Master's thesis

Master's Programme in Life Science Informatics

Deep Mutation Modelling in Cancer Driver Mutation and Cancer Driver Gene Detection

Katri Maljanen

May 27, 2021

Faculty of Science

University of Helsinki

Supervisor(s)

Esa Pitkänen, Ph.D., Docent

Examiner(s)

Prof. Ville Mustonen, Esa Pitkänen, Ph.D., Docent

Contact information

P. O. Box 68 (Pietari Kalmin katu 5)
00014 University of Helsinki,Finland

HELSINGIN YLIOPISTO – HELSINGFORS UNIVERSITET – UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Koulutusohjelma — Utbildningsprogram — Study programme
Faculty of Science	Master's Programme in Life Science Informatics

Tekijä — Författare — Author

Katri Maljanen

Työn nimi — Arbetets titel — Title

Deep Mutation Modelling in Cancer Driver Mutation and Cancer Driver Gene Detection

Ohjaajat — Handledare — Supervisors

Esa Pitkänen, Ph.D., Docent

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Master's thesis	May 27, 2021	80 pages

Tiivistelmä — Referat — Abstract

Cancer is a leading cause of death worldwide. Unlike its name would suggest, cancer is not a single disease. It is a group of diseases that arises from the expansion of a somatic cell clone. This expansion is thought to be a result of mutations that confer a selective advantage to the cell clone. These mutations that are advantageous to cells that result in their proliferation and escape of normal cell constraints are called driver mutations. The genes that contain driver mutations are known as driver genes. Studying these mutations and genes is important for understanding how cancer forms and evolves.

Various methods have been developed that can discover these mutations and genes. This thesis focuses on a method called Deep Mutation Modelling, a deep learning based approach to predicting the probability of mutations. Deep Mutation Modelling's output probabilities offer the possibility of creating sample and cancer type specific probability scores for mutations that reflect the pathogenicity of the mutations. Most methods in the past have made scores that are the same for all cancer types. Deep Mutation Modelling offers the opportunity to make a more personalised score.

The main objectives of this thesis were to examine the Deep Mutation Modelling output as it was unknown what kind of features it has, see how the output compares against other scoring methods and how the probabilities work in mutation hotspots. Lastly, could the probabilities be used in a common driver gene discovery method. Overall, the goal was to see if Deep Mutation Modelling works and if it is competitive with other known methods.

The findings indicate that Deep Mutation Modelling works in predicting driver mutations, but that it does not have sufficient power to do this reliably and requires further improvements.

Avainsanat — Nyckelord — Keywords

cancer, driver mutations, driver genes, deep learning

Säilytyspaikka — Förvaringsställe — Where deposited

Helsinki University Library

Muita tietoja — övriga uppgifter — Additional information

Algorithmic bioinformatics specialisation line

Contents

1 Introduction

Cancer is one of the leading causes of death in the world and its incidence is expected to grow [53]. In 2020 cancer caused nearly 10 million deaths in the world and further 19.3 million cases were diagnosed [49]. In 2018 cancer caused 12,730 deaths in Finland. In Finland it is estimated that 36% of women and 38% of men will get cancer during their lifetimes [37]. Cancer however is not a single disease but rather a group of diseases with a similar origin of uncontrolled expansion of a somatic cell clone. The somatic cell clone starts to proliferate when it gains mutations that confer a selective advantage to this cell clone over other cells in the tissue and the cell clone manages to evade normal immune surveillance. The mutations that confer the selective advantage are known as driver mutations and genes which contain these driver mutations are known as driver genes. As cancer cells accumulate numerous mutations not all of them can be driver mutations and some of them are instead passenger mutations that are selectively neutral. [29]. It is currently thought that all cancers originate from cells that have driver mutations in them. However, it is currently unclear how many driver mutations there are and how many are needed for tumorigenesis [26]. Currently it is thought that there are relatively few recurrently mutated driver mutations and many more much rarer driver mutations [10]. Additional questions arise from the specificity of driver mutations in different tumour types. Some tumours have recurrent mutation patterns while others have more interchangeable mutations in their key signalling pathways [14].

There have in the past two decades been several large-scale projects which have sequenced genomes from tumours. Two of such projects are the Pan-Cancer Analysis of Whole Genomes (PCAWG) [53] and The Cancer Genome Atlas (TCGA) [33]. The PCAWG project is an international collaborative project intended to discover common patterns of mutations from over 2,600 different tumour samples. The data in PCAWG project is based on International Cancer Genome Consortium whole genomes [53]. The TCGA project is a collaborative cancer genomics project between the National Cancer Institute and National Human Genome Research Institute. The program has produced different publicly available data on cancer including exome, the coding regions of the genome, sequence data and data on the somatic mutations present in the tumour samples [33].

There have been multiple different methods that have been used to identify driver muta-

tions and driver genes. In these studies, if there were mutations that had higher recurrence than expected they were generally classified as driver mutations [39]. The occurrence of driver mutations should differ from the background mutation rate as driver mutations are under positive selection while the background mutation rate should reflect the conditions under neutrality. For example, recurrence of a mutation in multiple tumour samples in a sequenced cohort can be a sing of positive selection [29].

Recently there has also been interest in using deep learning methods to identify driver mutations [24, 2]. Deep learning is a sub-field of machine learning that focuses on representation learning, where a system can be given raw data and it can automatically discover important features from this raw data on its own. Traditional machine learning methods are at a disadvantage compared to deep learning in that they have limited capabilities of processing the raw data and require human interference for the discovery of important features. The discovery of the important features by humans, however, requires significant domain knowledge and may still miss features that could be discovered from the data. Representation learning works by transforming the input with non-linear functions into more abstract representations that can be combined to classify data. In image data representation learning can for example mean that the system learns that there are edges in the image and the edges form shapes when they are combined. This allows the system to learn complex features and allows for the suppression of unimportant information [20].

Supervised learning is the most common form of machine learning and this applied to deep learning too. The other common type of machine learning is unsupervised learning. In supervised learning the model uses input which consists of either features or raw data, and a target value. Supervised learning is mostly used for classification tasks where the model tries to predict the target values. In unsupervised learning the model does not use the target values and instead it tries to group together similar inputs based on their features. Supervised learning is useful for classification and unsupervised learning is useful for clustering and discovering groups within the data [20]. Both supervised [2] and unsupervised learning have been used previously in the discovery of driver mutations and genes [19].

This thesis focuses on the exploration of the output, and potential use, of a deep learning method for modelling mutation probabilities called Deep Mutation Modelling (DMM) [8]. DMM is a deep learning based method that uses raw DNA sequences, which contain mutations, and sample information as its input. As it is a deep learning based method it should learn the genomic contexts of the mutations. It attempts to predict which mutation has occurred in the input sequence. The predicted probabilities of mutations form the

output. The output of DMM is potentially useful in the discovery of driver mutations and genes as it could be used as a scoring system for mutations [8]. These types of different scores have been developed for mutations before and they generally measure either the functional impact of the mutations or the possible pathogenicity level of them [46] [42]. Combined Annotation Dependent Depletion (CADD) [42] is one such commonly used tool which scores the deleteriousness of single nucleotide variants. CADD itself is based on the annotation differences between simulated mutations and alleles which have been fixed in the human genome since the human-chimpanzee divergence [42]. These types of scores are used to rank the mutations according to their potential pathogenicity or they can be used in other methods in aiding the discovery of driver genes [32]. However, these scores are usually the same for every cancer type and patient. DMM's output on the other hand is tumour sample specific and thus could be more flexible than existing scoring systems. This would allow for a more personalised medicine approach in discovering which mutations or genes are significant in an individual tumour sample. As the outputs of DMM are sample specific they can also be combined by cancer type or by some other sample feature to form specific scores for the feature in question.

This work had several objectives. The first one was to explore the basic features of the DMM output, the second was to compare the DMM output against other known scoring systems for mutations, the third was to see how the DMM output works with hotspot mutations, and lastly to see if the DMM outputs can be used in a previously established method of driver gene discovery as the scoring system for mutations. The main objective was to see if DMM works as intended and to see if it is competitive with other methods that have been used before.

I first explored single nucleotide variant datasets from the PCAWG and TCGA projects and the outputs that DMM produced for them. After this I compared other functional impact and pathogenicity score datasets with DMM for the PCAWG and TCGA mutations. The other score datasets used contained this information for non-synonymous, or synonymous, or both types of mutation depending on the dataset in question.

After exploration of the DMM output, I tested if the output could be used in a driver gene detection method known as OncodriveFML [32]. OncodriveFML computes how significant a gene is based on its observed mutation distribution versus a generated distribution. For this OncodriveFML needs a score file from where it can extract scores for the computations. [32] I used the DMM output to construct score files for four cancer types from the PCAWG and TCGA datasets. I compared these with the CADD score, which is the same for every

tissue type. In the analysis I used four known cancer genes *APC, BRAF, KRAS*, and *TP53* and six other genes.

The first chapters of this thesis focus on giving an overview of natural selection in cancer and background mutation rates in cells. After this the thesis covers the types of driver mutations present in cancer genomes and how these driver mutations and driver genes have been discovered in the past. The thesis will then move on to the experimental section which contains exploration of the DMM model's output and comparison to other scoring schemes that can be used to assess the impact of mutations. The experimental section also contains an experiment done by using the DMM output in OncodriveFML.

2 Background

2.1 Basic types of mutations

Cancer cells gain mutations stochastically through mutagens or DNA replication errors. Most of these mutations are passenger mutations which generally do not change the functionality of the cell. A small minority of the mutations give an advantage to the cells which allow them to escape the growth restraints on somatic cells. These cancer cells then end up proliferating. The mutations that provide a selective advantage to the cancer cells are known as driver mutations [53]. Mutations can be classified also into germline and somatic mutations. Germline mutations can be passed onto the next generation and they affect every cell in an individual's body causing a change in the genotype. Somatic refers to the non-germline cells that will also accumulate mutations over time, but they cannot be passed onto the next generation within the species [25].

Cancer causing genes can generally be classified into two groups: oncogenes and tumour suppressor genes. Oncogenes generally drive cancer by a gain of function mutations where the expression of the oncogene is increased or the protein it codes for changes into one that can cause cancer. Tumour suppressor genes tend to work in the opposite way where the cancer-causing mutations tend to inactivate the gene either by changing the regulatory regions or by making the protein product non-functional [5].

Mutations in the coding region of the genome, which makes up approximately 2% of the human genome [1], can alter the amino acid chain the genes code. The effects of non-coding mutations are harder to study as they do not alter the protein structure which makes their effects harder to detect. Non-coding mutations may affect the regulatory elements and processes present in the genome. Studying these mutations is difficult due to the still lacking knowledge on the functionality of the non-coding regions [43].

Single nucleotide variants (SNV) refer to variants of a single nucleotide. Single nucleotide variants can for example be single nucleotide substitutions [48]. Nucleotide substitutions can be classified into transversion and transitions. Transversions happen when substitutions result in a pyrimidine base, C and T, being replaced by a purine base, A and G, or vice versa. Transitions happen when a purine base is replaced by a different purine base

or when a pyrimidine base is replaced by another pyrimidine base [12]. Within the exome these mutations may act as missense, synonymous, or nonsense mutations. Missense mutations change the amino acid the codon triplet codes to another one. These mutations may also not result in a change in the amino acid in which case the mutations are known as synonymous mutations. Non-synonymous mutations can also cause the formation of an early stop codon truncating the protein. These truncating mutations are known as nonsense mutations [48]. Mutations may also cause a change in the protein if they cause a disruption in the splice sites near the exon-intron boundaries [50]. Within the non-coding genome SNVs can alter the binding motifs of different molecules or they may create new binding motifs [40]. A similar concept to SNVs is single nucleotide polymorphism (SNP) which refers to there being at least two alternative bases present in a population in a single locus in the genome [48].

Short insertions or deletions of nucleotides are called indels. As with substitutions indels can cause the protein structure to alter because of amino acid change or due to truncating of the sequence. Within the coding regions an indel, which length is not a multiple of three, may also cause a frameshift in the sequence altering the sequence following the mutation. A frameshift causes all following codon triplets to change. If the indel length is a multiple of three then it causes either the insertion of new amino acids or the deletion of them. Larger changes in the genome are called either copy number variants, variation between individuals in the number of over 50, base pairs long DNA sequences, or structural variants, changes that are longer than 50 base pairs. Inversions or translocations of DNA are, for example, examples of structural variants [48].

2.2 Natural selection in cancer

Natural selection is the driving force behind evolution. While natural selection is usually applied on species it does also apply to cancer cells. As such cancer follows Darwinian model of evolution. In the formation of a tumour cancer cells compete against the environment created by normal cells and if they manage to gain a competitive edge, they will form a tumour [53]. Natural selection can be classified into positive and negative selection. Positive selection occurs when a mutation gives a selective advantage to the organism and this genotype becomes the dominant one within the population. Negative selection on the other hand removes deleterious alleles from the population, in this case the organism has fewer offspring than its competitors leading to the disappearance of the deleterious allele.

In the case of cancer positive selection will result in the cancer cell proliferating. Negative selection in cancer would lead to the death of sub-clones due to deleterious alleles [27].

A major gap in understanding cancer is that it is unknown how many mutations are necessary for cancer formation [26]. One way to do this would be to count the mutations occurring in the known driver genes, but this is difficult because not all driver genes are known and the presence of passenger mutations in these genes would make this difficult. However, the number of non-synonymous base substitution driver mutations within tumours is estimated to be withing 1 to 10 range with the average being 4. Another gap in understanding is the lack of knowledge regarding negative selection in cancer evolution and to what extent do somatic lineages die off due to effects of deleterious mutations. Detection of negative selection is important because as it may help identify new driver genes and patterns of synthetic lethality. [27] Additionally, normal mutational burden in normal somatic cells is one of the unanswered questions in cancer. Other important questions are which mutation processes are operative in normal cells, the pattern of clonal expansion caused by first drivers, and the positive selection of competing clones within an organ. It has for example, been found that known driver genes can be under positive selection in normal tissue without there being a malignant tumour [28].

Genes under negative selection are expected to have lower number of mutations than would be expected under neutrality. Similarly, a gene under positive selection in cancer would have more driver mutations than expected. One common way of studying the number of mutations under selection is to compare the number of non-synonymous mutations in genes against the number of synonymous mutations. This is known as the dN/dS rate. The assumption with this rate is that the synonymous mutations are a good proxy for the background mutation rate. [27] Under neutrality the rate should be approximately 1, while rates greater than 1 indicate positive selection and rates under 1 indicate negative selection [54]. dN/dS rates have traditionally been used in population genetics to study species evolution, but it is also possible to use it in somatic cells. The use of dN/dS rate in somatic cells however requires adjustment so that it is more sensitive to more than simple missense mutations, the different mutation rate in different genes and possible presence of germline mutations. [27]

It is estimated that normal somatic cells gain 2 to 10 new mutations in their diploid genome in every cell division and that normal cells will accumulate somewhere between a hundred and a few thousand substitutions [26]. Genome wide dN/dS rates in normal cells when looking at germline mutations show that the rate is very low. In common human

germline polymorphism, the rate is around 0.08 dN/dS. Accordingly, negative selection characterises species evolution. Cancer evolution, however, generally shows genome wide dN/dS ratios a little above 1. This applies to multiple cancer types and applies to both missense and truncating mutations. Similar rate is also found from somatic mutations in healthy tissues. This means that in cancer mutations are under slight positive selection, but overall, the selection is close to neutrality. Thus, dN/dS ratio of approximately 1 characterises evolution in healthy somatic tissue and cancer. In hypermutated tumours the dN/dS ratio converges towards 1 and the number of driver mutations increase but with diminishing numbers as the mutation burden increases as most of the mutations are still passenger mutations [27].

Healthy tissues can exhibit similar patterns of mutations and selection as cancer cells. Healthy skin that has been exposed to UV-light has the same predominance of C→T and CC→TT mutations as skin cancers do. The same applies to C→A substitutions that are often seen in cutaneous squamous cell carcinoma cancers. Known cancer associated genes can additionally show positive selection even in healthy tissue. For example, a dN/dS analysis of UV-exposed skin showed that several genes belonging to the NOTCH group had excesses of protein-altering base substitutions. NOTCH receptors play a part in stem cell regulation and are often targets of inactivating mutations in epithelial cancers and activating mutations in lymphoid malignancies. *NOTCH1* is often inactivated in both alleles in squamous cell carcinomas, but this same inactivation can also be present in healthy skin so this is not only limited to cancers. In addition *TP53* was also identified as having driver mutations and *KRAS* was identified as having known canonical hotpot mutations in it even in this healthy skin. Overall, healthy skin was estimated to have on average 0.27 driver mutations while cutaneous squamous cell carcinoma tumours would have an average of 2.7 driver mutations [28]. Driver mutations can also be found from the blood cells of 10% of individuals older than 65. This mutation pattern is typical for leukaemia and these individuals have an elevated risk of getting cancer later in life [26]. And while sub-clones in healthy tissue can carry a few driver mutations this does not necessarily lead to malignancy. As such the question remains on what events are necessary for cells to turn malignant [28].

Some reasons for why cancer cells have a high tolerance for deleterious mutations could be that having two or more copies of genes protects against the deleterious effects. Alternatively, there may be alternative cellular pathways for the same process. Weakly deleterious mutations can hitchhike with driver mutations, so they become fixed in the cancer pop-

ulation as the more positive effects of the driver mutation masks the deleterious effects [27].

2.3 Background mutation rate in normal and cancer cells

Background mutation rate refers to how often a genomic site mutates. The rate can vary between different regions within the genome. Mutation rate can be used to estimate evolutionary distance and it can be used to reconstruct phylogeny and earlier stages of evolution. As such it is an important feature in understanding the evolution of organisms. It is also helpful in understanding cancer and cancer evolution. Cancer is a disease that can be caused by multiple different loci mutating. If certain loci mutate more often then there are more opportunities for the disease to manifest. On the other hand, it is easy to mistake sites and genes that mutate often with disease causing variants even if their effects are neutral. Most of the work on mutation rate has focused on germline mutations while some work has also examined the mutation rate of somatic mutations in cancer cells [12].

Within cancer histone modifications and chromatin openness are important factors influencing the background mutation rate. Many features that affect the mutation rate in the germline also affects the rate of somatic mutations in cancer [45]. Different cancer types can have different mutation patterns and be affected by different epigenomic features leading to different mutation rates. Up to 86% of the variance in mutation rate in cancer genomes can be explained by chromatin accessibility, modification, and replication timing. The best epigenomic features for assessing the local mutation density are features from the original cell type rather than the cancer cell lines [38] .

Background mutation rate is also dependent on the cancer type, sample and even region within the sample. Different cancer types have different whole genome mutation rates. Stomach cancer can have a mutation rate of 11.4 mutations/Mbp whereas medulloblastomas can have a mutation rate of 0.01 mutations/Mbp. Stomach cancers thus can have 800 times higher mutation rate. Even within the same cancer type the mutation rate may be different between samples. In breast cancer samples the mutation rates vary between 0.36 mutations/Mbp and 21.8 mutations/Mbp [23].

There are several causes that likely explain some of the reasons for the different mutation

rates. Regions that replicate later accumulate more DNA damage resulting in more mutations. Methylation pattern in the germline in certain sites may also play a part in the mutation rate as methylated cytosines in CpG sites are often unstable and may undergo deamination and change into thymine. This results in a C→T transition and this may explain why CpG sites and non-CpG sites have a different mutation rate. CpG sites can have 30 times the transition mutation rate compared to other sites in great apes. Transversion is also more common at CpGs. CpG islands are areas where there is an excess of CG dinucleotides within approximately 1kb region. These sites are often unmethylated and thus have a lower substitution rate for CpGs compared to CpGs outside CpG islands. The stability of the methylated cytosine within a CpG seems to also increase with the CG richness of the sequence and as a result the mutation rate may be up to three times lower with these CpG compared to the rest of the genome. Cancer genomes also show evidence of this CpG effect. [12].

Mutational context is also an important factor in mutation rate within the germline and the effects vary across the human genome. The importance of adjacent nucleotides effect on the mutation rate reduces the further away they are from the mutated site [12]. Nucleotide context is also an important factor in the mutation rate of somatic mutations in cancer. In a pan-cancer analysis of known 520 cancer genes the mutability is higher for those codon substitutions that were observed within the tumour samples, mean 3.9e-6. These observed mutations accounted only 1% of the theoretically possible substitutions. The mutability of substitutions that were not observed were three times lower according to the background model used, mean 1.3e-6. Both codon and nucleotide mutability were also higher for synonymous mutations, nucleotide mutability mean 3.9e-6, compared to the mutability of missense, nucleotide mean 3.2e-6, and nonsense mutations, nucleotide mean 3.10e-6 [5]. Additionally, there is evidence for context-independent mutation rate in the germline at certain sites. It has been observed that SNPs occur more often at orthologous [12], a gene inherited from a common ancestor that has since diverged [48], sites in humans, chimpanzees, and macaques than is expected if the process was random. This same pattern is also observed with single nucleotide substitutions between closely related species. The excess of coincident SNPs and coincident single nucleotide substitutions is seen in transversions and especially A→T transversion. These sites are not associated with ancestral variation, difference in sequencing coverage, or selection and such seem to be hypermutable on their own. Such sites could be influenced by motifs some distance away from the site that may influence the mutation rate in it [12].

