological hallmarks of cancer of unknown primary: where do we stand today?” In: *British Journal of Cancer* 122.8, pp. 1124–1132. ISSN: 1532-1827. DOI: 10.1038/s41416-019-0723-z. URL: https://www.nature.com/articles/s41416-019-0723-z (visited on 09/15/2020).

Reyna, Matthew A. et al. (Feb. 5, 2020). “Pathway and network analysis of more than 2500 whole cancer genomes”. In: *Nature Communications* 11.1. Number: 1 Publisher: Nature Publishing Group, p. 729. ISSN: 2041-1723. DOI: 10.1038/s41467-020-14367-0. URL: http://www.nature.com/articles/s41467-020-14367-0 (visited on 11/05/2020).

Rheinbay, Esther et al. (2020). “Analyses of non-coding somatic drivers in 2,658 cancer whole genomes”. In: *Nature* 578.7793, pp. 102–111. ISSN: 0028-0836. DOI: 10.1038/s41586-020-1965-x. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7054214/ (visited on 11/03/2020).

Rodin, Sergei N. and Andrei S. Rodin (Apr. 2005). “Origins and selection of p53 mutations in lung carcinogenesis”. In: *Seminars in Cancer Biology* 15.2, pp. 103–112. ISSN: 1044-579X. DOI: 10.1016/j.semcancer.2004.08.005.

Rosai, J. and L. V. Ackerman (Feb. 1979). “The pathology of tumors. Part II: Diagnostic techniques”. In: *CA: a cancer journal for clinicians* 29.1, pp. 22–39. ISSN: 0007-9235. DOI: 10.3322/canjclin.29.1.22.

Ruder, Sebastian (June 15, 2017). “An overview of gradient descent optimization algorithms”. In: *arXiv:1609.04747 [cs]*. arXiv: 1609.04747. URL: http://arxiv.org/abs/1609.04747 (visited on 11/27/2020).

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams (Oct. 1986). “Learning representations by back-propagating errors”. In: *Nature* 323.6088. Number: 6088 Publisher: Nature Publishing Group, pp. 533–536. ISSN: 1476-4687. DOI: 10.1038/323533a0. URL: https://www.nature.com/articles/323533a0 (visited on 11/05/2020).

Rüschoff, J. (2012). “Adenocarcinoma of the GEJ: gastric or oesophageal cancer?” In: *Recent Results in Cancer Research. Fortschritte Der Krebsforschung. Progres Dans Les Recherches Sur Le Cancer* 196, pp. 107–113. ISSN: 0080-0015. DOI: 10.1007/978-3-642-31629-6_7.

Sabarinathan, Radhakrishnan et al. (Apr. 14, 2016). “Nucleotide excision repair is impaired by binding of transcription factors to DNA”. In: *Nature* 532.7598, pp. 264–267. ISSN: 0028-0836. DOI: 10.1038/nature17661. URL: http://www.ncbi.nlm.nih.gov/pubmed/27075101 (visited on 05/16/2019).

Sabarinathan, Radhakrishnan et al. (Sept. 20, 2017). “The whole-genome panorama of cancer drivers”. In: *bioRxiv*. Publisher: Cold Spring Harbor Laboratory Section: New Results, p. 190330. DOI: 10.1101/190330. URL: https://www.biorxiv.org/content/10.1101/190330v1 (visited on 11/05/2020).

Salvadores, Marina, David Mas-Ponte, and Fran Supek (Apr. 15, 2019). “Passenger mutations accurately classify human tumors”. In: *PLOS Computational Biology* 15.4. Publisher: Public Library of Science, e1006953. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1006953. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006953 (visited on 11/05/2020).

Sample, Ian (July 3, 2018). “Sample, Ian. Routine DNA tests will put NHS at the ‘forefront of medicine’. The Guardian (3 July 2018).” In:

Schuster-Böckler, Benjamin and Ben Lehner (Aug. 2012). “Chromatin organization is a major influence on regional mutation rates in human cancer cells”. In: *Nature* 488.7412, pp. 504–507. ISSN: 1476-4687. DOI: 10.1038/nature11273. URL: https://www.nature.com/articles/nature11273 (visited on 09/14/2020).