Mutation events are also not always independent of each other. Passenger mutations may be generated by driver mutations as a side effect. This can happen if the driver mutation happens in DNA replication or repair gene. This causes DNA to accumulate mutations that may have been otherwise repaired. Within the germline the rate of point mutations also increases near sites that have undergone insertion or deletion. The effect is strongest near 50bp on either side of the indel while the effect can still be observed over a few hundred base pairs. Even very short indels with a length of only a few base pairs have an effect but the effect of longer indels is stronger. It has been estimated that the mutation rate around indels might be sixfold in humans [12]. There are also somatic mutation processes that can generate multiple mutations in one event. Double-stranded DNA breaks can result in the rearrangement of the DNA even between chromosomes. Kataegis is a hypermutation process that results in locally clustered nucleotide substitutions. Kataegis is common in cancers and in a pan-cancer analysis of tumour genomes 60.5% of all cancers had it. Lastly chromotripsis is a process where hundreds of DNA breaks occur simultaneously in one or a couple of chromosomes and the fragments are arranged together near randomly. In the same pan-cancer analysis chromotripsis regions overlapped with 3.6% of identified driver mutations [53].

Aside from variation of mutation rate within smaller regions there is also larger scale variation within the chromosomes within the germline. Some of the differences could be due to selection but an estimated 3% of the intergenic and intronic human genome is thought to be under selection. Some causes for this variation are replication time, male recombination rate, distance from the telomeres, CG content, nuclear lamina binding sites, and simple repeat content. None of these factors explain the differences in mutation rate between regions alone, but replication time and distance from the telomeres are thought to be more influential than the others. Chromosomes also have differing mutation rates [12].

Somatic mutation rates measured from cancer genomes correlate moderately with those inferred from human-chimpanzee sequence divergence in the germline. On the mega base scale mutation rate depends strongly on chromatin organisation levels. Euchromatin and inaccessible heterochromatin especially affect the mutation rate. In cancer it is known that the highest mutation rates are in the inaccessible heterochromatin like regions and lowest in euchromatin like regions. single nucleotide variant density is positively correlated with histone H3K9me3 modification and negatively associated with features that are associated with open chromatin. In cancer cells H3K9me3 could account up to 40% of the variance in

single nucleotide variant density [45]. There are some differences in which histone marks affect which cancer type. Mutation density in hepatocytes follow H3K4me1 levels for example, while melanocyte mutation rate follows H3K4me1 levels [38].

Interestingly it seems that the epigenomic features of the original healthy cell lines are better at predicting the mutation density rather than the epigenomic features from cancer cell lines. This applies at least to liver cancer and melanoma. It is possible that this is because the somatic mutations arise before the epigenomic changes associated with tumour progression or the tumour undergoes epigenetic changes later that differ it from other tumours of the same cancer type. Thus, it is important to consider the cancer type when investigating the mutation rate within tumour genomes [38].

2.4 Driver Mutations

Driver mutations are mutations that give a growth advantage to a somatic cell clone leading to clonal expansion and tumour formation. Driver mutations can be challenging to find as not all mutations in a tumour are driver mutations and instead most are passenger mutations. A pan-cancer analysis of 2658 whole tumour genomes found that 91% of tumours had at least one driver mutation. On average the tumours had 4.6 driver mutations. On average there were 2.6 coding point mutations per tumour. In the same analysis, 13% of point mutations were non-coding, and 25% of tumours had a potential non-coding driver mutation [53]. The number of driver mutations is also dependent on the cancer type in question as the estimated number of driver mutation varies between one and ten. Endometrial cancers have an estimated little above 10 driver mutations while Thyroid cancer has approximately 1 driver mutation. Many other cancer types fall between these numbers [27].

Most studies on driver mutations have focused on protein altering driver mutation within the coding region but there have also been studies into the non-coding regions [43] and into synonymous mutations within the coding region [46]. Studying the driver mutations in these regions is however still challenging due to the poorer understanding of the background mutation processes and functionality of the non-coding genome. One known target of non-coding driver mutations is the promoter of *TERT* gene [43].

A prevailing theory on driver mutations is that there are a small number of highly recurrent driver mutations, and a larger number of rare driver mutations that are rarely recurrent. These rare driver mutations form a so called long-tail of driver mutations. These long-tail

mutations are much harder to detect as they are likely to be cancer type specific and current tumour genome datasets are possibly too small to detect them if we are detecting driver mutations by observing the number of recurrent mutations [10].

While the number of driver mutations present in the tumours is an interesting question another interesting question is how cancer type specific are the driver mutations. Different cancer types tend to accumulate mutations in different cancer genes. For example, breast cancer and melanoma tend to have mutations accumulate in tissue-specific genes whereas lung cancer has a more diffuse mutation pattern [14]. Some driver genes, such as *TP53*, are active in multiple different cancer types while other genes are much more specific. For example, *MYC* is active in Burkitt lymphomas. Individual tumours even within the same cancer type can have different mutation profiles. Breast adenocarcinomas can have mutations in hundreds of genes, or they may only have mutations in only a couple of handfuls of genes [29]. Different types of driver mutations are also present in different quantities in different tumour types. Structural variants are more common in breast adenocarcinomas, mean 6.4 structural variants versus mean 2.2 point mutations, whereas point mutations are more common in colorectal adenocarcinomas, mean 2.4 structural variants versus mean 7.4 point mutations [53].

Driver mutations have not been identified in all tumours. The reason for this could be either technical or biological. Technical issues could arise from poor quality samples, sequencing errors or failures in the algorithms used in detecting driver mutations. Biological causes could include yet unknown driver mutations. Chromophobe renal cell carcinoma and pancreatic neuroendocrine cancers have been identified as cancer types with high fractions of patients with no detected driver mutations. A total of 5.3% tumours had no known driver mutations in a pan-cancer analysis of somatic mutations [53].

Non-synonymous driver mutations

The most common way of discovering driver mutations and genes is to search the coding regions of the genome for recurrently mutated genes in a tumour sample cohort. Some common assumptions regarding coding region mutations are that if they affect functionally important amino acid residues then they are likely more deleterious. Functional impact of amino acid residues is usually inferred from evolutionary conservation and protein domain analysis [21]. Long-tail driver mutations are harder to detect than highly recurrent mutations. They can potentially be identified is by considering recurrence of mutations in clusters of spatially close residues in proteins instead of only in linear sequences. They

could, for example, be found by looking for missense mutations that cluster together in 3D proximity in protein structures above expected background rate [9].

Non-synonymous mutations are mutations that cause an amino acid change within the amino acid chain. These can be either missense, nonsense, or indel mutations. Genes known to be targets of non-synonymous mutations are for example *BRAF, APC, TP53* and *KRAS*. Cancer genes are generally classified into oncogenes and tumour suppressor genes. Oncogenes and tumour suppressor genes have different pattern of somatic mutations. Tumour suppressor genes have an excess of inactivating mutations and oncogenes have mutations near specific amino acid residues [36]. This is because tumour suppressor genes cause cancer by being inactivated and oncogenes by becoming more active or by changing the protein, they code into an oncogenic one [5]. Analysis of mutations within cancer genomes have found that recurrent missense mutations have a recurrence rate of around 29.1%. Missense mutations often target highly conserved sites within genes, and they are also enriched in charged amino acids compared to non-charged amino acids [46].

Tumour suppressor genes are generally affected by different types of inactivating mutations. *TP53* is somewhat different from the usual pattern of tumour suppressor genes as it is often inactivated by missense mutations. Exons 4 to 9 are especially common locations of the driver mutations as they contain the coding sequence for the binding site of the p53 protein. The main effects of the various *TP53* mutations is that it results in p53 being unable to bind to its targets. *APC* is a more typical tumour suppressor gene in that it is more likely to be inactivated by a deletion or a nonsense mutation [44]. A potential tumour suppressor gene which may contain rare driver mutations is *PTEN*. Rare driver mutations are unlikely to be recurrent and instead they may cluster within certain domains in proteins. For example, a 3D analysis of these kind of clusters found that *PTEN* had 15 clusters in it that may fit this profile. The clusters were within the flanking regions surrounding a phosphatase catalytic core motif, which is necessary for *PTEN* activity. [9]

In some cases, tumour suppressor genes can have mutations in both alleles disabling them both. In a pan-cancer analysis of somatic mutations, *TP53*, had both alleles mutated in 77% tumours that had driver mutations in the gene. 96% of these two-hit events in *TP53* had a somatic point mutation in one allele and a deletion in the other one. Other cancer predisposition genes in the same analysis were hit by biallelic inactivation due to protein-truncating germline mutation and a somatic mutation in 4.5% of the patients [53]. Li-Fraumeni syndrome is a cancer predisposition syndrome where most families with this disorder have an underlying *TP53* germline mutation. Families with missense mutations

in the *TP53* core DNA binding domain have higher incidence and earlier onset of cancer in comparison to families with other types of mutations in *TP53* [44].

Microsatellite instability (MSI) is a type of process that leads to the accumulation of large numbers of indels in the genome. It is typical to some cancers such as colorectal and endometrial cancers where MSI is present in 15% of the tumours. Germline mutations in DNA mismatch repair genes predisposes individuals to these cancers. *BRAF* and *KRAS* have also been implicated as driver genes in MSI cancers. Additionally genes *CTNNB1* and *PIK3CA* are mutated in 1.6% to 4.7% of MSI colorectal cancers, and over half of the mutations in them tend to accumulate in hotspots. All of these genes have been identified to have driver point mutations in MSI colorectal cancers [18].

Approximately 50% of skin cancers have *TP53* mutations. Different skin cancers types have different *TP53* hotspot mutations, but they all have arginine codon 248 mutation in common. In UV-induced skin cancers these mutations seem to form early, and they can drive the formation of precancerous lesions [44]. Another commonly mutated driver gene in melanoma skin cancers is *BRAF*. 50% of melanomas have a V600 mutation in this gene. Another common driver gene for melanomas is *NRAS*. The V600 mutation in *BRAF* and mutations in *NRAS* are generally mutually exclusive in melanoma. Melanomas have driver mutations in other genes too. One driver mutation in the *RAC1* gene in melanomas is a c.85C>T transition that results in P29S amino acid change. This gene plays a part in cytoskeleton rearrangement and by extension cellular adhesion, migration, and invasion. The prevalence of this mutation is estimated to be 3.9% in melanomas. Its predicted effect is that it leaves RAC1 in an active, GTP-bound state. Normal RAC1 switches between active GTP-bound and inactive GDP-bound states. The common *BRAF* and *NRAS* mutations in melanoma do not conform to the common UV-light induced pattern of C→T mutations, but this *RAC1* mutation could be a mutation induced by UV-light damage [13]. *RAC1* has another activating mutation at A159. This gene has also been identified to have two potential rare driver mutations in the same residue as the previously identified mutations. The potential additional mutations are G15S and C18Y and they both result in increased RAC1 expression levels [9].

Synonymous mutations as driver mutations

The genetic code uses 61 different codon triplets to encode 20 amino acids. This means that there is overlap between codons and the amino acids they encode. Mutations that change a nucleotide but do not result in a change in the amino acid are known as syn-

onymous mutations. These synonymous mutations can however influence the messenger RNA (mRNA) translation accuracy and speed, mRNA folding, mRNA splicing, or through translational pausing how the protein folds. As such while the mutation may be silent in respect to the amino acid it may have other effects on the function of the sequence and the protein. Synonymous mutations thus may play a part in cancer formation [50]. Synonymous mutations are the second most frequent type of point mutation after missense mutations, and they have been studied less than missense mutations in cancer [46]. Driver mutations have often been distinguished from passengers by comparing the frequency of protein-coding changes to the frequency of synonymous mutations within genes. This approach is obviously not suitable for detecting synonymous mutations which are under positive selection.[50].

A pan cancer analysis of synonymous mutations found that the fraction of recurrent synonymous mutations, 26.8%, was similar to the fraction of recurrent missense mutations. The most frequent synonymous mutation was *NCOA6* c.807G>A which was found 63 times. The most frequent synonymous mutation that has not been listed as polymorphism was found 45 times. The mutation in question was *CHEK2* c.1176G>T, it is located in a tumour suppressor gene. The genes *TTN* and *KCNJ12* both had multiple occurrences of synonymous mutations. For *TTN* this is likely explained by its length [46].

Oncogenes that are known to be activated by missense mutations are also enriched by synonymous mutations. These oncogenes have 23% to 30% excess of synonymous mutations in their exons compared to genes with similar genomic features. This same phenomenon is not seen in tumour suppressor genes and instead they may have less than expected synonymous mutations in their exons. Synonymous mutations are also enriched in the whole genomes and not just in the exons. Variation within the local mutation rate is thus unlikely to explain the enrichment of synonymous mutations. This enrichment is also not seen in oncogenes that are amplification or translocation activated [50]. Synonymous mutations are also enriched within known cancer genes [46].

The enrichment of synonymous mutations cannot be explained by these mutations hitchhiking on positively selected missense mutations in missense-activated oncogenes. Oncogenes with missense mutations are unlikely to contain synonymous mutations in the same gene within the same tumour sample. The avoidance of missense and synonymous mutations in the same gene are strongest in oncogenes mutated within a particular tissue. This avoidance is not seen in tumour suppressor genes. Some cancer types are more likely to have synonymous mutation enrichment and others have lower enrichment. The enrich-

ment is generally lower in cancer types with a high mutational load such as melanoma or head and neck cancers. Cancers that have low mutational load such as leukaemia and breast and ovarian cancer have more synonymous mutation enrichment. Within individual tissues the synonymous mutations and missense mutations are enriched in the same tissue-specific oncogenes. These mutations however are not seen together in the same gene within the same sample [50].

Analysis of a group of receptor tyrosine kinases, *IL7R* and *TSHR* receptors, the structural protein *ELN*, cytoplasmic kinases *JAK3* and *ITK* and the transcription factors *GATA1* and *RUNX1T1* and a few other genes showed that synonymous mutations target evolutionarily conserved synonymous sites. Within these genes the synonymous mutations are likely to appear within five codons of another synonymous mutations creating clusters. This is similar to the phenomenon of missense mutation clustering in oncogenes [50]. Both missense and synonymous mutations are generally depleted near the 5' and 3' ends of the exons. At 5' end the strongest depletion is seen within the first 10% of the coding sequence but the effect is still seen within the first 50 codons. Similar pattern is seen in first, last, and monoexonic transcripts. Within internal exons synonymous mutations are depleted near intron and exon boundaries [46].

While conserved sites are targets of synonymous mutations, they do not seem to influence oncogene activity by changing codon optimality, mRNA folding or by targeting micro RNA (miRNA) binding sites. Instead, they seem to be enriched near exon boundaries and preferentially result in the gain of exonic splicing enhancer or the loss of exonic splicing silencer. This phenomenon is stronger in sites where the flanking splice site sequences diverge much from the consensus sequence resulting in weaker splice sites. Oncogenes with synonymous mutations were associated with abnormal splicing compared to oncogenes with non-synonymous mutations or non-cancer cells with synonymous mutations [50]. In a pan cancer analysis 26.8% of synonymous mutations caused a change in predicted exonic splicing enhancer or the loss of exonic splicing silencer. More than 40% of synonymous mutations seem to target conserved sites. Synonymous mutations aside from affecting splicing can also affect alternative promoter usage. Alternative events associated with synonymous mutations are mostly exons known to be subject to alternative splicing [46].

Synonymous mutations are similar to missense mutations in their nucleotide changes. Common nucleotide changes are C→T and G→A substitutions. While missense and synonymous mutations were similar in nucleotide changes, they were different on amino acid level. Synonymous mutations were enriched in hydrophobic amino acids. For example,

synonymous mutations were enriched in phenylalanine codons. When assessing structural impact on mRNA, transversions involving G had the largest structural impact. C→T transitions on the other hand had smallest impact [46].

So far synonymous mutations have been mostly associated in changes within alternative splicing. Structural changes caused by synonymous mutations in mRNA could affect its stability, translation efficiency, and folding. Stable mRNA pseudoknot structures cause translational pausing and regulate translation speed. This impacts co-translational protein folding and the interactions with cellular components. The first codons of mRNA are generally depleted of synonymous mutations. The mutations that are present within the first codons are likely to have stronger structural impact [46].

KRAS belongs to a family of oncogenes called *RAS* and they encode small GTPases. Mutations in *RAS* genes can lead to inhibition of GTP hydrolysis. Mutations in *RAS* are quite common in human tumours as 30% of them have a *RAS* mutation. As such *KRAS* is a commonly mutated oncogene. *KRAS* c.30A>C mutation is found near the 5' end of the sequence. This mutation affects the secondary structure of the resulting protein locally. Synonymous mutations within *KRAS* codons 12 and 13 affect the expression rate of *KRAS* with other mutations increasing it and others decreasing it. These synonymous mutations could affect the oncogenicity of *KRAS* [46].

While tumour suppressor genes in general are not enriched in synonymous mutations. *TP53* however is an exception as it can be enriched for synonymous mutations. In *TP53* the synonymous mutations target nucleotides directly adjacent to splice sites. Most of these mutations are within the two terminal nucleotides of an exon. 3' terminal nucleotide synonymous mutation of *TP53*'s 4th exon is a recurrent target. In germline variant it is known to cause Li-Fraumeni syndrome producing an aberrantly spliced mRNA. The mutation causes intron retention and activation of cryptic splice sites. Another exon where this happens is the 6th exon. In this exon a mutation can cause activation of a cryptic splice site 5 nucleotides downstream from the end of the exon resulting in a frameshifted mRNA. The terminal guanine is also often lost in exon 9 [50].

Latent and mini-driver mutations

The sequencing of tumour genomes has revealed few new driver genes that have been found to be mutated at high frequencies and instead there are many genes that are mutated at lower frequencies. Thus, it is possible that rather than there being few driver mutations

with major impact on the fitness of cancer cells there are many mutations that provide only small selective advantages to the cancer cells. These mutations that have only small effects on the fitness of the tumour are called either mini-driver mutations [6] or latent-driver mutations [35]. As such the mini-driver mutations do give an advantage to the tumour cells but are not crucial for them. As the tumorigenesis progress the number of mini-driver mutations would increase in the tumour cells [6]. While individual mini-driver mutations will not confer a selective advantage there is some evidence so far that the presence of mini-drivers will worsen outcomes of lung adenocarcinoma patients [4].

Individual mini-driver mutations are likely be only present in small proportion of tumours, but tumours are likely to have multiple mini-driver mutations in them and sub-clones are likely to have their own mini-driver mutations that are not necessarily present in all cells of the tumour. As the mini-driver mutations have only a small effect it would be unlikely for all sub-clones to have the same ones. Sub-clones could have convergent evolution between their mini-drive mutations too and while the mini-driver mutations might not give an advantage to the tumour, they could give a competitive edge to a specific sub-clone against the other sub-clones within the tumour. The opposite could also be possible where some sub-clone has mini-driver mutations in it that are deleterious. The targets of mini-driver mutations would likely include non-coding genomic features such as regulatory elements, transcript switching, and RNA stability [6].

Multiple mini-driver mutations could also substitute for a more major driver. In this case their effects would likely be weaker than that of the major driver. Most mutations in *KRAS* in cancer are found from its codons 12 and 13 with occasional ones found from codon 61, however codons 146 and 117 can have atypical clonal mutations in them. It is possible that these atypical mutations could act as mini-driver mutations [6].

Mini-driver mutations could also potentially adjust the effects of mutations that are already present in the tumour. Such adjustments could be for example, that the mini-driver mutation amplifies the effects of a major driver or that it compensates for deleterious mutations that the tumour has accumulated already. These adjustments could happen to major drivers too, because the major driver may have affected the tumour growth positively before but as the environment and mutations in the tumour have changed this may no longer be so. Mini-driver mutations could also remove functions in the cancer cell that it no longer needs. Passenger mutations could also change into mini-driver mutations as the tumorigenesis progresses further [6].

Alternatively, these types of mini or latent drivers could work so that all additional latent

drivers have a small effect individually, but their cumulative effects start to cause changes. It is also possible that individual latent drivers do not have an effect and instead two or more mutations are needed for any change. In this case if two mutations are needed then having either mutation alone would leave them as passenger mutations. Regular driver mutations by contrast would have an effect even if they were present alone [35].

Cancer symptoms usually take a long time to emerge compared to drug resistance which may emerge quickly within cancer. The reason for this may be these latent driver mutations. Most research considers driver and passenger mutation distinction to be binary. However, under some conditions a passenger mutation could act as a driver instead. In drug resistance it may be that the mutations already exist as passengers in the tumour and the change in selective pressure changes them into driver mutations leading to the growth of the tumour [35]. There is so far some evidence that the accumulation of mini-drivers within genes, that are part of extracellular matrix organisation, epithelial-mesenchymal transitions, and blood vessel development and circulation in lung adenocarcinoma leads to worse survival of patients with this disease [4].

Identifying mini-driver mutations is a challenge as they are likely to be uncommon across cancer types and they may only be present in sub-clones and be present in cancers with high background mutation rates [6]. Conventional methods that are used for identifying driver mutations tend to focus on mutations that have a significant effect on the target gene or protein or ones that have a significantly higher recurrence that would be expected. These models are unlikely to work on mini-driver mutations as they do not have significant effects on their targets alone. Analysis of larger modules that accumulate mutations however seems promising [4].

Non-Coding driver mutations

Non-coding driver mutations have been less studied in comparison to coding mutations [53]. One reason for the smaller focus on non-coding regions is that it has been difficult to annotate the variants found from the non-coding regions [23]. An estimated 98% to 99% of the human genome is non-coding [1]. The non-coding regions contains multiple interesting regulatory elements such as promoters, enhancers, and insulators which regulate gene expression within the non-coding regions that could be targets of driver mutations. Promoters bind core transcriptional complex to itself while enhancers bind transcription factors, which can affect gene expression rates greatly. [40] Driver mutation identification is more difficult in non-coding regions due to poorer understanding of the mutational pro-

cesses there. Another challenge is in building a background mutation model which can be difficult due to lack of easily interpreted neutral events the mutations could be compared to [43]. In a pan-cancer analysis of somatic mutations a fourth of tumours had at least one potential non-coding driver mutation. The most well characterised non-coding region affected by driver mutations is the *TERT* promoter and it was affected in 9% of the tumours. *TERT* is a gene responsible for maintaining telomere length. According to the same analysis individual enhancers and promoters are infrequent targets for driver mutations [53].

Over expression of *TERT* can be caused by point mutations that lead to the formation of a new transcription factor binding site, insertions of viral enhancers upstream of the gene, or through chromosomal amplification [53]. In *TERT* promoters non-coding mutations can cause new binding sites for the ETS family of transcription factors. This leads to over expression of *TERT* and subsequently increased telomere length maintenance and cell survival [40] [36]. *TP53* may occasionally be affected by non-coding mutations. SNVs and deletions in its promoter can affect the transcription start site or donor splice site of the first intron. These mutations often occur with loss of heterozygosity and result in lower mRNA expression. This type of inactivation of *TP53* by non-coding mutations is still infrequent in tumours [43].

Amplification activated oncogenes show enrichment of mutations in the 3' untranslated region (UTR). This enrichment in 3' UTR is generally seen in cancer genes that have elevated expression levels in the tumour [50]. Mutations in the 3' UTR of different genes can cause either the expression of the gene to decrease, as happens with *TOB1*, or to increase, as happens to *NFKBIZ*. *TOB1* encodes an anti-proliferation regulator and affects migration and invasion in gastric cancers. *NFKBIZ* 3' UTR mutations concentrate in a hotspot near the stop codon and upstream of conserved miRNA binding sites [43]. Recurrent mutations in 3' UTR of *NOTCH1* result in aberrant splicing in chronic lymphocytic leukaemia. 3' UTR mutations of *CD274* on the other hand in gastric cancer patients have been shown to disrupt miRNA mediated degradation of mRNA resulting in overexpression [36]. Mutations in the promoter or 5' UTR of *MTG2*, have been associated with lower expression of *MTG2* and the effect of these mutations has previously been studied in vitro showing similar results [43].

Non-coding mutations in haematological malignancies have been identified to be active in the pre-transcription regulation of genes. Enhancer activity is often targeted towards a specific promoter by containing the enhancer-promoter loop with insulator elements.

These insulator elements anchored by CCCTC-binding factors known as CTCFs. Mutations in the CTCF elements, hypermethylation of CTCF binding motifs, or mutations in the cohesin complex can result in loss of insulation. This then can result in aberrant gene expression of the neighbouring genes [40]. One of the genes associated with CTCF mutations is *TGFB* and it promotes blood vessel formation and tumour cell migration in melanoma [22]. Different tumour types have different pattern of mutation in cohesin-binding sites. Pol ε exo$^-$ tumours have a reduced rate of somatic mutations near the site while microsatellite stable (MSS) tumours tend to have a spike of mutations at 17-bp CTCF binding motif. MSI samples have a somewhat lower mutation frequency overall near the cohesin-binding sites [15].

In blood cancers there are mutations within the non-coding genome that lead to aberrant regulation of gene expression. These kinds of mutations are present in T-cell acute lymphoblastic leukaemia. Such mutations are for example present in *TAL* and *LMO* transcription factors. The mutations cause these genes to be dysregulated. These transcription factors are expressed in early T-cell differentiation and mutations, in the aforementioned genes, result in T-cell differentiation disruption and the proliferation of malignant precursors. The mutations do not always have to be present within the transcription factor gene or binding site and instead they can be found from elsewhere in the genome. In the case of *TAL1* it can be influenced by a non-coding mutation 7kb upstream. Indels of length 2 to 18 bp at this location can cause transcription factor MYB to bind at this location. MYB is a master transcription factor and it is capable of recruiting TAL-LMO complex at this aberrant site resulting in expression of *TAL1* from this location [40].

Medulloblastoma is a highly malignant paediatric brain tumour, and it can be classified into four subgroups. Subgroups 3 and 4 are the mostly poorly understood ones and *GFI1* and *GFI1B* are highly prevalent oncogenes in these groups. Structural variants at 9q34 cause upregulation of *GFI1B* by moving it under the regulation of super enhancers. As with *GFI1B* the activation of *GFI1* is also often associated with structural variants. *GFI1* is also associated with enhancers as it can translocate to highly expressed sites. Both *GFI1B* and *GFI1* activate when they end up near enhancers and they can work together with *MYC* to induce cancer. This type of enhancer hijacking may be present in other oncogenes too [34].

Non-coding RNAs may also play a part in cancer. The non-coding RNA *RMPR* can have significantly mutated exon and promoter in multiple tumour types. These mutations tended to accumulate in sites that can affect the secondary structure of proteins [43]. There

are several different RNAs and possible modifications in the non-coding genome such as: mRNA splice sites, non-coding RNA genes, and long non-coding RNA (lncRNA). The last ones are especially interesting as lncRNAs can control gene expression. For example, *HOTAIR* lncRNA is highly expressed in breast tumours and metastases and [21] lncRNA *MALAT1* has been shown to be mutated in oestrogen receptor positive breast cancers [36]. mRNA expression can also be altered by mutations in if the mutation happens in miRNA that regulates mRNA or in the mRNA where the miRNA is supposed to bind to. [21].

2.5 Detecting driver mutations and genes

Driver genes and driver mutations have both been looked for by multiple different algorithms, but they tend to follow similar principles for the most part [39] [29]. The algorithms can either focus on gene level detection of driver genes or sub-gene level and look for individual driver mutations or clusters of driver mutations. Both levels of algorithms can focus on detecting driver elements from either linear sequence, usually either DNA or amino acid sequences, or they focus on detection from 3D structures of proteins [39] [29].

Most algorithms in driver detection work by comparing the background mutation rate and the expected distribution of mutations or the expected types of mutations against the observed mutations. A challenging aspect within these methods is that the background mutation rate is different between different cancer types and even between different individuals. Thus, building a model that can accurately model the background rate well is critical [29]. Most tumours can also have thousands of mutations and usually these mutations are unique to that tumour and other tumours are unlikely to share them. This is explained by the driver-passenger division of mutations, where most of the mutations are passenger mutations that are rarely seen across different tumours. In this view the rare but recurrent mutations are likely to be driver mutations while the more common but not recurrent mutations are passengers [39].

On the gene level driver genes can be detected by detecting positive signals of selection for the gene in the cancer. First, the number of observed mutations can be compared to the number of expected mutations. Secondly the mutational clusters present can signify positive selection as there can be clusters either in domains of interest in the linear sequence or there may be clusters in the protein 3D structure, which in the linear sequence may be far away from each other. Normally in the case that there is no positive selection the

expected distribution of these mutations would be more random. Instead of clusters it is also possible to see biases in the mutations in some way. For example, the mutations in a gene may have higher than expected number of mutations with high functional impacts. There could also be bias in the nucleotide context of the mutations in that some nucleotide contexts accumulate more mutations than expected [29].

The sub-gene level has similar algorithms to the gene level. The sub-gene level algorithms also mostly focus on finding clusters from either linear sequences or 3D structures. These sub-gene algorithms can be classified into four subgroups. The first group consists of algorithms that look for clusters of mutations in the sequence. The main differences between methods in this class is the background model they use. The second group consists of methods that look for the mutational clusters from 3D structures of the proteins. The exact way these methods work can vary in how they interpret the 3D data they use. The third group is like the first one in that these methods also look for linear sequences but the sequences in this case must have known functional regions. In this group the methods compare how large portion of the mutations are within the functional region compared to the rest of the sequence. The third group also has methods that analyse the same domain in multiple different proteins to find recurrently mutated positions. The fourth group is like the second and third and the methods in this group look for externally defined regions in the 3D structures of proteins [39].

The algorithms that rely on 3D protein structures and externally defined interesting domains have their limitations in practical use. Experimentally determined protein structures are available for only a limited number of proteins. Similarly, the need for domains limits the studied sequences to those for which there is knowledge on the functional domains in the sequence [39].

On the sub-gene level algorithms that belong to the same group tend to get similar results. Overall, sub-gene algorithms are more likely to find oncogenes than tumour suppressor genes. This could be because these algorithms tend to focus on finding clusters of mutations within genes, and oncogenes potentially have more mutational clusters than tumour suppressor genes, while gene level methods focus more on mutational recurrence of a gene in a cohort [39]. Different sub-gene and gene level algorithms can be used together to complement each other and to ensure that they look at variety of features of the mutational pattern in genes. This ensures that a wider variety of driver genes are found [39] [29].

While many algorithms can be classified in the above-mentioned categories, in the principle on which they operate on when detecting signals of positive selection of mutations in genes,

there are many other algorithms that work differently. For example, deep learning can be used to detect driver mutations based only on the raw sequence data. Deep learning allows the model to learn independently the important features from the input data. This means that the model focusing on detecting driver mutations from pure DNA sequences should learn the features that separate driver mutations from sequences that do not contain driver mutations [2]. Deep learning is further discussed after the different ways of detecting and scoring mutations.

Introduction to previously used driver mutation and gene finding frameworks

One algorithm that belongs to the sub-gene algorithm category 1 algorithms is OncodriveCLUST [39]. OncodriveCLUST uses clusters of mutations within specific regions of amino acid sequence to detect driver mutations. It compares the measured bias of mutation clustering compared to its background model, which consists of coding region synonymous mutations. As the synonymous mutations are assumed to not be under selective pressure, they should reflect the baseline of somatic mutation clustering. OncodriveCLUST consists of five steps where it retrieves all non-synonymous mutations and positions, the ones that have mutation frequency above the background threshold rate are identified as cluster seeds. After this these cluster seeds are joined together to form a cluster if they are close enough to each other and mutations which fall into these clusters are added to them if they were not identified as cluster seeds. Lastly, a score is calculated for the cluster, the score is directly proportional to the fraction of mutations within the cluster and inversely proportional to the cluster length. The final score for a gene is calculated by summing over all clusters in it. OncodriveCLUST is better at finding recurrently mutated drivers and less likely to work well for tumour suppressor genes as they may have more even distribution of mutations [51].

LARVA is a computational framework designed to find highly mutated regions in the genome and especially in the non-coding regions of the genome and regulatory elements. LARVA treats the mutation count in the given regulatory regions as beta-binomial distributed random variable. LARVA's workflow starts by pre-processing mutations through quality control and intersecting them with annotations for regulatory regions in the non-coding regions. The processing itself consists of counting all variant intersections with the annotation categories and corrected for covariates. Significance analysis is then performed by fitting a background model, beta-binomial distribution, and then p-values are calculated for the regions under analysis. LARVA is for example capable of detecting that the

TERT promoter is a significant region [23].

deepDriver is a framework that can predict driver genes based on somatic mutations using convolutional neural network architecture. deepDriver uses a similarity network between mutations as its input. The similarity network is constructed by calculating Pearson's correlation coefficient between 12-feature vectors constructed for genes. The feature vector, for example, contains the fractions of different mutation types. The similarity network is then constructed by calculating correlation coefficients between the genes and connecting them to each other. A gene is connected to k other genes with the k highest correlation coefficient values by k-nearest neighbours' algorithm. The similarity network is then represented as a matrix that is used as the input in a convolutional neural network consisting of two convolutional layers and a fully connected layer [24].

Convolutional neural networks can also be used to detect driver mutations from the DNA sequence alone. Training a convolutional neural network to detect driver mutations will also need sequences which do not contain driver mutations as the negative examples. The convolutional neural network should directly learn the important features for separating sequences with a driver mutation from sequences with no driver mutations. The length of the sequence is also important as too long sequences may contain unimportant information that may end up confusing the neural network. Methods like this are still somewhat limited as the linear DNA sequence does not capture potential three dimensional effects that may contribute to the mutation processes [2].

Mut2Vec is a pipeline that generates distributed representations of mutations and it can separate passenger and driver mutations from each other. In this case if an unidentified mutation is close to many known driver mutations in the distributed space, then it too is likely a driver mutation. Mut2Vec is based on Skip-Gram model. The Skip-Gram model is a multi-layered neural network. It consists of an input layer, an embedding lookup layer and the prediction layer. Its goal is to predict surrounding entities based on the entity embedded within the network. This is achieved by training the model on an entity and its surrounding contextual entities. In the case of Mut2Vec the Skip-Gram is first trained on PuBMed abstracts to extract the context of genes and the lookup layer is initialised with gene word vectors. After this the model is trained with gene-level mutation vectors. Lastly, the model can be retrofitted with protein-protein interaction network data from BioGRID. This results in gene-level mutation vectors. These vectors can then be used to represent mutations in a two-dimensional space and passenger mutations and driver mutations end to cluster to their own groups [17].

ParsSNP is a parsimony based unsupervised method for prioritising driver mutations. ParsSNP assumes that the proportion of driver mutation drops as the number of mutations in a sample increase. ParsSNP uses an expectation-maximisation framework. In the training phase ParsSNP assign a random label between 0 and 1 for each variant and calculates descriptors for the variants. The descriptors can be calculated or obtained from other methods. During the expectation step the labels are updated using the Bayes Law and the belief that in a sample with N mutations, between 1 and $log_2(N)$ mutations are drivers. The maximisation step has a probabilistic model that updates the labels according to the descriptors. The expectation and maximisation steps iterate until convergence is achieved. The final labels are then used to train a separate neural network, which can then take new input variants and score them. The final score correlates with the recurrence of a mutation and the score can be used to indicate potential driver mutations [19].

OncodriveFML

OncodriveFML is one widely used framework for detecting coding and non-coding driver mutations. The idea behind OncodriveFML is that tumours have a bias towards accumulating high impact mutations in certain genomic regions [32].

OncodriveFML uses functional impact score (FI) and functional mutation score (FM). Functional impact means that the score can for example measure the mutations effect on the protein structure or the binding affinity of different molecules and their binding sites. Functional impact can also measure effects on the mRNA. Functional impact score can be any score of this type that can be calculated for all positions in the genomic elements under analysis. For example, a score that can be used as FI score is the CADD score [32, 42]. The only requirement for FI is that it is meaningful for the genomic elements in question [32].

OncodriveFML starts by first gathering the FI scores for the observed mutations from a FI score file provided for it. Then it calculates the average FI score for the observed somatic mutations. After this OncodriveFML samples N times the same number of mutations randomly from the FI score file and calculates the mean FI scores for these N instances. The sampling of the mutations can be done by having the same probabilities for all mutations or the mutation probabilities can be calculated from the observed mutations. It is also possible to set custom probabilities. Thirdly OncodriveFML compares the observed average FI score with the mean FI values from the N sampling instances. OncodriveFML calculates an empirical p-value for the observed mean FI score by comparing it to the N

mean FI values. This is done by calculating how many times the observed mean FI value is bigger than the randomised mean FI value. This value is then normalised by dividing it with the number of randomisations performed. Lastly, OncodriveFML performs multiple-testing correction with Benjamini-Hochber correction to rule out false positives that may have been generated during the N sampling procedure [32].

2.6 Scoring methods for functional impact and pathogenicity of mutations

There have been multiple attempts to score mutations according to their functional impact or by their pathogenicity. Below are some examples on what kind of scores have been made.

Cancer Mutation Census

Cancer Mutation Census (CMC) is a project that classifies the coding mutations in Catalogue Of Somatic Mutations In Cancer (COSMIC) and attempts to find variants that are driver mutations in cancer. CMC ranks the mutations into mutation significance tiers based on the available evidence if the mutation is pathogenic or not. The tiers are 1,2,3 and 'Other'. Tier 1 is the highest pathogenicity tier and mutations in this category are classified as pathogenic. Mutations are classified into this category if they have clinical significance 5 in Clinvar cancer-related diseases and it is a recurrent missense mutation in oncogene or a loss-of-function mutation in a tumour suppressor gene. Significance level 2 mutations can be like category 1 mutations, but the clinical significance level is 4 instead of 5. Alternatively level 2 mutations may have clinical significance level of 4 or 5, but it is not recurrently present in known oncogenes or tumour suppressor genes, but still shows evidence of positive selection by a dN/dS algorithm. Alternatively level 2 mutations may be recurrently present in known cancer genes and show evidence for positive selection but have no Clinvar evidence. Level 3 mutations show one category of evidence as they can be either recurrently present in known oncogenes or tumours suppressor genes, or show evidence of positive selection, or have Clinvar significance of 4 or 5. Mutations in the 'Other' category either have no available evidence or show it from any of the categories [47].

Combined Annotation-Dependent Depletion

Combined Annotation-Dependent Depletion (CADD) is a score that can be used to measure the deleteriousness of a variant. The scores are based on various features such as sequence context, gene model annotations, evolutionary constraint, epigenetic features, and functional predictions. CADD uses a machine learning model, such as logistic regression, to then predict a score for a variant based on these features [42]. For training CADD uses alleles that have fixed since human-chimpanzee lineages diverged as likely benign or neutral variants as otherwise we might expect them to have disappeared if they were very deleterious. Another group of variants CADD uses in training are a group of simulated mutations that are free of selective pressures. Most of these simulated mutations are also likely benign but a group of deleterious mutations should also arise from them. The difference in their annotation features is used in making the final score when training the machine learning model [42]. A limitation of CADD is that some of the simulated variants are neutral but it is unknown which. Because of this CADD's performance is evaluated on known curated datasets of different diseases or functional effects [42].

CADD score can also be divided into the RAW and PHRED scores. The RAW score is the raw output from the model, and they are not comparable between training rounds or different models. High RAW score means that the variant has been more likely derived from the simulated variant dataset rather than the neutral variant dataset. This means that higher RAW score indicates higher deleteriousness. PHRED scores are scaled so that the score indicates in which percentile it is. This means that PHRED scores can be compared between each other as regardless of the details in the model a score of 20 or greater would indicate that the score is in the top 1% of scores [42].

Synonymous Mutations In Cancer database

Synonymous Mutations In Cancer database (SynMICdb) is a score that was created for 659,194 synonymous mutations from 13,935 tumour samples to measure the functional impact of synonymous mutations. The synonymous mutations were curated from COSMIC v76. This score consists of mutation frequency, mutational signature, mutation load, evolutionary conservation, cancer gene annotation, SNP annotation, FATHMM-MKL score, CADD score, and predicted RNA secondary structure changes. The frequency indicates the frequency of the mutation after correcting for mutational bias. The mutational load indicates the mutational load of samples affected by the mutation. The SynMICdb score

ranges from -4 to +12, with higher score indicating a higher likelihood of the mutation having a functional impact. A score higher than 0.89 would put the mutation into the top 50% while a score of 4.38 puts it in the top 1% [46].

Oncogenic driver Variants

Oncogenic driver Variants (OncoVar) is a platform for driver genes and driver mutations. The cancer data is gathered from the TCGA data and International Cancer Genome Consortium (ICGC). The platform has cancer type and dataset specific files for download. The interesting features of the OncoVar platform are the OncoVar scores for mutations and genes, the consensus-score and driver level. The OncoVar scores measure the 'driverness' of a mutation or a gene. Driver mutations were defined as mutations that had high scores and occurred in at least two samples. Driver genes were based on a consensus score from multiple other driver gene sets, including Cancer Gene Census, and if it had high enough score it was assigned as a driver. OncoVar scores were adjusted also on a Gaussian model so that they could be compared. A higher OncoVar score indicates higher pathogenicity. [52].

2.7 Deep Learning

Machine learning methods attempt to learn relationships between features from the data without having to predefine what they might be beforehand [3]. Deep learning is a subfield of machine learning where the methods are representation learning based. Representation learning allows for the automatic discovery of features from raw data without human interference. This contrasts with classical machine learning methods where features had to be carefully engineered by domain experts. This however did not always yield very good results hence deep learning techniques were developed to deal with these issues [20]. The representations in deep learning methods are multi-layered where each level of representation is obtained by composing non-linear functions on the representation from the previous layer. With enough layers very complicated features can be learned. The higher-level layers will learn which features are important for discriminating different classes from each other and which ones are irrelevant in classification tasks. Deep learning methods for example may learn from image data edges and what kind of patterns and shapes these edges form [20].

A common type of deep learning model is the deep neural network. The architecture of a deep neural network consists of an input layer, hidden layers, and an output layer. The input layer simply takes the input and moves it to the next hidden layer. Hidden layers in neural networks consist of neurons that get an input from the previous layer [3], multiplies it by the weight of the neuron and passes it on through an activation function to the next layer. If the neurons of one layer are all connected to all neurons from the next layer, then the layer is called fully connected layer [20]. In the case the neurons are only connected to a few neurons in the next layer the layer is called a convolutional layer. Convolutional layers consist of filters, a set of weights, that are moved over the input and they pick up features from the data. The result is that in the next layer there are smaller inputs that the filters from the previous layers created. The convolution is generally followed by the activation function that decide if the neuron is going to be activated or not [3]. A common activation function is called ReLU, rectified linear unit, its function is max(0, z) [20] and if the value of the neuron is below or at zero the neuron will not fire [3].

Convolution is generally followed by a pooling layer in convolutional neural networks. The purpose of a pooling layer is to reduce the dimensionality of the feature maps from the convolutional layers by merging similar information from close by points. A common method of pooling is called max-pooling, where the maximum values from local regions are passed onto the next layer [3]. Pooling layers are generally followed by another convolutional layer, which in turn is followed by another pooling layer. The combination of convolutional and pooling layers allows the model to learn features from the data [20]. The feature learning stage is generally followed by classification. The output has usually a SoftMax activation function. SoftMax transforms the output so that every possible output category has a value between 0 and 1 and all output values sum up to 1 [11].