Setlow, R. B. and W. L. Carrier (May 1, 1966). “Pyrimidine dimers in ultraviolet-irradiated DNA’s”. In: *Journal of Molecular Biology* 17.1, pp. 237–254. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(66)80105-5. URL: http://www.sciencedirect.com/science/article/pii/S0022283666801055 (visited on 11/04/2020).

Shaw, Alice T. et al. (Oct. 2011). “Effect of crizotinib on overall survival in patients with advanced non-small-cell lung cancer harbouring ALK gene rearrangement: a retrospective analysis”. In: *The Lancet. Oncology* 12.11, pp. 1004–1012. ISSN: 1474-5488. DOI: 10.1016/S1470-2045(11)70232-7.

Sheahan, K. (1993). “Metastatic adenocarcinoma of an unknown primary site: a comparison of the relative contributions of morphology, minimal essential clinical data and CEA immunostaining status”. In: *Am. J. Clin. Pathol.* 99. DOI: 10.1093/ajcp/99.6.729. URL: https://doi.org/10.1093/ajcp/99.6.729.

Shen, S. Y. (2018). “Sensitive tumour detection and classification using plasma cell-free DNA methylomes”. In: *Nature* 563. DOI: 10.1038/s41586-018-0703-0. URL: https://doi.org/10.1038/s41586-018-0703-0.

Siegel, Rebecca L., Kimberly D. Miller, and Ahmedin Jemal (2020). “Cancer statistics, 2020”. In: *CA: a cancer journal for clinicians* 70.1, pp. 7–30. ISSN: 1542-4863. DOI: 10.3322/caac.21590.

Sikandar, Bushra et al. (2017). “Increased Tumour Infiltration of CD4+ and CD8+ T-Lymphocytes in Patients with Triple Negative Breast Cancer Suggests Susceptibility to Immune Therapy”. In: *Asian Pacific Journal of Cancer Prevention : APJCP* 18.7, pp. 1827–1832. ISSN: 1513-7368. DOI:

10.22034/APJCP.2017.18.7.1827. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5648386/ (visited on 11/06/2020).

Snoek, Jasper, Hugo Larochelle, and Ryan P. Adams (2012). "Practical Bayesian Optimization of Machine Learning Algorithms". In: *Advances in Neural Information Processing Systems* 25, pp. 2951–2959. URL: https://papers.nips.cc/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html (visited on 11/05/2020).

Soh, K. P. et al. (2017). "Predicting cancer type from tumour DNA signatures". In: *Genome Med.* 9. DOI: 10.1186/s13073-017-0493-2. URL: https://doi.org/10.1186/s13073-017-0493-2.

Sokolenko, Anna P. and Evgeny N. Imyanitov (Aug. 27, 2018). "Molecular Diagnostics in Clinical Oncology". In: *Frontiers in Molecular Biosciences* 5. ISSN: 2296-889X. DOI: 10.3389/fmolb.2018.00076. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6119963/ (visited on 08/21/2020).

Sorscher, Steven M. and Frank Anthony Greco (May 10, 2012). "Papillary Renal Carcinoma Presenting as a Cancer of Unknown Primary (CUP) and Diagnosed through Gene Expression Profiling". In: *Case Reports in Oncology* 5.2, pp. 229–232. ISSN: 1662-6575. DOI: 10.1159/000339130. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3369261/ (visited on 11/04/2020).

Sparano, Joseph A. et al. (July 12, 2018). "Adjuvant Chemotherapy Guided by a 21-Gene Expression Assay in Breast Cancer". In: *New England Journal of Medicine* 379.2. Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMoa1804710, pp. 111–121. ISSN: 0028-4793. DOI: 10.1056/NEJMoa1804710. URL: https://doi.org/10.1056/NEJMoa1804710 (visited on 11/03/2020).