These types of neural networks are generally used for supervised learning and classification tasks. In supervised learning the model is trained on data pairs that consists of the input features and a target value. For example, an image of a dog would be the input and the word ‘dog’ the target value. The point of supervised learning is for the model to learn to minimise the difference between the real target values and the ones it predicts for the input. The function it needs to minimise for this task is called the objective or loss function. During the training of the model the weights of the neurons are updated after each training round when the difference between the real and predicted values are known. After training the model is done, the model can be used to predict the target values of new data [20]. To speed up the training the data can be batch normalised before the activation

function. This zero centres and normalises the data allowing higher learning rates during the training [3].

In addition to supervised learning there is unsupervised learning where the target values are not used. In this case the aim is to discover patterns from the data. Various clustering methods belong to this class of learning [3].

3 Materials and Methods

3.1 Deep Mutation Modelling

Deep Mutation Modelling (DMM) is a method for modelling the probabilities of specific types of mutations at a given locus in a DNA sequence. It uses a deep neural network architecture to achieve this. The types of mutations DMM can model are single nucleotide variants, multi nucleotide variants, indels, structural variants and mobile elements. It also considers the possibility of there being no mutation at the given locus. DMM can also learn the mutational probabilities given the tumour type and sequence context. This gives DMM the potential advantage of giving cancer or even sample specific probabilities for mutations at given sites. These could then be used to form cancer and sample specific scores that can be used in further analysis. The input for the DMM model consists of mutations with associated site and sample level features. DMM has three main modules it consists of: the mutation event site, sample, and integrative site. Figure 3.1 shows the modules and how they connect to each other [8].

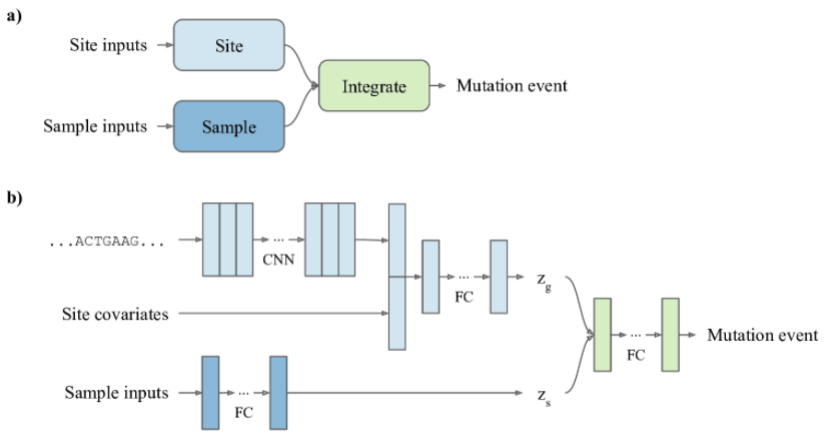


Figure 3.1: DMM - a deep neural network model architecture to predict somatic mutation events, Pitkänen *et al.* unpublished

The site module deals with the sequence input and site features, such as genic orientation, replication timing or whether the site is in an exon or non-coding region, in a convolutional neural network to learn these features and to predict the likelihood of mutational outcomes. The input sequences are one-hot encoded as binary matrices. The site module in the end

concatenates the sequence features extracted by the convolutional layers and processes them and the site features with fully connected layers. The output of this module is a feature vector. This module thus learns the sequence context of the mutations. [8]

The sample module processes sample related information such as such as tumour type, patient age and sex, sequence coverage etc. These features are processed by one-hot encoding the categorical inputs. They can then be given as input to another fully connected layer in the sample module which will output a sample level feature vector. The sample level output is the same for all input mutations that are from the same sample [8].

The third module is the integrative module which uses the feature vectors from the two previous modules as its input and processes them with fully connected layers. The layers use ReLU as the activation function except in the last layer where SoftMax is used instead. The output from this module is, given a mutation alphabet the size of m, a matrix of shape (c x m), where c is the length of the sequence around the central mutation that is considered. The rows in the output matrix represent the nucleotides and the columns the possible mutational outcomes [8].

The loss function DMM attempts to minimise is $L = H(Y, \hat{Y}) + \lambda \|z_s\|_1,\ \lambda > 0$ where z_s is the feature vector from the sample module. H is the cross entropy of the real and predicted mutations [8].

Overall, the model learns the probability of the mutational outcomes given the sequence context and sample representation. As driver mutations are expected to occur in more unusual mutational contexts, they are expected to be more difficult to predict with DMM than passenger mutations and thus should have lower accuracy for the true mutation prediction and lower probability of predicting any mutation [8].

DMM is trained by giving it sequences that contain a mutation in the sequence centre position. DMM is also given sequences where the centre position does not have a mutation as negative examples. The negative examples are generated by drawing a set of mutations for each sample from any other sample $s \neq \hat{s}$. These random mutations would then be represented by a mutation string with no mutation and the sample being s. In short, the negative examples are real mutation in some other sample but are represented as being an event with no mutation in the original sample s. Based on these examples like other neural networks DMM should learn the features from the sequence that affect the mutation probability alongside with the sequence features. DMM also has the capacity to learn which samples have high number of mutations and which have lower number of mutations. Figure 3.2 shows what kind of sequences DMM can take as its input [8].

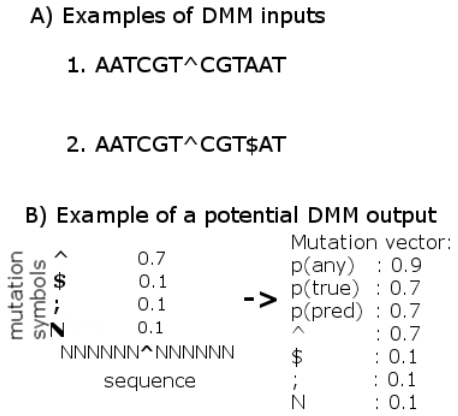


Figure 3.2: DMM input and output examples, in subfigure A the first sequence contains a C→T mutation represented by a wedge in the middle of the sequence, while the second sequence has the same mutation and an additional C→A represented by $ symbol in it. DMM uses the central mutation as its input. In subfigure B the output of DMM is originally a matrix which has probabilities for different predicted mutations for the central mutation. The matrix can be changed into a mutation vector that has the same probabilities. Additionally, it has probabilities for any mutation, the true mutation, and the highest predicted value.

The mutational outcome is given so that each mutation has a predicted probability for every possible mutation type as seen in Figure 3.2. Additionally, the probabilities for any mutation to have happened (p_{any}), the probability of the true mutation to have happened (p_{true}), and the probability of the mutation type that has the highest probability (p_{pred}) [8].

3.2 Materials

TCGA and PCAWG datasets

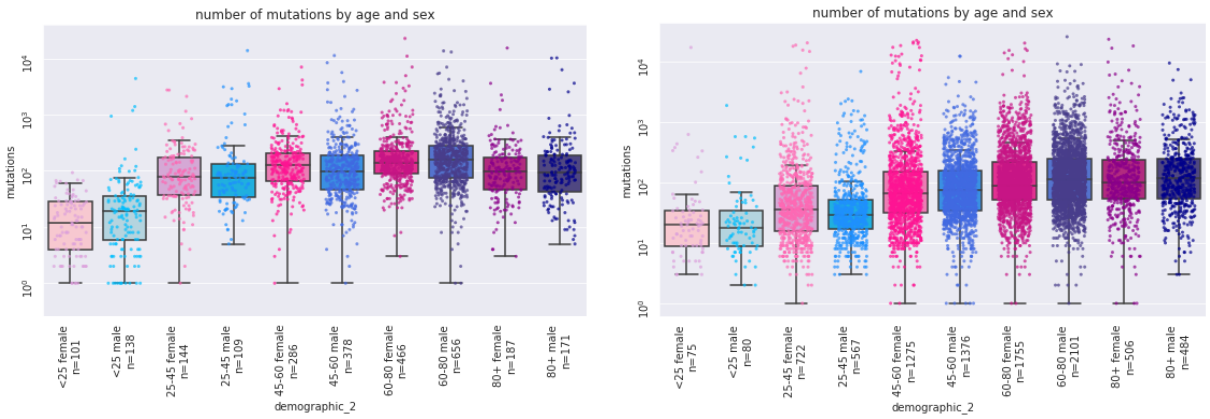
The main data I used were the Pan Cancer Analysis of Whole Genomes (PCAWG) [53] and The Cancer Genome Atlas (TCGA) datasets. The PCAWG dataset consists of whole genome sequence data from tumours while the TCGA is only based on exome sequence data from tumours. Both datasets used GRCh37 genome assembly.

The datasets had been annotated by DMM previously so that the data consisted of the mutations, associated sequence, and a few sequence features such as replication timing. Mutations that had no sample ID in associated sample information files were dropped.

The PCAWG DNA sequences I used, were 2048 base pairs long and the TCGA sequences were 256 base pairs long. I limited my analysis only to single nucleotide substitutions. In both cases I only examined mutations that, according to DMM's annotation process, were within exonic and genic regions, had a known replication timing, and were not duplicates of each other in the same sample. I also noticed some mutations within the TCGA dataset that had no known cancer type or sex of the patient associated with them and I dropped these from the analysis. The TCGA dataset also contained mutations that had reference or alternative allele sequence longer than one base pair, these mutations were also dropped.

In the end I had 739,341 mutations in the PCAWG dataset and 2,547,192 mutations in the TCGA dataset that I used for further analysis. The substitution counts can be seen in Table 4.1. According to this table C→T : G→A substitutions were the most common ones. These mutations covered 2636 different tumour samples in the PCAWG dataset and 8941 samples in the TCGA dataset. Figure 3.4a shows the number of samples and mutations per cancer type in the PCAWG dataset while 3.4b shows the number of samples and mutations in the TCGA dataset.

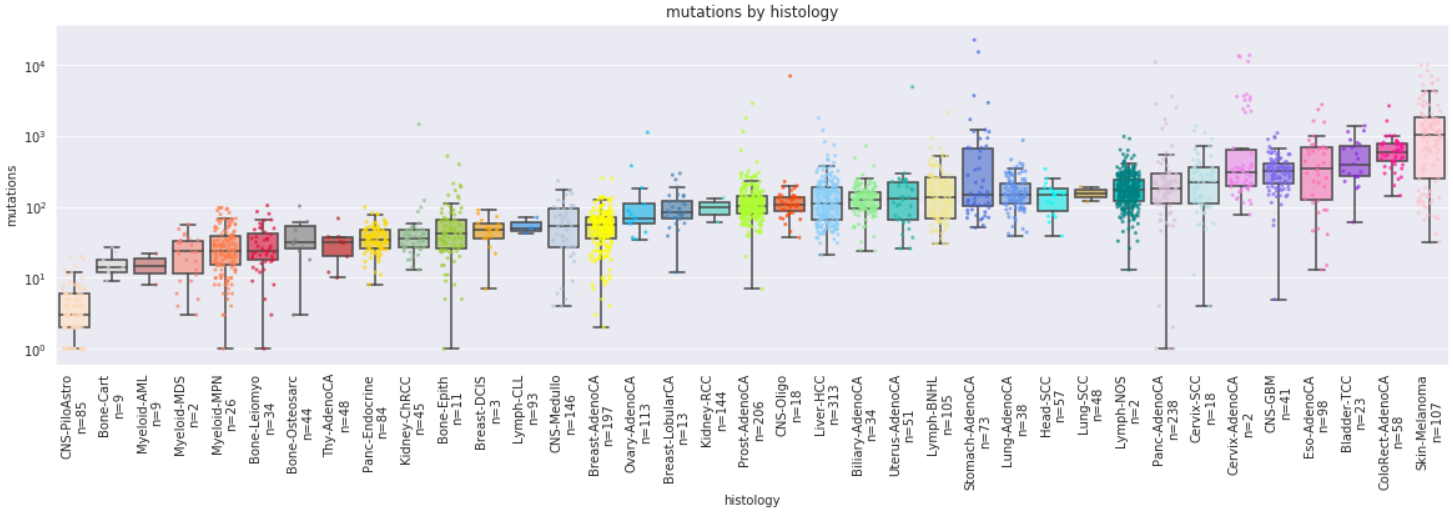
Both datasets had more samples from males than females. The male samples outnumbered female samples by 275 in the TCGA dataset and by 268 in the PCAWG dataset. The number of male and female samples and the mutations in them are shown in Figure 3.3. There were more samples from individuals of European backgrounds than from other ethnicities. In the TCGA dataset I had 6545 samples classified as white and in PCAWG dataset there were 1945 samples classified as European.



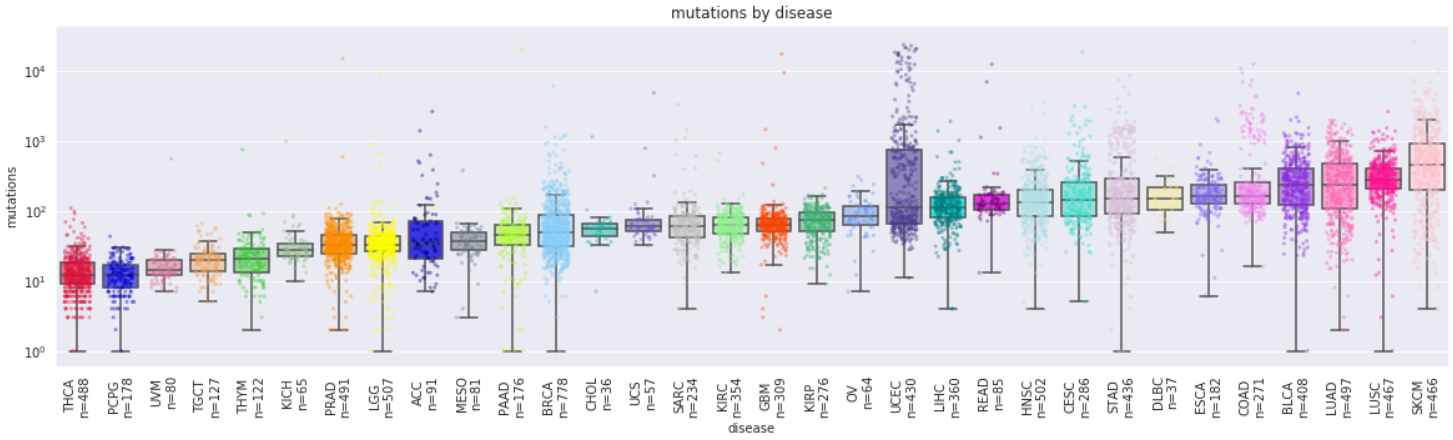
(a) PCAWG mutations

(b) TCGA mutations

Figure 3.3: Number of mutations in in the PCAWG and TCGA dataset age and sex brackets



(a) PCAWG mutations



(b) TCGA mutations

Figure 3.4: Number of mutations in different cancer types in the PCAWG and TCGA datasets

Functional impact and pathogenicity scores

I also used other data and scores meant for analysing the pathogenicity of mutations. The scores I used were SynMICdb, a score for the functional impact of synonymous mutations [46], CADD v. 1.6 is a score for the predicted deleteriousness of mutations [41], Cancer Mutation Census (CMC) contains classification of mutations into different classes based on their pathogenicity [47] and I used CMC data v92 for my analysis, and OncoVar is also a score for the pathogenicity of mutations [52]. OncoVar has data based on different cancer types and datasets, I used pan-cancer TCGA data for my analysis from the OncoVar database. Additionally, I used Ensembl Variant Effect Predictor (VEP) v.100 to annotate the PCAWG and TCGA datasets with the basic options and with CADD v. 1.6. [31].

I used SynMICdb with CMC mutations and first combined the CMC mutations with the PCAWG and TCGA datasets based on genomic position and reference and alternative alleles and then combined the SynMICdb based on the mutation ID available for both CMC and SynMICdb mutations. I compared the SynMICdb score to the DMM p_{any} and p_{true} by calculating the Spearman correlation between them and by making a scatterplot between the different scores. PCAWG had a total of 22323 synonymous mutations for which the SynMICdb score was available and for TCGA there were 114063 synonymous mutations for which SynMICdb score was available. Similarly, to CMC I merged the OncoVar mutations with my PCAWG and TCGA datasets based on the mutation position and alleles for the mutation data and based on the gene for the gene level data. The SynMICdb data is available for download at http://synmicdb.dkfz.de

I downloaded a cancer hotspot v2 mutation dataset from https://www.cancerhotspots.org. The hotspots in this dataset were obtained from 24,592 cancer samples representing more than 300 cancer types. The dataset contains 1165 significant hotspot mutations. 80% of these mutations occurred in 1 in 1,000 or fewer samples [7]. The hotspots were in the downloaded file grouped by codon. I separated the individual nucleotides from the codons for my analysis. For each hotspot mutation I kept its location and reference and alternative alleles. I used both positions only to merge the data to the PCAWG and TCGA datasets.

OncodriveFML input data

OncodriveFML [32] requires multiple different files for it to work. I used OncodriveFML version 2.3. It requires a file with the input mutations, a region file that contains the regions under analysis, and a score file that contains a score for every base within the region file. Additionally, OncodriveFML requires a configuration file for running the program.

The real observed mutations were extracted from the curated PCAWG and TCGA datasets mentioned earlier, by taking the mutations that were found from the skin-melanoma samples (PCAWG), MSI samples (PCAWG), and hepatocellular carcinoma samples (TCGA), and skin cutaneous melanoma samples (TCGA) and written into cancer type specific files. In total I had 107 skin-melanoma samples and 25 potential MSI samples from the PCAWG dataset, 350 hepatocellular carcinomas and 468 skin cutaneous melanomas from TCGA. In the case of the TCGA melanomas there were two more samples than in the mutation dataset as these two did not have any mutations that fit the filtering categories and were included by mistake. The genes I analysed were *TP53, APC, KRAS, BRAF, ATR, TTN, PCLO, BMPR2, GRIN2B*, and *SLC9A8*. Out of these *TP53, APC, KRAS*, and *BRAF*

are known cancer genes. The number of observed mutations per gene in the datasets is shown in the Table 3.1.

Input mutations from the dataset into oncodriveFML				
Gene/Dataset	PCAWG Skin-Melanoma	PCAWG MSI	TCGA hepatocellular carcinoma	TCGA Skin Cutaneous Melanoma
APC	17	22	13	51
ATR	9	12	13	52
BMPR2	17	13	4	19
BRAF	73	7	0	301
GRIN2B	207	40	4	191
KRAS	9	15	4	14
PCLO	238	35	58	723
SLC9A8	10	12	1	9
TP53	23	11	82	77
TTN	676	175	163	2945
Total:	1279	342	342	4382

Table 3.1: Input mutations for OncodriveFML

OncodriveFML also requires a region file which contains the gene regions to be included in the analysis. The genomic locations for the test genes were obtained from BasePlayer [16] genome browser using GRCh37-hg19 human genome assembly. The region file was made by taking the genomic locations and removing the non-genic and non-exonic regions according to DMM and the positions for which no replication timing was available. Positions which were not genic or exonic were dropped as OncodriveFML does not recommend using coding and non-coding regions together in the same analysis. The regions were then defined as being the continuous position left. I also removed the continuous regions that according to BasePlayer genome browser belonged to other genes or genomic elements. The region file contains the chromosome of the continuous stretch of positions and the start and end positions as well as the gene name and a segment number for the separate segments for each gene.

A third important file for the function of OncodriveFML is the score file. Based on this, I created a vcf file containing all reference alleles, from a GRCh37 reference genome, within the regions and all three possible substitutions for each reference allele. Then the vcf file containing fake mutations was annotated with DMM. The annotation adds known

replication timing and information on if the position is genic and exonic for each position. Annotation also adds the sequence surrounding the mutation, sequence length of 2048 was used for the PCAWG mutations while TCGA mutations got sequence length of 256. Positions which were both genic and exonic and had a known replication timing were used in further analysis. The files were then mapped with DMM to obtain the probability outputs for all possible substitutions. Mapping was done individually for all combinations of individual samples and individual genes. After mapping was done for each sample and gene the probabilities for p_{true} and p_{any} were processed into score files. These files were created for each four cancer types individually so that a $-log_{10}$ was applied on every probability. In the case where probability was zero it was replaced by the lowest non-zero probability found in the mapped gene in the sample. After applying the logarithmic on the probabilities, the modified probabilities were summed together and divided by the number of samples to obtain an average score for the cancer type. Separate probability files were made for p_{true} and p_{any} for all cancer types, the file for p_{true} included the reference alleles and the substitution specific probability score in addition to the chromosome and position. The file for p_{any} contained only the chromosome, position, and score for the position without information regarding the substitution.

The configuration file contains the information for OncodriveFML on which signature method to use and how to calculate the statistics. In this analysis the 'complement' method was used, and statistics used arithmetic mean. The minimum number of sampling was set to 100,000 while the maximum was one million. Minimum number of observations was ten while sampling chunks was one hundred million. Multi nucleotide polymorphic mutations were not included in the analysis. When running OncodriveFML the PCAWG skin-melanoma and MSI datasets were given the whole genome setting while TCGA hepatocellular carcinoma and skin cutaneous melanoma were used with the whole exome setting.