Srivastava, N. et al. (2014). "Dropout: a simple way to prevent neural networks from overfitting". In: *J. Mach. Learn. Res.* 15.

Stadler, Zsofia Kinga et al. (May 20, 2020). "Targeted therapy based on germline analysis of tumor-normal sequencing (MSK-IMPACT) in a pan-cancer population." In: *Journal of Clinical Oncology* 38.15. Publisher: American Society of Clinical Oncology, pp. 1500–1500. ISSN: 0732-183X. DOI: 10.1200/JCO.2020.38.15_suppl.1500. URL: https://ascopubs.org/doi/abs/10.1200/JCO.2020.38.15_suppl.1500 (visited on 11/02/2020).

Stamatoyannopoulos, John A. et al. (Apr. 2009). "Human mutation rate associated with DNA replication timing". In: *Nature Genetics* 41.4, pp. 393–395. ISSN: 1546-1718. DOI: 10.1038/ng.363. URL: https://www.nature.com/articles/ng.363 (visited on 09/14/2020).

Su, Andrew I. et al. (Oct. 15, 2001). "Molecular Classification of Human Carcinomas by Use of Gene Expression Signatures". In: *Cancer Research* 61.20, pp. 7388–7393. ISSN: 0008-5472, 1538-7445. URL: https://cancerres.aacrjournals.org/content/61/20/7388 (visited on 09/20/2020).

Supek, Fran and Ben Lehner (May 2015). "Differential DNA mismatch repair underlies mutation rate variation across the human genome". In: *Nature* 521.7550, pp. 81–84. ISSN: 1476-4687. DOI: 10.1038/nature14173. URL: https://www.nature.com/articles/nature14173 (visited on 09/14/2020).

Swanton, Charles et al. (July 2015). "APOBEC enzymes: mutagenic fuel for cancer evolution and heterogeneity". In: *Cancer discovery* 5.7, pp. 704–712. ISSN: 2159-8274. DOI: 10.1158/2159-8290.CD-15-0344. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4497973/ (visited on 11/04/2020).

Sève, Pascal et al. (Jan. 15, 2007). "The role of 2-deoxy-2-[F-18]fluoro-D-glucose positron emission tomography in disseminated carcinoma of unknown primary site". In: *Cancer* 109.2, pp. 292–299. ISSN: 0008-543X. DOI: 10.1002/cncr.22410.

Tomkova, Marketa and Benjamin Schuster-Böckler (Aug. 1, 2018). "DNA Modifications: Naturally More Error Prone?" In: *Trends in Genetics* 34.8, pp. 627–638. ISSN: 0168-9525. DOI: 10.1016/j.tig.2018.04.005. URL: http://www.sciencedirect.com/science/article/pii/S0168952518300817 (visited on 11/04/2020).

Torrey, Lisa and Jude Shavlik. *Transfer Learning.*

Tothill, R. W. (2013). "Massively-parallel sequencing assists the diagnosis and guided treatment of cancers of unknown primary". In: *J. Pathol.* 231. DOI: 10.1002/path.4251. URL: https://doi.org/10.1002/path.4251.

Travis, Glenn S (May 19, 2011). "Field guide to next-generation DNA sequencers - GLENN - 2011 - Molecular Ecology Resources - Wiley Online Library". In: URL: https://onlinelibrary-wiley-com.myaccess.library.utoronto.ca/doi/full/10.1111/j.1755-0998.2011.03024.x (visited on 11/22/2020).

Travis, Lois B. (Nov. 2006). "The epidemiology of second primary cancers". In: *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* 15.11, pp. 2020–2026. ISSN: 1055-9965. DOI: 10.1158/1055-9965.EPI-06-0414.

Tsimberidou, Apostolia M. et al. (June 1, 2020). "Review of precision cancer medicine: Evolution of the treatment paradigm". In: *Cancer Treatment Reviews* 86. ISSN: 0305-7372, 1532-1967. DOI: 10.1016/j.ctrv.2020.102019. URL: https://www.cancertreatmentreviews.com/article/S0305-7372(20)30057-8/abstract (visited on 09/09/2020).