3.3 Methods

I first used DMM to map the PCAWG and TCGA sequences to generate mutation probabilities for the mutations 739,341 mutations in the PCAWG dataset and 2,547,192 in the TCGA dataset. The output consists of each mutation having a probability for each mutation type provided by the DMM model used. I also used DMM to annotate the fake substitutions needed to make the score file for OncodriveFML and the regions needed

for the region file. Then I used DMM to map the fake substitutions to get mutation probabilities for all of them.

For visualisation I used Uniform Manifold Approximation and Projection (UMAP) [30]. It is a dimensionality reducing technique that can be used for visualisation and is especially useful for non-linear reductions. UMAP works by finding a topological representation of a lower dimension that is close to the original higher dimensional topological representation. I used UMAP to reduce the dimensionality of the output from DMM to see how the mutation probabilities and sample information cluster in two-dimensional space. I coloured the mutation probabilities with various features to illustrate how the features are distributed in the two-dimensional space.

I also calculated test statistics for different categories by doing Student's T-test for independent samples to see if there were differences between the means of the different categories. For correlations between different features and datasets I used Spearman's correlation coefficient, this way the distribution of the data does not need to be normally distributed. Additionally, I used ordinary least squares regression to fit a linear model on the mutation probabilities and sample features to examine how much of the variance in the results can be explained by the sample features.

Most of the datasets I had I combined by using the chromosomal position and the reference and alternative alleles to the PCAWG and TCGA datasets. In the case of CMC data and the SynMICdb score I used a mutation ID that was available for both in place of the mutation position and alleles. I used VEP to get the CADD score for all mutations.

The probability outputs from DMM can be used as scores in OncodriveFML. I tested the probability scores p_{any} and p_{true} from DMM for the ten genes and compared them to the CADD RAW and CADD PHRED scores. I used in OncodriveFML the files described in the materials section.

4 Results

The first experiments I did with the data was to check how the substitutions I had in the PCAWG and TCGA datasets fell into the six basic substitution categories. I was also interested if there were differences between the substitution types on how many of them were predicted correctly and incorrectly. Other question regarding these basic six substitution types were how they were distributed in the two-dimensional space in relation to each other so I used UMAP to reduce the dimensionality of the DMM probability output vector so that I could visualise them. The visualisation is then based on the predicted probabilities but coloured by the true mutation rather than the predicted ones. I also used UMAP to reduce the dimensionality of the integrated layer vector, which contained sample feature-based values, from DMM and coloured it the same colours as the probability vectors. Additionally, I used boxplots to see if there is a difference between the predicted probabilities for sequences which only contain the necessary input mutation within the sequence and sequences that have at least one additional mutation somewhere in the sequence.

The predicted probabilities from DMM are the most interesting output it has. I mostly focused on the predicted probabilities for any mutation and for the true mutation. I coloured the same UMAPs with these probabilities to see if the probabilities are clustered in some areas of the image. I also calculated the mean probabilities for each sample and compared them to the number of mutations in each sample to see how dependent the DMM models are on the number of mutations when they predict the probabilities.

Since the integrated layer output also contains information not only from the sequence but from the sample too, I coloured the UMAP reductions with the cancer types. This allowed the visualisation of how different cancer types were distributed within the two-dimensional space. Additional sample information I thought important to check was how the cancer type, age, sex, and ancestry affected the prediction of the DMM probabilities. I did this with ordinary least squares regression. I also created boxplots for the distribution of the sample mean DMM predicted probabilities for the cancer types, and for different age and sex groups.

I compared the different pathogenicity and functional impact scores against the DMM probabilities. I did this to see if DMMs probabilities show differences between the likely

pathogenic mutations versus mutations that are not pathogenic or at least are not known to be pathogenic. The hypothesis with DMM is that it should do worse in predicting driver mutations than passenger mutations, because driver mutations happen in unusual contexts and thus DMM will not learn these mutations as well leading to worse predictions. In the comparison images the pathogenic classes should show lower DMM probabilities. Since these scores generally use a higher score to mean more pathogenic there should be negative correlation between DMM probabilities and the scores. I also compared some of the scores against each other to see if they agreed with each other on the pathogenicity of the mutations. SynMICdb and CMC significance tier levels could not be compared against each other as all synonymous mutations in CMC are ranked as 'Other' I did this kind of comparison also with the hotspot mutations as they should also have lower DMM probabilities for the hotspot mutations.

The last experiment I ran was one where I used OncodriveFML to see if it could separate out the known cancer driver genes *APC, BRAF, KRAS* and *TP53* from the other genes I tested. If DMM works well it should be able to do this if the probabilities work well and have differences in the probabilities for the true hotspots in them and lower probabilities for the rest of the gene regions and lower probabilities for the non-cancer genes. I also compared the DMM probabilities against the CADD scores in their ability to separate the genes.

4.1 Exploration of the Deep Mutation Modelling output

Basic information on the substitutions

Based on the reference and alternative alleles I classified all mutations into six basic substitution types and calculated which ones DMM predicted with high accuracy and which ones had lower accuracy. Table 4.1 contains the percentages of the accurately predicted substitutions. A mutation is correctly predicted when the probability of the true mutation p_{true} and the highest predicted probability p_{pred} are the same. Based on the table C→T : G→A substitution is the easiest one to predict correctly. Overall, the accuracy follows the number of mutations where the higher number of mutations means higher accuracy with the only exception being C→G : G→C where it has more mutations that T→G : A→C, but lower accuracy. T→A : A→T on the other hand seems to pose a

challenge for the DMM models as it has the lowest percentage of correctly classified cases. Table 4.2 shows that overall the p_{true} probabilities are quite low for the substitutions, but C→T : G→A sticks out as one with higher chance of being predicted as true. As in Table 4.1, Table 4.2 shows that T→A : A→T has the lowest mean probabilities of it being predicted correctly. Table 4.2 also shows the mean probability of any mutation and it shows that the probabilities are much higher than for any of the true mutations, but the probabilities for any mutation re also much more uniform with less differences between the substitutions. Overall, the DMM model's predictions follow the number of substitutions and in the case of p_{any} the predictions between different substitutions are quite uniform.

Percentage of correctly predicted substitutions and the number of substitutions						
Substitution	PCAWG correct	TCGA correct	PCAWG subs.	PCAWG %	TCGA subs.	TCGA %
C→A : G→T	42.5	55.8	105,955	14.3	465,080	18.3
C→G : G→C	13.6	33.4	56,098	7.6	187,775	7.4
C→T : G→A	74.7	73.9	400,022	54.1	1,397,337	54.9
T→A : A→T	11.8	19.4	41,166	5.6	100,705	4.0
T→C : A→G	36.6	53.7	91,532	12.4	271,757	10.7
T→G : A→C	31.3	43.6	44,568	6.0	124,538	4.9

Table 4.1: Percentages of correctly predicted substitutions and the number and percentage of the substitutions

Mean p_{any} and p_{true} for the substitutions				
Substitution	PCAWG p_{true}	PCAWG p_{any}	TCGA p_{true}	TCGA p_{any}
C→A : G→T	0.34	0.83	0.42	0.70
C→G : G→C	0.14	0.77	0.28	0.70
C→T : G→A	0.64	0.87	0.58	0.71
T→A : A→T	0.13	0.82	0.21	0.65
T→C : A→G	0.25	0.84	0.39	0.63
T→G : A→C	0.25	0.86	0.32	0.68

Table 4.2: Mean predicted probabilities for the substitutions

The input sequences in the analysis for DMM contain at least one mutation in the centre of the input sequence, but the sequence may contain more mutations elsewhere in the sequence. The effect of these mutations according to Figure 4.1 is that the extra mutations increase DMM's ability to predict either any mutation or the true mutation. The effect

is less apparent in the TCGA dataset. Based on Student's T-test all of the four different subfigures in Figure 4.1 had statistically significant differences as they all had p-values close to zero and the test scores were approximately -205.74, -183.25, -46.12, and -61.83.

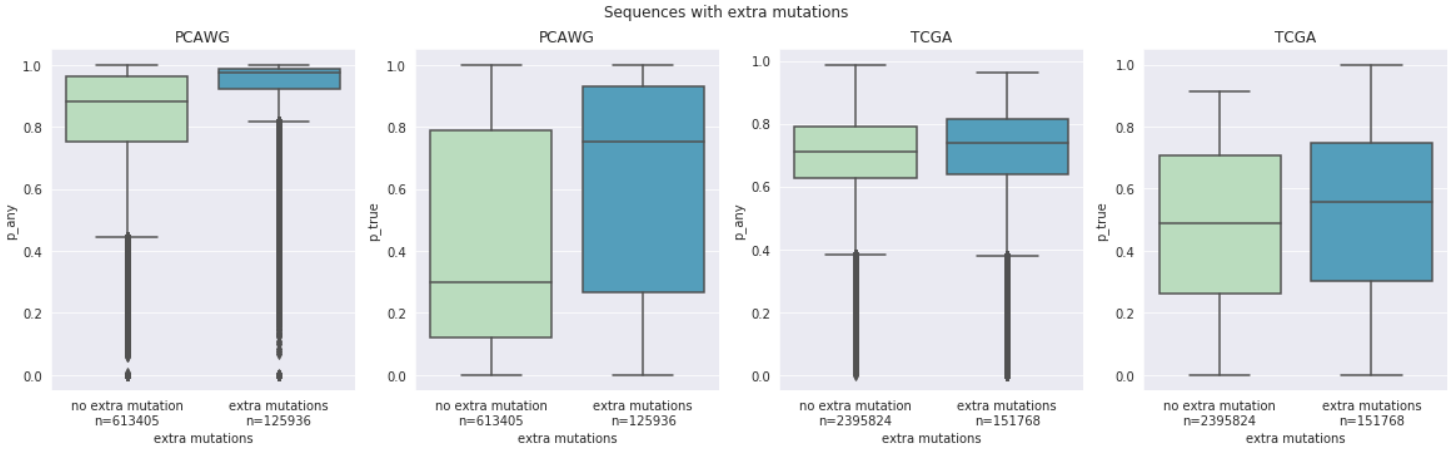
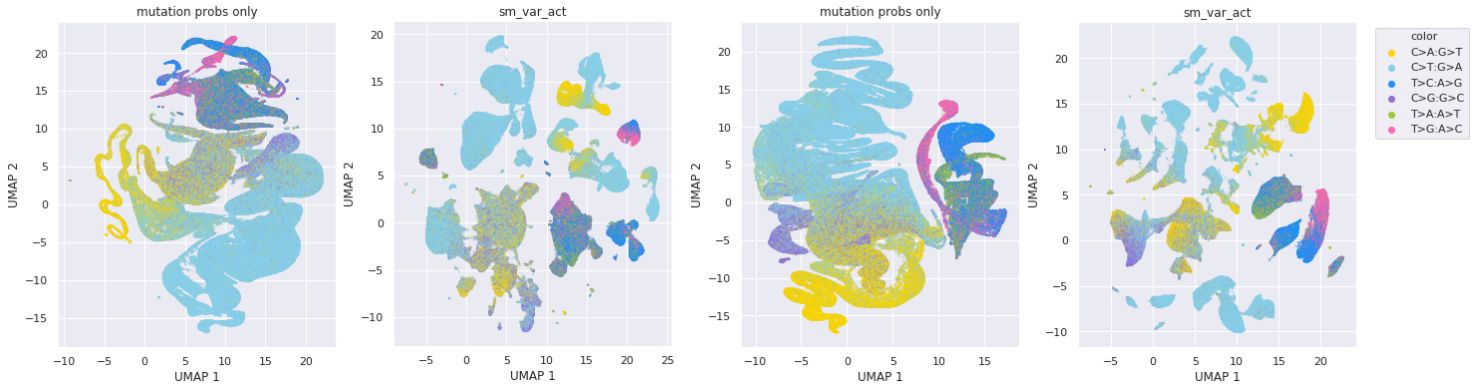


Figure 4.1: Effect of additional mutations in sequence on p_{any} and p_{true}

For basic visualisation of the probability vectors of the different mutations and the combined sample and sequence integrated feature vectors I used UMAP to reduce the dimensionality to two. In Figure 4.2 the individual points are made out of the probability vectors of the mutations so that the location of an individual point in the image is based on the probability vector of a mutation. The colour on the other hand is the true mutation of the same point. Based on Figure 4.2 the probability vectors and the UMAP reduction separate the six different substitutions into different regions within the UMAP reduction. C→T : G→A substitutions clearly form their own large regions within the plots.

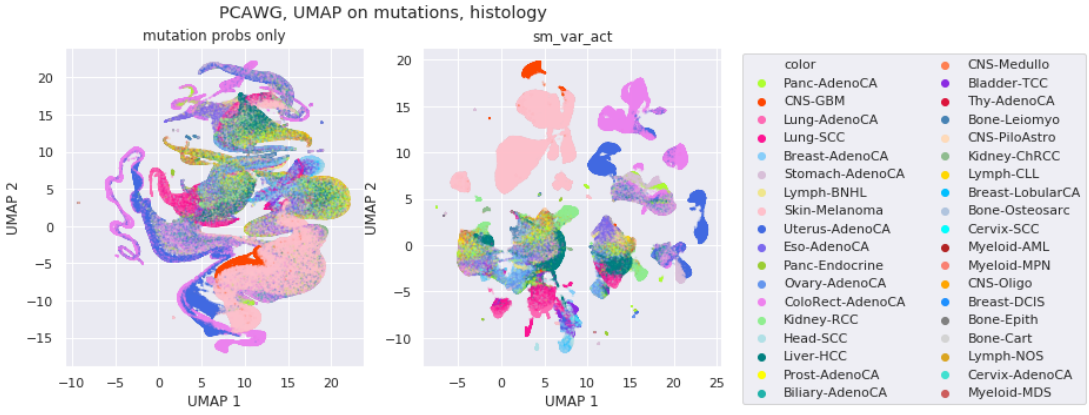


(a) PCAWG **(b)** TCGA

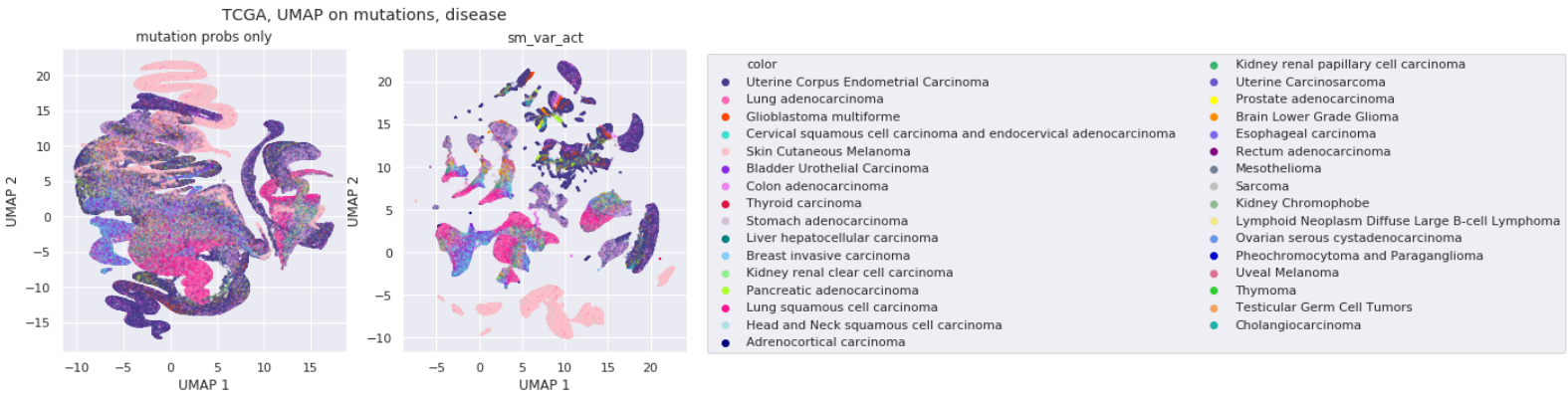
Figure 4.2: UMAP of the PCAWG and TCGA probability and feature vectors, colorised with the substitution type

Histology distributions

The clusters in Figure 4.2 sample and sequence feature vector, right-hand-side subfigures, UMAP reduction become clearer when they are coloured by cancer types as the different cancer types are distributed in many cases into their own clusters while others share clusters. For example melanoma mutations form their own very distinct clusters and these clusters mostly overlap with C→T : G→A substitutions as can be observed when comparing Figures 4.2 and 4.3.



(a) PCAWG



(b) TCGA

Figure 4.3: UMAP of the PCAWG and TCGA probability and feature vectors, histology

The separation of melanomas into four distinct clusters in the Figure 4.3a right-hand-side subfigure is mostly based on the substitution type present in the sequence. Two of the clusters have mostly C→T : G→A substitutions alongside with a few other C base reference substitutions while the other two clusters have the T reference substitutions as can be seen in Figure 4.4. Otherwise, the clusters shared features so that clusters 5 and 12 had similar features and clusters 11 and 16 had similar features.

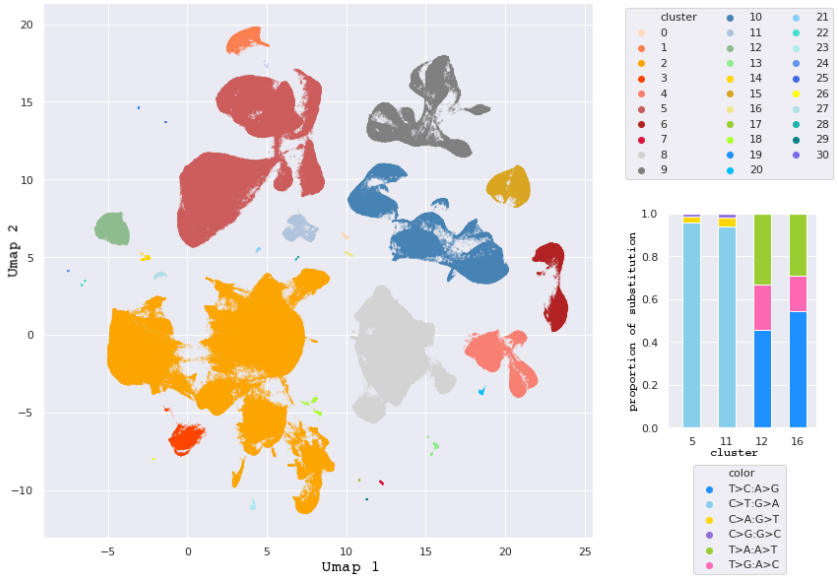
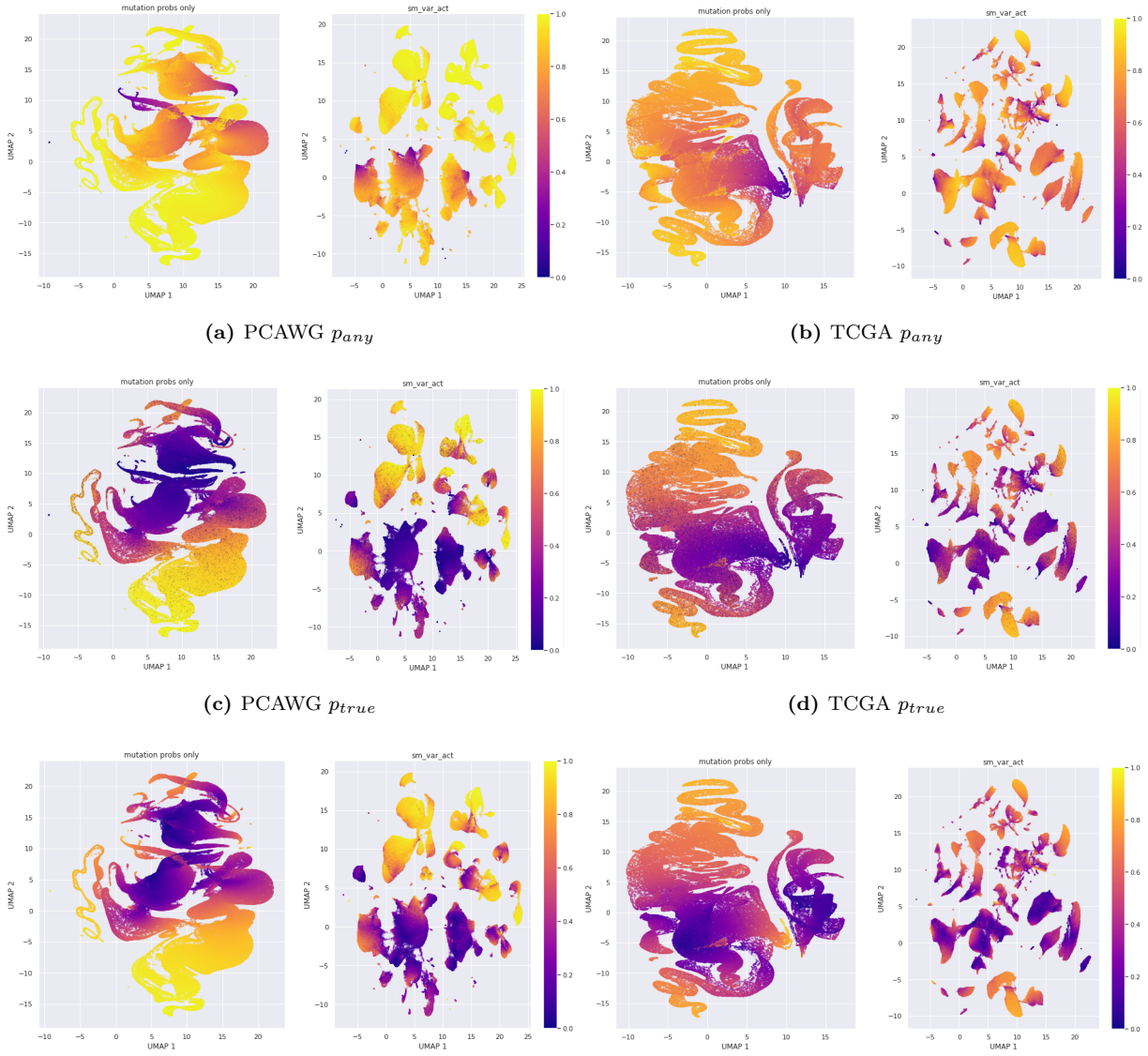


Figure 4.4: Clustering of the PCAWG data and substitutions in melanoma clusters

Probability outputs

Using the predicted values of p_{any}, p_{true}, and p_{pred} to colour the UMAPs results in the high probabilities clustering together in the p_{true} images as seen in Figure 4.5. When comparing the images in Figure 4.5 to the real substitutions in Figure 4.2 we can see that the p_{true} has the highest values in the same regions where the C→T : G→A and C→A : G→T substitutions are. p_{any} on the other hand means that any mutation is possible in this position. Based on the p_{any} Figures 4.5a and 4.5b the interesting areas are the wedge shaped regions in the centre of the subfigures on the left as the DMM model predicts that in these areas no mutation has happened. The p_{pred} and p_{true} images for the most part look similar in Figure 4.5.