Uzilov, Andrew V. et al. (Dec. 2016). "Development and clinical application of an integrative genomic approach to personalized cancer therapy". In: *Genome Medicine* 8.1. Number: 1 Publisher: BioMed Central, pp. 1–20. ISSN: 1756-994X. DOI: 10.1186/s13073-016-0313-0. URL: http://genomemedicine.biomedcentral.com/articles/10.1186/s13073-016-0313-0 (visited on 11/03/2020).

Vaicenavicius, Juozas et al. (Feb. 19, 2019). "Evaluating model calibration in classification". In: *arXiv:1902.06977 [cs, stat]*. arXiv: 1902.06977. URL: http://arxiv.org/abs/1902.06977 (visited on 09/27/2020).

Van Rossum, Guido and Fred L. Drake (2009). *Python 3 Reference Manual.* Scotts Valley, CA: CreateSpace. ISBN: 1441412697.

Varadhachary, Gauri R (Nov. 2007). "Carcinoma of unknown primary origin." In: *Gastrointestinal cancer research : GCR* 1.6. Publisher: International Society of Gastrointestinal Oncology, pp. 229–35. ISSN: 1934-7820. URL: http://www.ncbi.nlm.nih.gov/pubmed/19262901 (visited on 08/18/2017).

Varadhachary, Gauri R. et al. (Sept. 20, 2008). "Molecular Profiling of Carcinoma of Unknown Primary and Correlation With Clinical Evaluation". In: *Journal of Clinical Oncology* 26.27, pp. 4442–4448. ISSN: 0732-183X. DOI: 10.1200/JCO.2007.14.4378. URL: https://ascopubs.org/doi/10.1200/JCO.2007.14.4378 (visited on 08/21/2020).

Varghese, A M et al. (Dec. 2017). "Clinical and molecular characterization of patients with cancer of unknown primary in the modern era". In: *Annals of Oncology* 28.12, pp. 3015–3021. ISSN: 0923-7534. DOI: 10.1093/annonc/mdx545. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5834064/ (visited on 09/06/2020).

Vaswani, Ashish et al. (Dec. 5, 2017). "Attention Is All You Need". In: *arXiv:1706.03762 [cs]*. arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762 (visited on 11/05/2020).

Vodanovich, Domagoj Ante and Peter F M Choong (2018). “Soft-tissue Sarcomas”. In: *Indian Journal of Orthopaedics* 52.1, pp. 35–44. ISSN: 0019-5413. DOI: 10.4103/ortho.IJOrtho_220_17. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5791230/ (visited on 11/02/2020).

Vogt, Alexia et al. (May 2, 2017). “Multiple primary tumours: challenges and approaches, a review”. In: *ESMO Open* 2.2. ISSN: 2059-7029. DOI: 10.1136/esmoopen-2017-000172. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5519797/ (visited on 09/21/2020).

Volkova, Nadezda V. et al. (May 1, 2020). “Mutational signatures are jointly shaped by DNA damage and repair”. In: *Nature Communications* 11.1, pp. 1–15. ISSN: 2041-1723. DOI: 10.1038/s41467-020-15912-7. URL: https://www.nature.com/articles/s41467-020-15912-7 (visited on 05/04/2020).

Volta, Alberto Dalla et al. (Dec. 17, 2018). “Transformation of Prostate Adenocarcinoma Into Small-Cell Neuroendocrine Cancer Under Androgen Deprivation Therapy: Much Is Achieved But More Information Is Needed”. In: *Journal of Clinical Oncology* 37.4. Publisher: American Society of Clinical Oncology, pp. 350–351. ISSN: 0732-183X. DOI: 10.1200/JCO.18.01055. URL: https://ascopubs.org/doi/full/10.1200/JCO.18.01055 (visited on 11/02/2020).

Waters, Lauren S. et al. (Mar. 2009). “Eukaryotic Translesion Polymerases and Their Roles and Regulation in DNA Damage Tolerance”. In: *Microbiology and Molecular Biology Reviews : MMBR* 73.1, pp. 134–154. ISSN: 1092-2172. DOI: 10.1128/MMBR.00034-08. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2650891/ (visited on 11/04/2020).

Wiencke, John K (Oct. 15, 2002). “DNA adduct burden and tobacco carcinogenesis”. In: *Oncogene* 21.48, pp. 7376–7391. ISSN: 0950-9232. DOI: 10.1038/sj.onc.1205799. URL: http://www.ncbi.nlm.nih.gov/pubmed/12379880 (visited on 05/16/2019).

Wilson, Andrew Gordon and Pavel Izmailov (Apr. 27, 2020). “Bayesian Deep Learning and a Probabilistic Perspective of Generalization”. In: *arXiv:2002.08791 [cs, stat]*. arXiv: 2002.08791. URL: http://arxiv.org/abs/2002.08791 (visited on 06/02/2020).

Wu, G. (2014). “The genomic landscape of diffuse intrinsic pontine glioma and pediatric non-brainstem high-grade glioma”. In: *Nat. Genet.* 46. DOI: 10.1038/ng.2938. URL: https://doi.org/10.1038/ng.2938.

Yatabe, Yasushi et al. (Mar. 2019). “Best Practices Recommendations for Diagnostic Immunohistochemistry in Lung Cancer”. In: *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer* 14.3, pp. 377–407. ISSN: 1556-0864. DOI: 10.1016/j.jtho.2018.12.005. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6422775/ (visited on 09/08/2020).

Yin, Li et al. (June 9, 2020). “Triple-negative breast cancer molecular subtyping and treatment progress”. In: *Breast Cancer Research* 22.1, p. 61. ISSN: 1465-542X. DOI: 10.1186/s13058-020-01296-5. URL: https://doi.org/10.1186/s13058-020-01296-5 (visited on 11/02/2020).

Yoon, Hongjun et al. (2019). “Tumor Identification in Colorectal Histology Images Using a Convolutional Neural Network”. In: *Journal of Digital Imaging* 32.1, pp. 131–140. ISSN: 1618-727X. DOI: 10.1007/s10278-018-0112-9.

Yuan, Yuchen et al. (Dec. 23, 2016). “DeepGene: an advanced cancer type classifier based on deep learning and somatic point mutations”. In: *BMC Bioinformatics* 17.17, p. 476. ISSN: 1471-2105. DOI: 10.1186/s12859-016-1334-9. URL: https://doi.org/10.1186/s12859-016-1334-9 (visited on 11/05/2020).

Zapatka, Marc et al. (Mar. 2020). "The landscape of viral associations in human cancers". In: *Nature Genetics* 52.3, pp. 320–330. ISSN: 1546-1718. DOI: `10.1038/s41588-019-0558-9`. URL: `https://www.nature.com/articles/s41588-019-0558-9` (visited on 08/11/2020).

Zehir, Ahmet et al. (2017). "Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients". In: *Nature Publishing Group.* DOI: `10.1038/nm.4333`. (Visited on 09/19/2017).

Zhang, Jinghui et al. (June 2013). "Whole-genome sequencing identifies genetic alterations in pediatric low-grade gliomas". In: *Nature genetics* 45.6, pp. 602–612. ISSN: 1061-4036. DOI: `10.1038/ng.2611`. URL: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3727232/` (visited on 11/05/2020).

Zviran, Asaf et al. (July 2020). "Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring". In: *Nature Medicine* 26.7. Number: 7 Publisher: Nature Publishing Group, pp. 1114–1124. ISSN: 1546-170X. DOI: `10.1038/s41591-020-0915-3`. URL: `http://www.nature.com/articles/s41591-020-0915-3` (visited on 11/08/2020).

Appendix

All large data files described below are available as supplementary Excel files.

A.1 Tables related to Chapter 2

Table A.1: This table shows a summary of the tumour samples from PCAWG that were used for training and testing the classifiers.