Figure 4.6 shows the sample means of p_{any} and p_{true} in PCAWG and TCGA datasets. Each dot is one sample and the colours are the same cancer types as in Figure 4.3. In Figure 4.6 we can see that the TCGA model is quite flat in the p_{any} subplot and even the p_{true} is not very different between different cancer types. The only exception is the group that contains skin cutaneous melanomas (light pink) which hover above the rest of the samples. The PCAWG model shows heavy dependence on the number of mutations in the p_{any} subplot of Figure 4.6. In this subplot we can clearly see the effect of the number of mutations on p_{any} so that the higher number of mutations there are the higher the probability of p_{any} is. The p_{true} on the other hand while dependent on the number of mutations does not rely on it as much. The interesting point in this whole figure is that



(a) PCAWG p_{any} **(b)** TCGA p_{any}

(c) PCAWG p_{true} **(d)** TCGA p_{true}

(e) PCAWG p_{pred} **(f)** TCGA p_{pred}

Figure 4.5: UMAP of the PCAWG and TCGA probability and feature vectors, probabilities

TCGA for the most part is flat while PCAWG shows that it is more dependent on the number of mutations.

Figure 4.7 shows the mean p_{any} and mean p_{true} for samples in the PCAWG dataset. Based on this figure the PCAWG model seems to work the best for melanoma samples as melanomas have the highest mean p_{any} and p_{true}. Based on Figure 4.7 the model may work well for melanomas from PCAWG, but many of the other cancer have very low p_{true} probabilities. I made a similar image for TCGA but like in Figure 4.6 the model looked flat and like in PCAWG the melanomas had highest p_{any} and p_{true} probabilities.

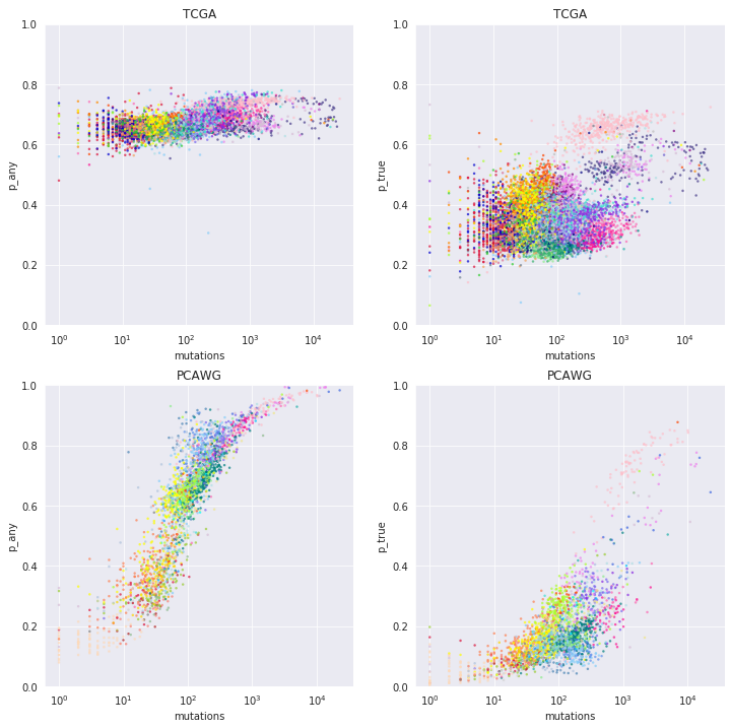


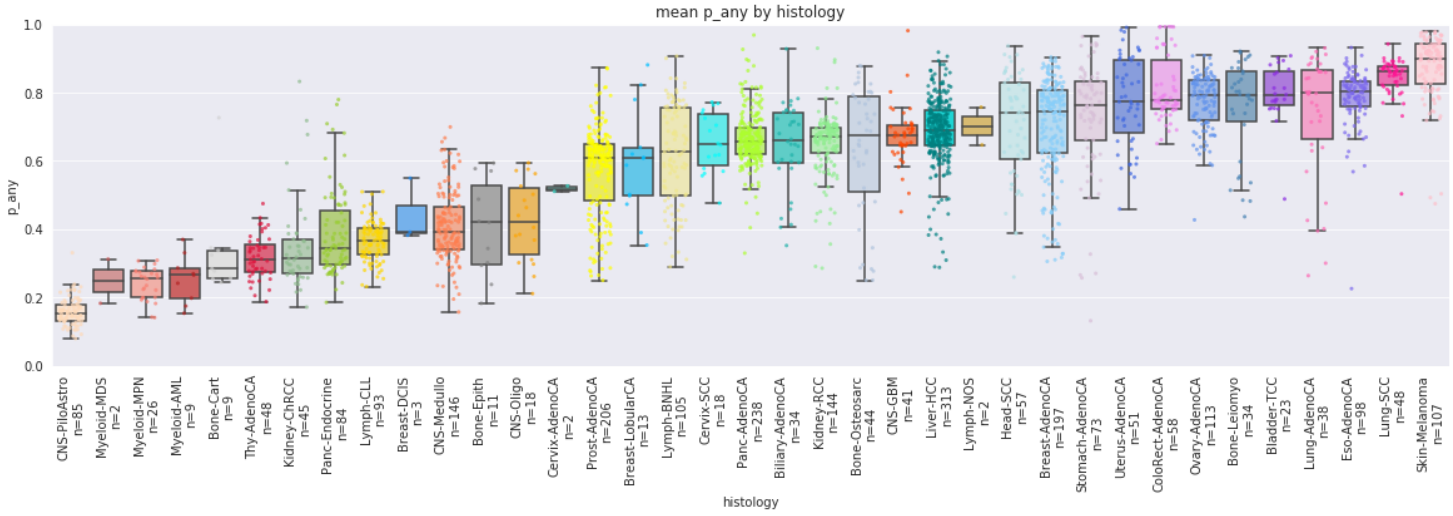
Figure 4.6: Mean p_{any} and p_{true} in samples compared to the number of mutations in the samples in TCGA and PCAWG, the colours are the same as in the cancer types

Sex differences in the probability outputs

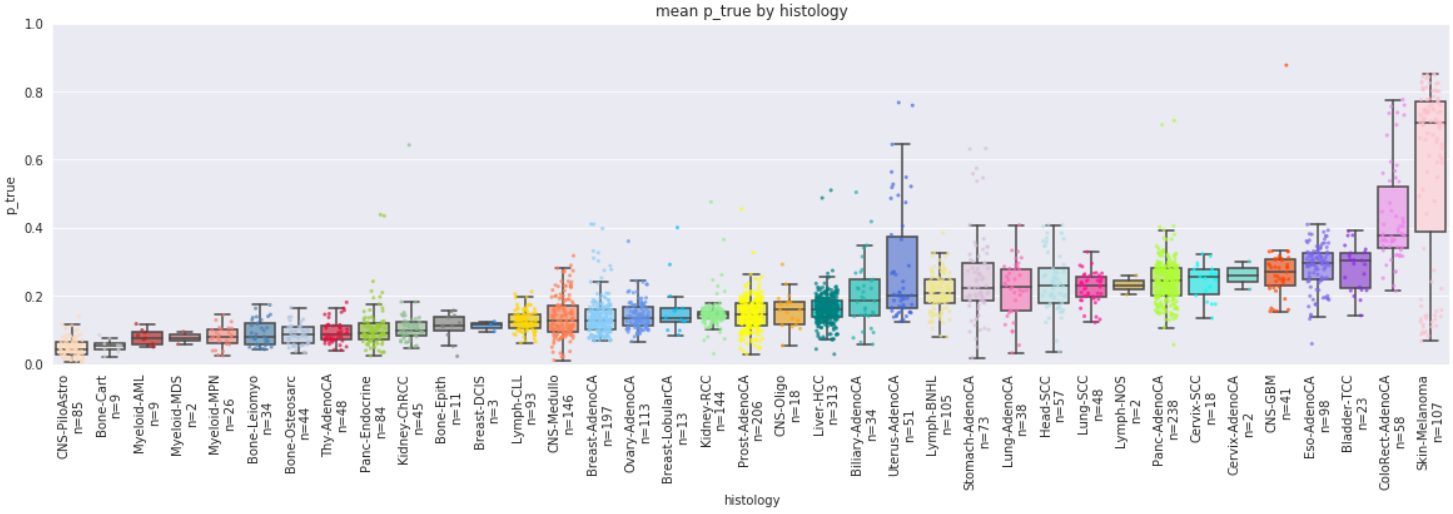
Figure 4.8 shows sample mean p_{any} and p_{true} probabilities for the PCAWG and TCGA datasets by sex and age group. Figure 4.8a shows that in the PCAWG model the probabilities become higher with age, except for the over 80 groups where it decreases a little again. This increase in prediction ability is especially notable in p_{any}. The TCGA model on the other hand looks flat again, but the mean probabilities are higher than in the PCAWG. The TCGA probabilities on the other hand show a smaller range of probabilities than the PCAWG model. The TCGA mean p_{any} are between 0.6 and 0.8 while in the PCAWG there are samples where the mean p_{any} reaches nearly 1. In the p_{true} subfigures the outlier samples that have a mean probability higher than 0.6 are mostly melanomas in both the PCAWG and the TCGA subfigures.

Ordinary least squares for sample features

I performed an OLS regression on the data where my dependent value was either p_{any} or p_{true} and the independent values were the age and sex of the patient, the ancestry of the patient, and the cancer type. Based on the OLS analysis the R^2 statistic varied widely



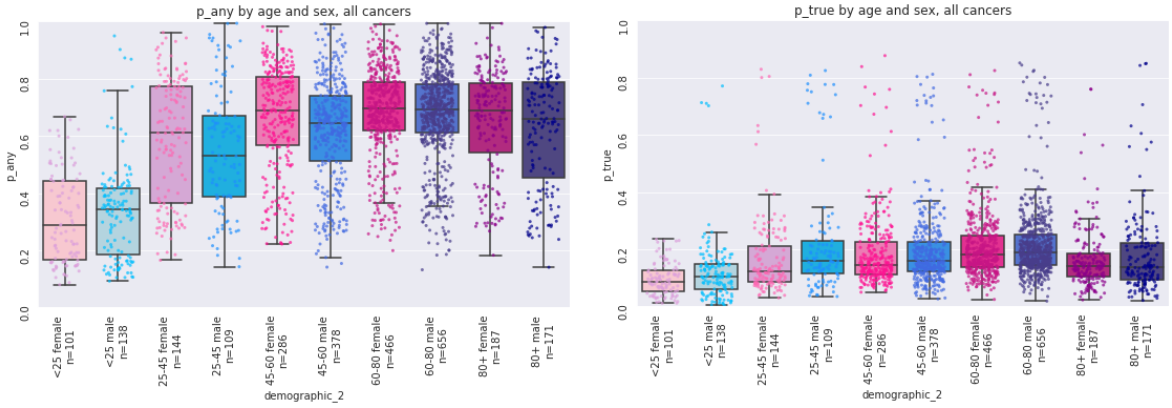
(a) PCAWG p_{any}



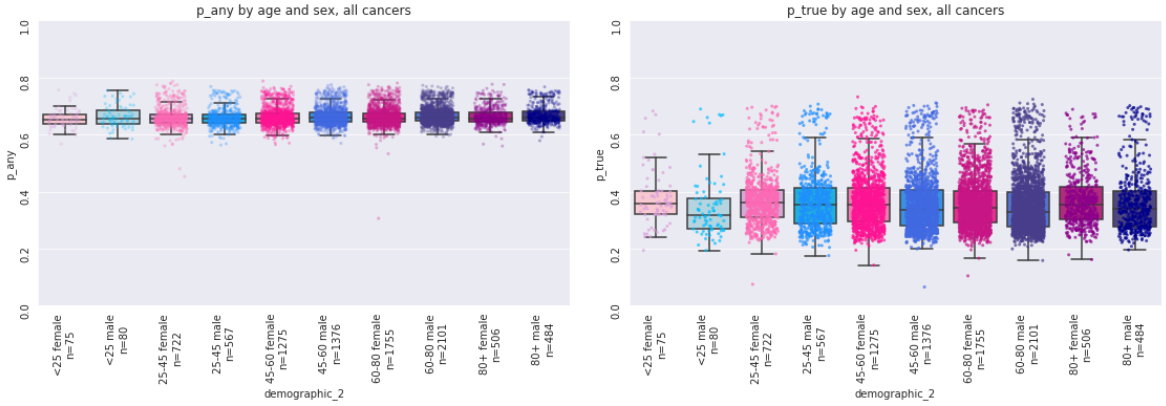
(b) PCAWG p_{true}

Figure 4.7: PCAWG histology

between the TCGA and PCAWG datasets. For TCGA p_{any} the R^2 was 0.04 and for p_{true} it was 0.24. For the PCAWG data the R^2 was 0.56 for p_{any} and for p_{true} it was 0.49. In all cases male sex was associated with slight increase in the probabilities but even the highest coefficient for male sex was only 0.007 in the PCAWG p_{true} model. The various cancer types had different coefficients, but skin melanoma in all models had a positive coefficient. For the p_{true} TCGA model it was 0.28 and for p_{any} it was 0.05. For the PCAWG p_{true} model it was 0.38 and for the p_{any} it was 0.12. Age had different effects in the PCAWG and TCGA regressions. In PCAWG age had a low but positive coefficient, p_{true} 0.0001 and p_{any} 0.0005, but in TCGA the coefficients were negative but still low, p_{true} -0.0004 and p_{any} -0.0003.



(a) PCAWG mutations



(b) TCGA mutations

Figure 4.8: Probabilities in the PCAWG and TCGA dataset age and sex brackets

4.2 Functional impact and pathogenicity scores

CMC significance tiers

I was also interested in how the CMC mutation significance tier affects the p_{any} and p_{true} values in the PCAWG and TCGA datasets. In both cases I combined the mutations in the two different datasets based on the position and reference and alternative alleles. Mutations that were present in the TCGA and PCAWG datasets but had no CMC information were marked as 'NaN' mutations. The results are shown in Figure 4.9. In Figure 4.9a it is visible that tier 1 has lower p_{any} and p_{pred} probabilities than the other tiers while in the p_{true} subfigure tiers 1,2, and 3 have closer medians, but the bulk of the tier 1 mutations still have lower probabilities. According to Student's T-test all significance triers differed significantly from the 'Other' category. Between tiers 1,2, and 3 the differences were smaller. In the p_{any} probabilities the test statistic was -2.80 and p-value 0.005 for tiers 1 and 2

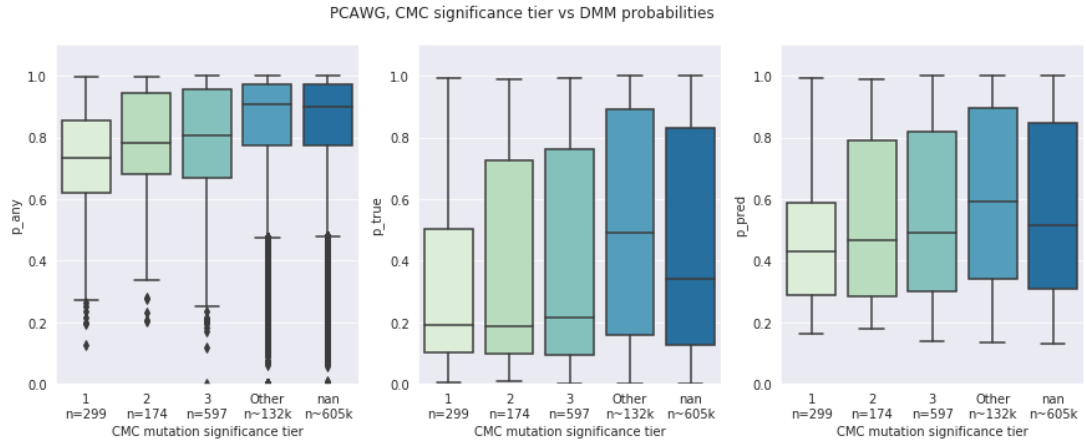
while between tiers 1 and 3 the test statistic and p-value were -4.07 and 5.183e-05. For p_{any} there was no significant differences between tiers 2 and 3. There we no significant differences between tiers 1 and 2 or between tiers 2 and 3 in the p_{true} probabilities. There were some differences between tiers 1 and 3 in the p_{true} probability and for that the test statistic and p-value were -3.15 and 0.002, this difference in the case of p_{pred} was -4.15 and 3.728e-05. For p_{pred} and tiers 1 and 2 the test statistic was -2.29 and the p-value was 0.023.

In the TCGA subfigure 4.9b most of the tiers have similar looking medians and the differences do not look as great as the ones in the PCAWG subfigure. Again, in all cases all tiers were statistically significantly different from the 'Other' group. In the case of p_{any} tiers 1 and 2 the test statistic was 2.95 and the p-value was 0.003, for tiers 1 and 3 the test statistic was -3.02 and p-value was 0.003. For tiers 2 and 3 the test statistic was -5.76 and the p-value was 3.710e-20. For p_{true} the T-test gave the following values for tiers 1 and 2: test statistic was 3.68 and p-value 0.0002 and for tiers 2 and 3 the test statistic was -2.68 and p-value 0.007. Tiers 1 and 3 had no significant p-value. In the case of p_{pred} tiers 1 and 2 the test statistic was -2.78 and p-value 0.006. For tiers 1 and 3 it was -7.36 and 2.116e-13. For tiers 2 and 3 it was -2.99 and 0.003.

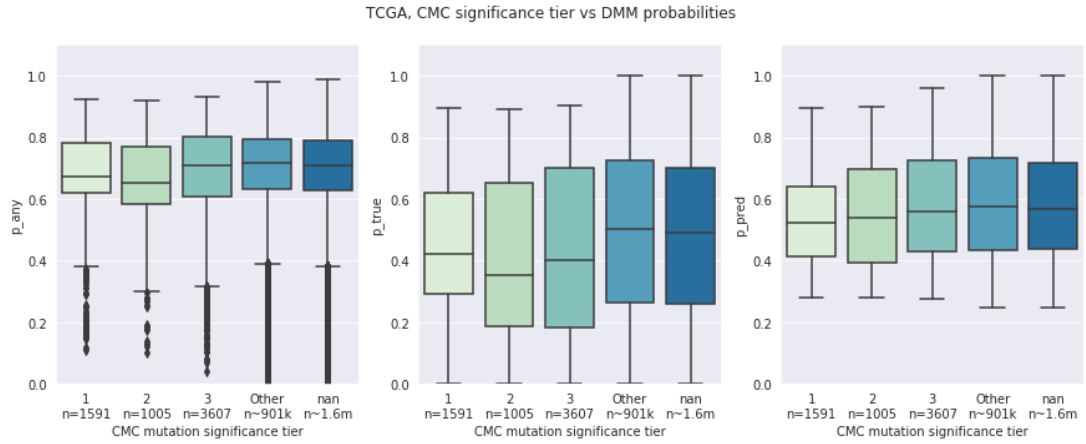
SynMICdb

Figure 4.10 shows the scatterplots between the SynMICdb score and the DMM probabilities p_{any}, p_{true} and p_{pred}. The figures also show a regression model fitted on the data. The figures in Figure 4.10 show that there exists negative correlation between the DMM output and the SynMICdb score. PCAWG shows stronger correlation in Figure 4.10a than TCGA does in Figure 4.10b. The Spearman correlation between the PCAWG p_{any}, p_{true}, p_{pred}, and SynMICdb score were -0.45, -0.49, and -0.46 and they all had p-values close to zero. For TCGA the same correlations were -0.08, -0.25, and -0.25 and the p-values were close to zero.

Figure 4.11 shows the mutation probability vector UMAPs coloured by the SynMICdb score values. Most of the higher scores cluster in the figures within the wedge shaped regions in the middle of the figures. When comparing these UMAPs to Figures 4.5 it is visible that these regions are the same ones where the predicted probabilities are lower.



(a) PCAWG

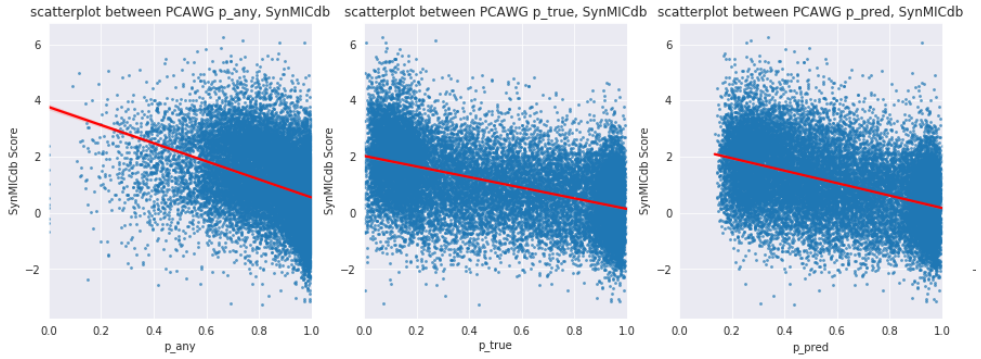


(b) TCGA

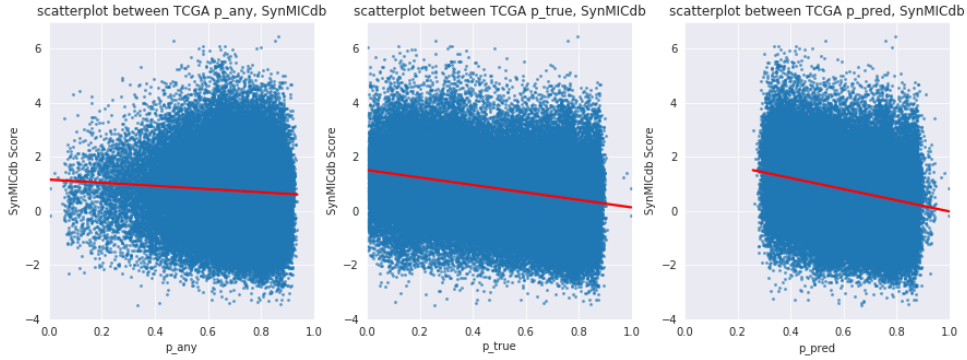
Figure 4.9: Probabilities in the PCAWG and TCGA dataset by CMC significance tiers

CADD

I calculated the Spearman's correlation also between p_{any} and p_{true} for CADD scores. For the PCAWG dataset the correlation between p_{any} and CADD RAW score was -0.02 and p-value 3.646e-47 and for p_{true} they were -0.02 and p-value 3.012e-38. For the TCGA dataset the correlation between p_{any} was 0.02 and the p-value was 3.185e-130 and for p_{true} it was -0.01 and the p-value was 1.338e-40. The values for CADD PHRED score did not differ from the values for the RAW score in any significant way and they were almost identical.



(a) PCAWG



(b) TCGA mutations

Figure 4.10: Probabilities in the PCAWG and TCGA dataset vs SynMICdb score

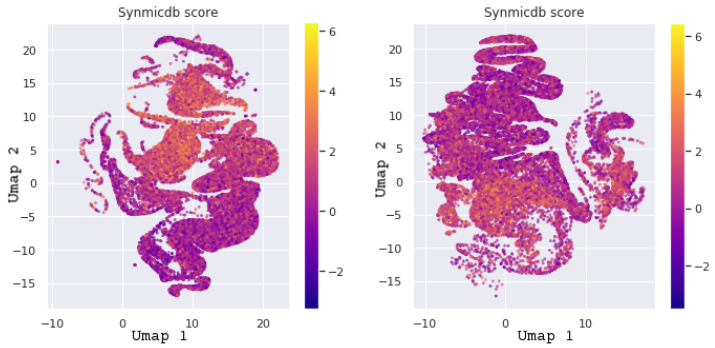


Figure 4.11: UMAP of PCAWG and TCGA colored by SynMICdb score

OncoVar

I also compared the DMM outputs to OncoVar scores. The Spearman's correlation between OncoVar mutation level score for PCAWG p_{any} and p_{true} were -0.01 with p-value 0.0004 and -0.001 and p-value 0.841. For TCGA the correlations were for p_{any} 0.001 and p-value 0.854 and for p_{true} -0.01 and p-value 0.058. For gene level OncoVar score and for PCAWG gene level mean p_{any} the Spearman's correlation was 0.04 and the p-value was 4.065e-08

and for mean p_{true} they were -0.01 and 0.245. For TCGA the correlation for mean p_{any} was -0.01 and the p-value was 0.061 and for mean p_{true} the correlation was 0.09 and the p-value was 4.093e-36. OncoVar also features a consensus score, and I computed the Spearman's correlation between it and the gene level mean p_{any} and p_{true} probabilities. For PCAWG the correlation for mean p_{any} was 0.01 and p-value was 0.324 and for p_{true} they were 0.01 and 0.261, respectively. For TCGA the correlation for mean p_{any} was -0.03 and p-value was 4.098e-06 and for mean p_{true} they were -0.07 and 3.219e-20.

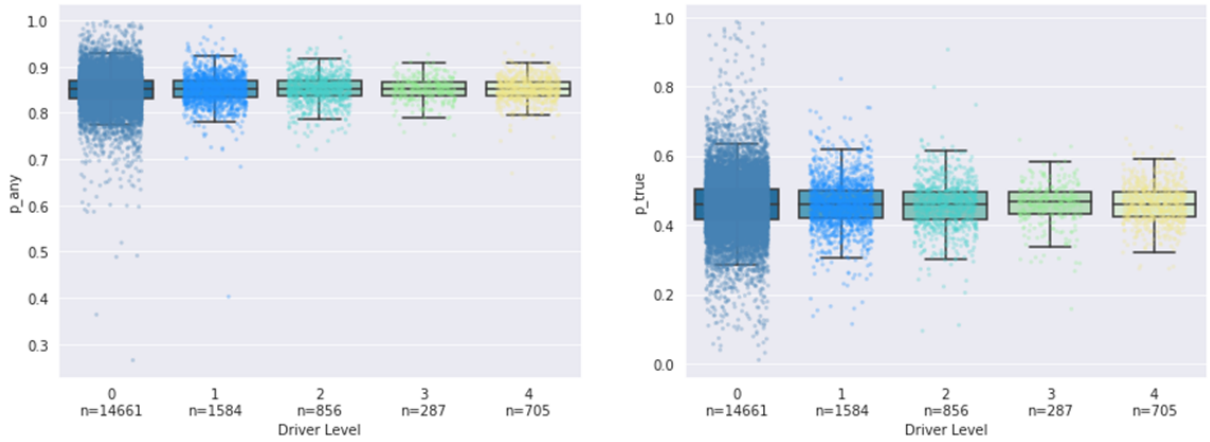
Figure 4.12 shows the OncoVar driver levels with the mean DMM probability outputs per gene. In both the PCAWG and TCGA subfigures there is not much visible differences between the different driver levels in the median values for the probabilities. Level 4 is the highest level and genes in this category are classified as drivers.

CADD and OncoVar in CMC significance tiers

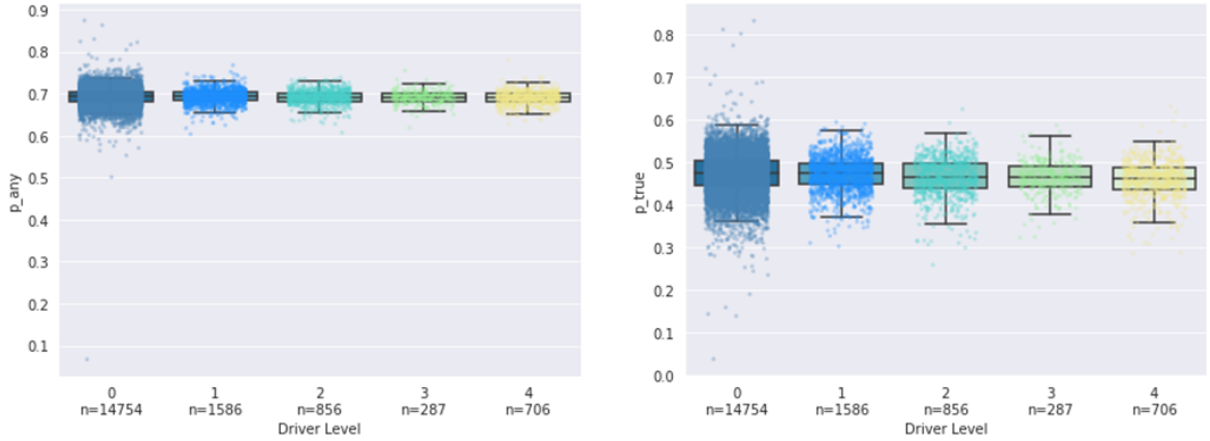
For comparison's sake to the DMM probabilities I also made boxplots for the CMC significance tiers and instead of the DMM probabilities I used the CADD PHRED and OncoVar scores. Both had higher median values for the CMC tiers that have some evidence of them being pathogenic compared to the tier that has no known evidence of being pathogenic. The effect can be seen in Figures 4.13 and 4.14.

4.3 Score levels of mutation hotspots

I also analysed the DMM p_{any} and p_{true} outputs in relation with known cancer hotspots. Figure 4.15 shows the p_{any} and p_{true} values for the known hotspot mutations and mutations which hotspot status is not known. In both the PCAWG and TCGA subfigures it is visible that hotspot mutations have lower median probabilities than the mutations for which hotspot status is not known. Figure 4.16 shows the hotspot mutations in four known cancer genes *TP53, KRAS, APC*, and *BRAF* and their probabilities. In this case the hotspots are defined simply by the chromosome and position of the mutation. In the PCAWG subfigures it is visible that the mutations in known hotspot positions have lower p_{any} than the mutations in not hotspot positions. The same is visible for p_{true} except for *TP53* where the probability is higher for known hotspot mutations. In Figure 4.16a the genes that have somewhat statistically significant differences between the hotspot and unkown categories are *TP53* and *KRAS*. Their test statistics and p-values from a T-test



(a) OncoVar driver levels, PCAWG

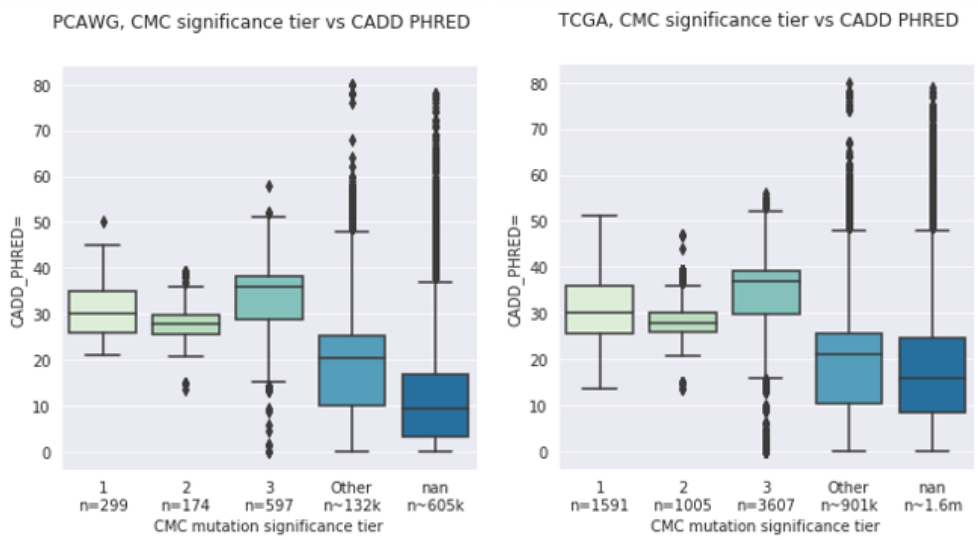


(b) OncoVar driver levels, TCGA

Figure 4.12: Mean probabilities in the PCAWG and TCGA genes vs OncoVar driver level

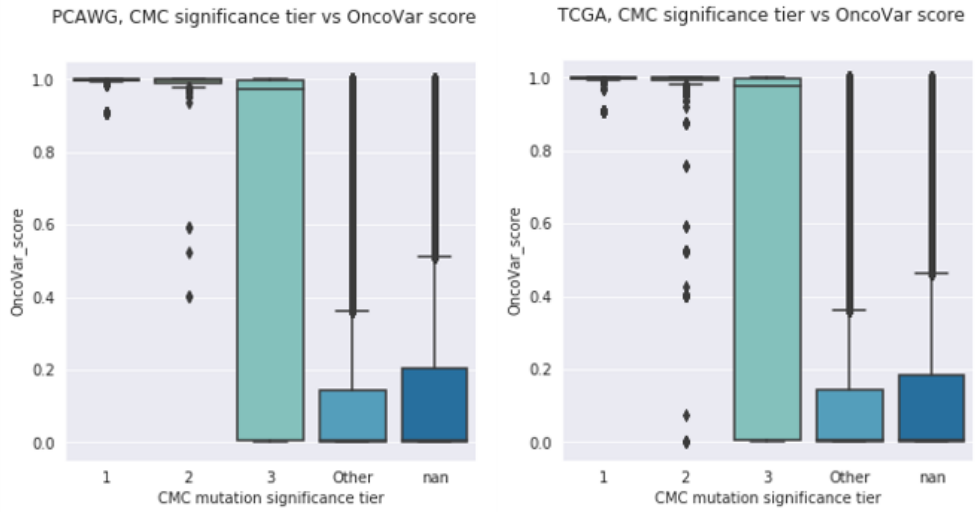
are -2.40 and 0.017 for *TP53* and -2.03 and 0.043 for *KRAS*. In the TCGA p_{any} subfigure the median between the known and unknown hotspot mutations in different genes is more variable. In *TP53* the median is roughly similar. In *KRAS* and *BRAF* on the other hand the p_{any} is higher for known hotpot mutations while in *APC* it is lower for known hotpot mutations than for the unknown ones. For the p_{true} *TP53* is the only gene for which the median p_{true} is higher for the known hotspot mutations than the unknown mutations. For the other three genes the known hotpot mutations have lower median p_{true}. The test statistic from T-tests indicated that for TCGA the differences were significant for the genes in the p_{true} case and in p_{any} only *APC* and *KRAS* are significant.

Additionally, I did this with CADD scores to see if they behave like expected with the



(a) CMC significance tiers, PCAWG **(b)** CMC significance tiers, TCGA

Figure 4.13: CADD PHRED in CMC significance tiers for PCAWG and TCGA



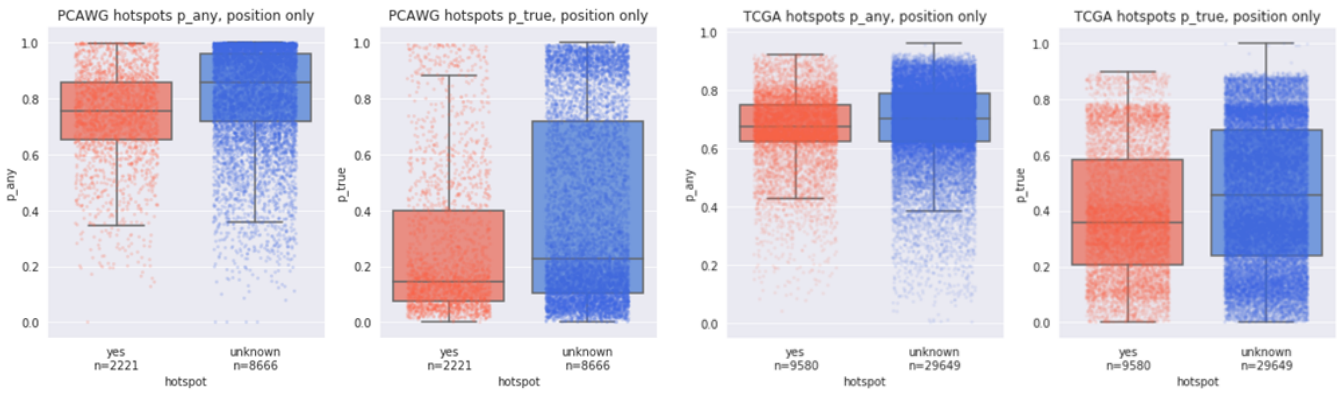
(a) CMC significance tiers, PCAWG **(b)** CMC significance tiers, TCGA

Figure 4.14: OncoVar score in CMC significance tiers for PCAWG and TCGA

hotspots too. The results can be seen in Figure 4.17 and in both subfigures the median is higher for the mutations in the hotspots than for the unknown mutations.

4.4 OncodriveFML experiment

Figure 4.18 shows the output of the OncodriveFML for all of the scores and genes for the PCAWG melanoma samples. From the CADD scores in Figures 4.18c and 4.18d it is clear



(a) PCAWG hotspots

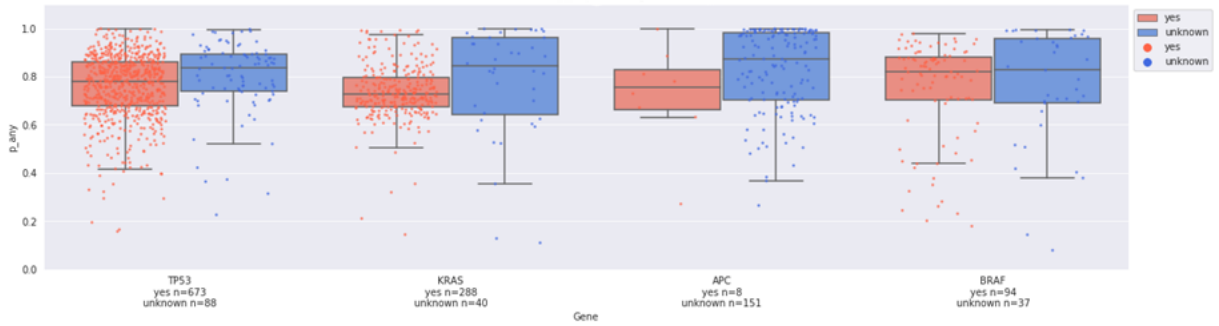
(b) TCGA hotspots

Figure 4.15: Hotspots in the PCAWG and TCGA datasets

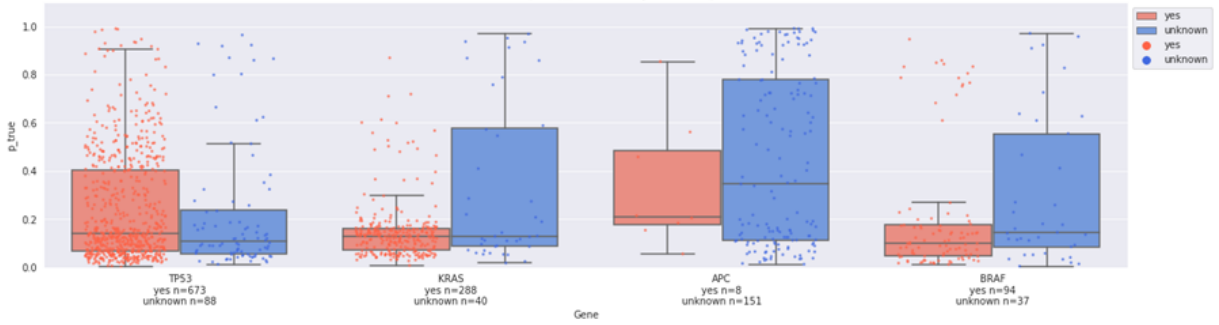
that the genes *TP53* and *BRAF* have achieved statistical significance as their p-values, both 1e-06, are the smallest possible ones. *APC* and *KRAS* have also small p-values, in Figure 4.18c 0.037 and 0.042 respectively and in Figure 4.18d 0.034 and 0.028 respectively. In contrast in 4.18b p_{true} only shows *BRAF* achieving significance with p-value being the smallest possible one, 1e-06. *TP53* also shows some degree of significance as its p-value is 0.024. In 4.18a however, we can see p-value inflation as all of the datapoints float above the central diagonal line. The diagonal line represents data that is equally distributed. In this case the p-values for *TTN, PCLO* and *TP53* are 0.003, 0.005, and 0.007 respectively.

In Figure 4.19 it is again visible that the CADD scores achieve much better results as they are able to separate the known cancer genes from the rest of the genes. *KRAS* especially looks like it has very small p-values in the CADD score subplots. The p-values for *APC, BRAF, KRAS* and *TP53* in Figure 4.19c are 0.055, 0.016, 2e-05, and 0.011 respectively. In Figure 4.19d they are 0.019, 0.031, 4.9e-05, and 0.005 respectively. Meanwhile the p_{any} does not result in any gene being considered significant with the exception of *TTN*, which has a p-value 0.026. In p_{true} subfigure *BRAF* is again considered significant even if not as much as *KRAS* in the CADD subfigures. *BRAF* however does have smaller p-values in the p_{true} subfigure than in the CADD subfigures. The p-values in Figure 4.19b for *BRAF, TP53, ATR, TTN* and *KRAS* are 0.003, 0.043, 0.050, 0.057, and 0.070.

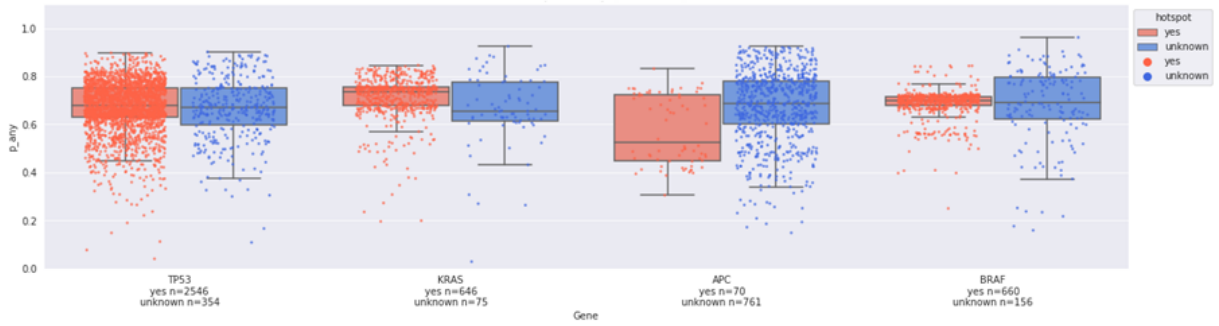
In Figure 4.20 CADD scores achieve good success but in this case neither of the DMM scores shows any gene as significant. In the CADD subfigures *TP53* is the gene that achieves smallest p-values. Interestingly *GRIN2B* is also supposedly significant in the CADD subfigures. In Figure 4.20c the p-values for *TP53, GRIN2B, KRAS* and *APC* are 1e-06, 0.0003, 0.001, and 0.078. In Figure 4.20d they are 1e-06, 0.001, 0.001, and 0.042



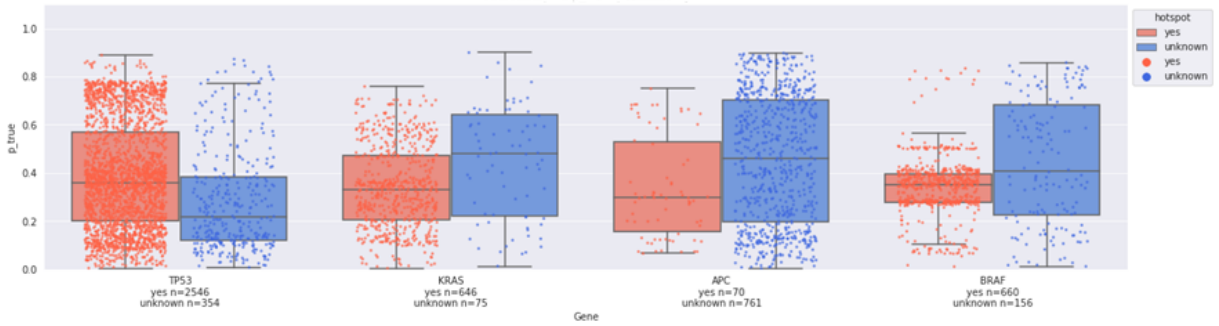
(a) PCAWG hotspots p_{any}



(b) PCAWG hotspots p_{true}



(c) TCGA hotspots p_{any}



(d) TCGA hotspots p_{true}

Figure 4.16: Hotspots in four cancer genes in the PCAWG and TCGA datasets

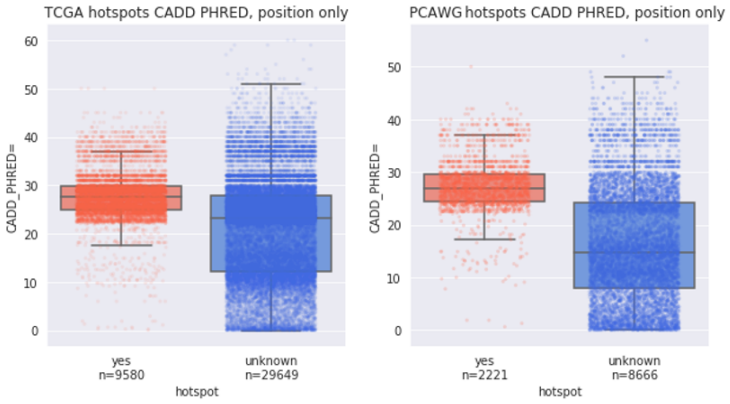
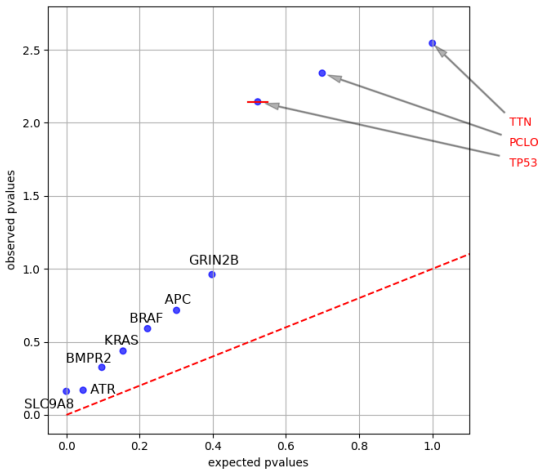


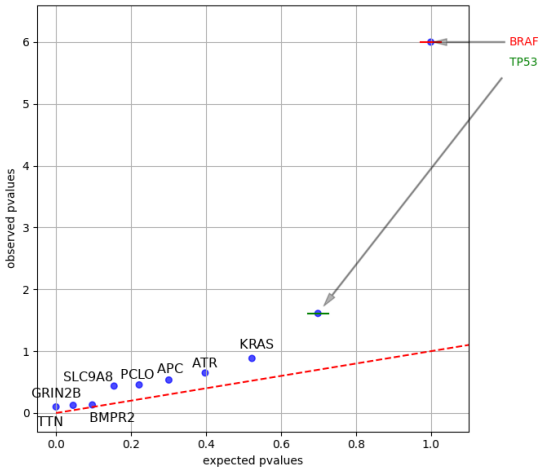
Figure 4.17: Hotspots with CADD PHRED score for TCGA and PCAWG

respectively.

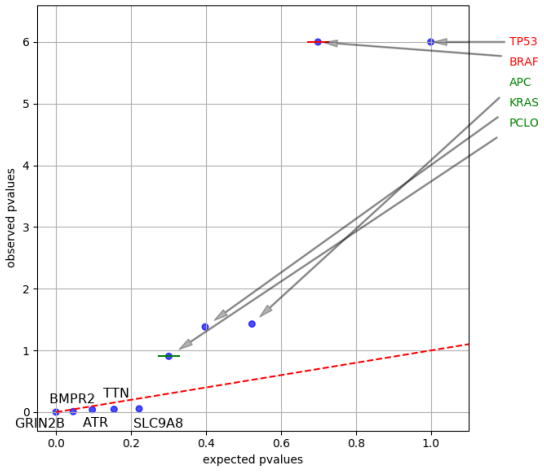
Figure 4.21 shows that in the CADD score subfigures almost all genes supposedly have significant p-values with the exceptions being *TTN* and *ATR*. In these figures *TP53, PCLO, KRAS, GRIN2B* and *BRAF* all have the minimum p-value 1e-06. In Figure 4.21a three of the known cancer genes *APC, KRAS*, and *TP53* separate above the main bulk of the genes even if they do not have very significant p-values. Their p-values are in this figure 0.067, 0.032, and 0.016. In Figure 4.21b all of the known four cancer genes have separated themselves from the main bulk of the genes and *BRAF* achieves very good p-values while *KRAS* has somewhat worse but still okay looking p-value. The p-values for *BRAF, KRAS, APC* and *TP53* in this subfigure are 1e-06, 0.0001, 0.013, and 0.049.



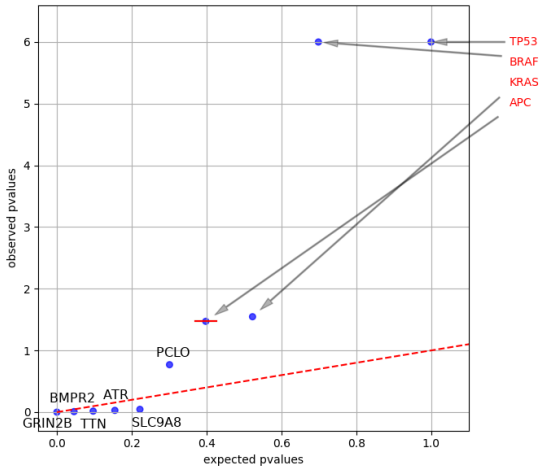
(a) DMM p_{any}



(b) DMM p_{true}

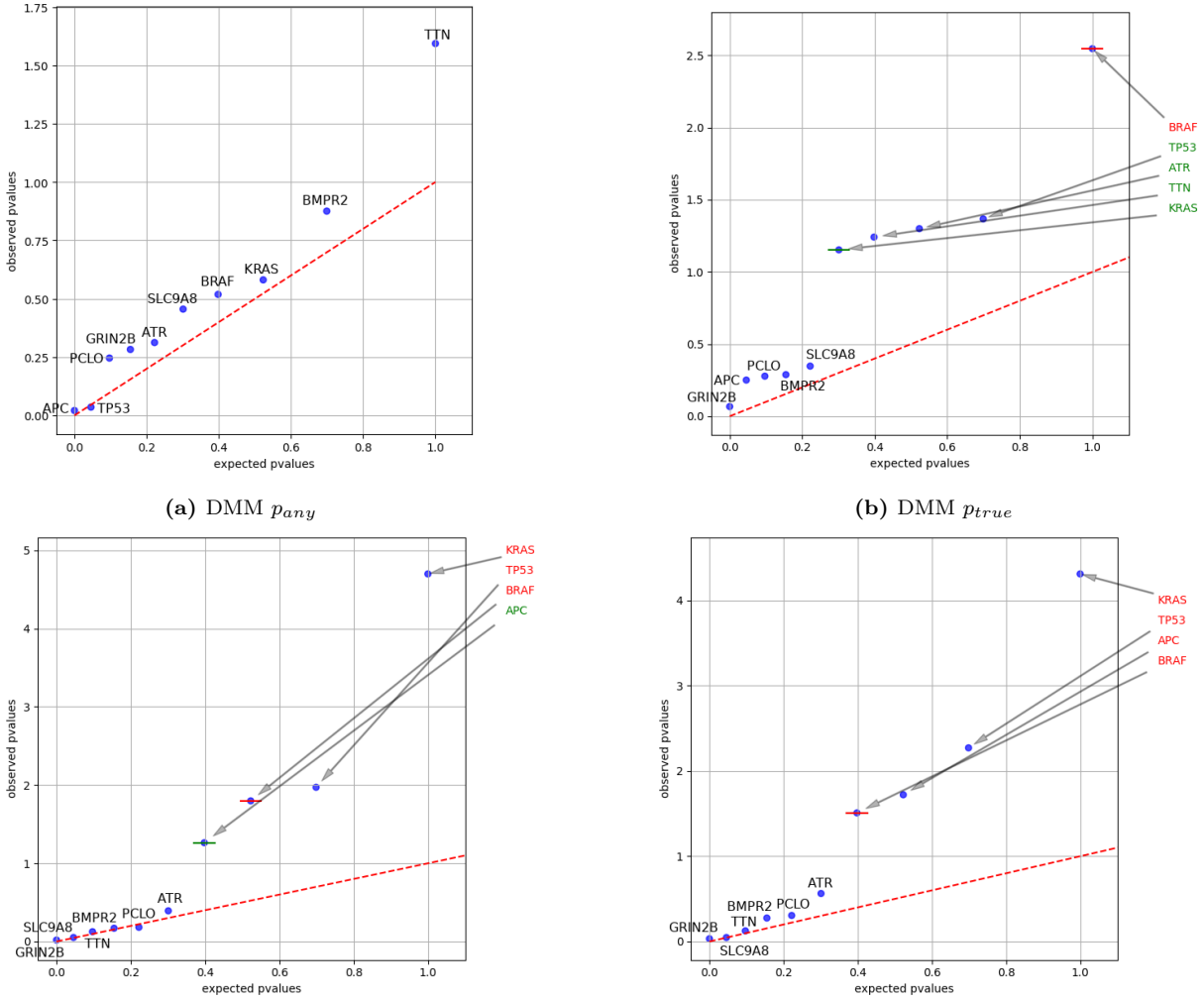


(c) CADD PHRED



(d) CADD RAW

Figure 4.18: The comparison between DMM scores and CADD scores in PCAWG Skin-Melanoma samples



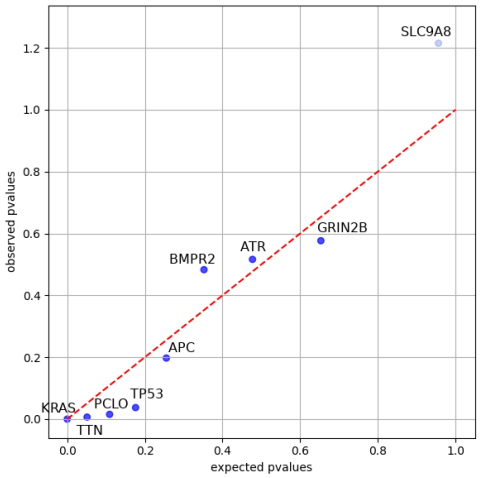
(a) DMM p_{any}

(b) DMM p_{true}

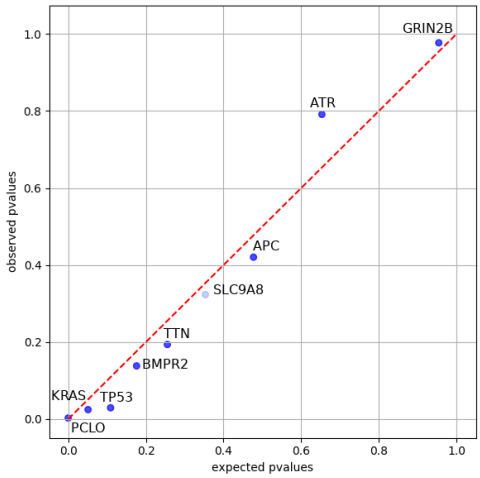
(c) CADD PHRED

(d) CADD RAW

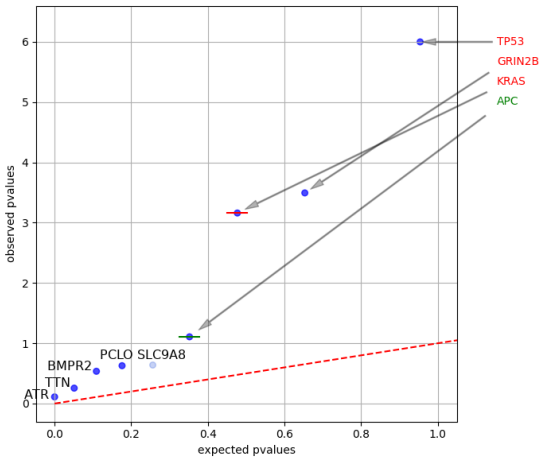
Figure 4.19: The comparison between DMM scores and CADD scores in PCAWG MSI samples



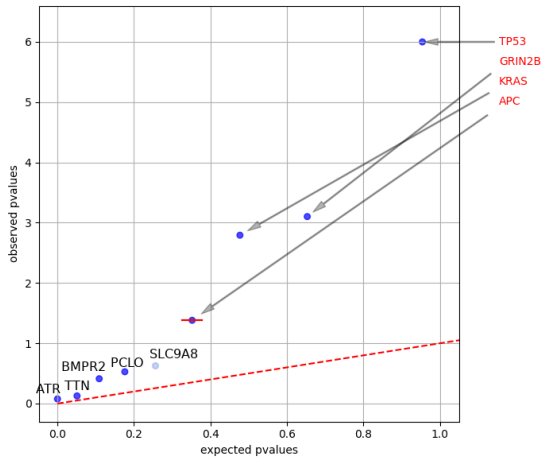
(a) DMM p_{any}



(b) DMM p_{true}

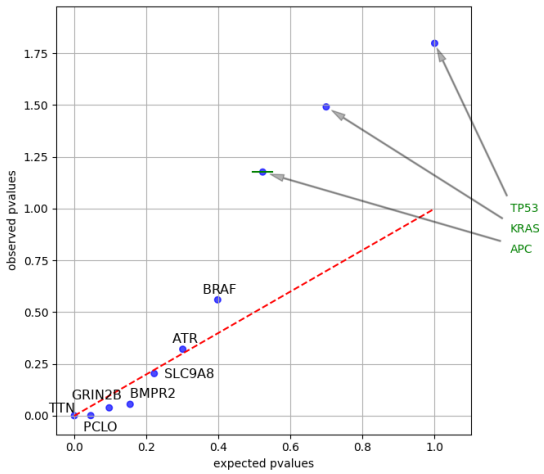


(c) CADD PHRED

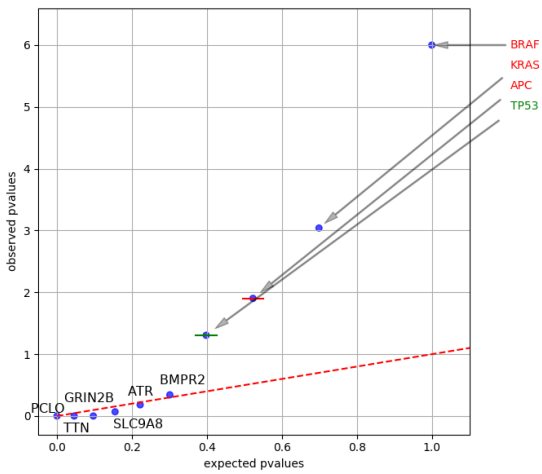


(d) CADD RAW

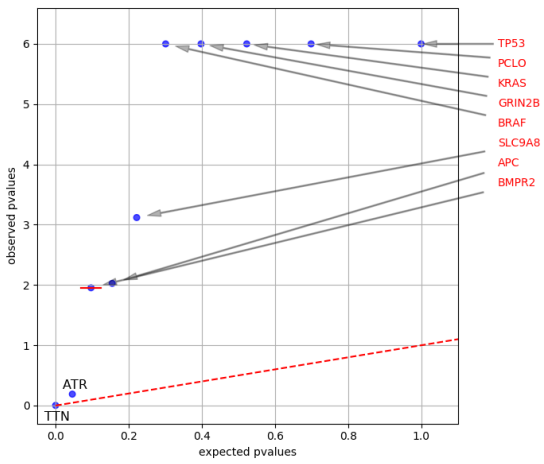
Figure 4.20: The comparison between DMM scores and CADD scores in TCGA hepatocellular carcinomas



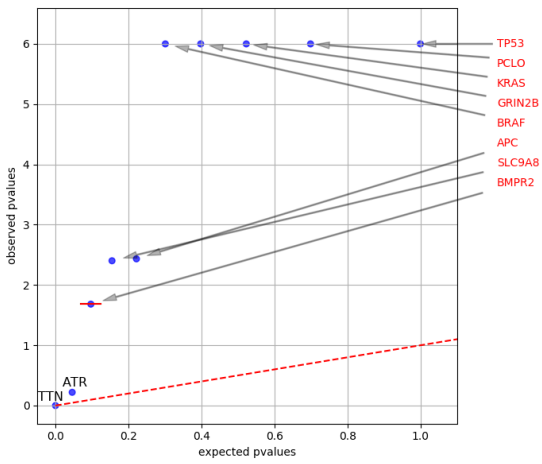
(a) DMM p_{any}



(b) DMM p_{true}



(c) CADD PHRED



(d) CADD RAW

Figure 4.21: The comparison between DMM scores and CADD scores in TCGA melanoma

5 Discussion

5.1 Commentary on individual results

The results shown in Tables 4.1 and 4.2 show that the more common a substitution is the higher the percentage of correctly predicted mutations and the higher the predicted probabilities are. This is likely due to the fact that the number of mutations in the test sample reflect the proportions of the mutations in the training set and because of this the model gets exposed to a higher number of common mutations and it becomes better at predicting these over the rarer mutations. Similarly Figure 4.1 shows that sequences with more than one mutation have higher predicted probabilities might be due to the fact that the model sees these regions with multiple mutations close by multiple times and the model thus might start learning these regions more compared to other regions. Overall, the result of more common mutations having higher predicted probabilities is not surprising in this light. The substitutions that C→T : G→A and that C→A : G→T especially show this effect where the large number of these mutations have made them easy to predict. The regions in Figure 4.2 which contains these substitutions overlaps with the regions for high p_{true} in Figure 4.5.

In the UMAPs in Figures 4.2, 4.3 and 4.5 it is visible that C→T : G→A substitutions tend to be located in the same regions as melanoma samples and high values for p_{true}. This is not surprising as shown in Tables 4.1 and 4.2 these substitutions are predicted correctly often, and they have the highest mean p_{true} values. These substitutions are also very common in melanomas and this can be observed from the fact that the regions for these substitutions overlap, for the most part, with the regions for melanoma samples. It is not surprising then that in Figures 4.7a and 4.7b that melanoma samples have the highest mean p_{any} and p_{true} out of the cancer types for PCAWG. Both in the PCAWG and TCGA datasets melanoma samples have the most mutations as seen in Figure 3.4 and based on Figure 4.6 that the number of mutations in melanomas has made predicting probabilities for mutations in melanoma easier compared to other cancer types. Especially based on Figure 4.6 the PCAWG models predictive ability is dependent on the number of mutations while the same effect is not seen in the TCGA model. While the different cancer types in TCGA have different number of mutations this seems to have little effect

on the predicted probability of p_{any} or p_{true} and instead the probabilities are close to each other.

It is difficult to say if the flatness of the TCGA model is because it has gotten enough examples from every substitution type to predict them all well or if there is something wrong with the model. Nevertheless, in almost all cases the TCGA model has smaller differences between the predicted probabilities for the mutations. According to the OLS regression the sample features age, sex, ancestry, and cancer type do not explain that much of the variance either when it comes to the TCGA model. It is possible that the TCGA model leans more towards the sequence features. On the other hand, according to OLS these sample features explain approximately half of the variance in PCAWG data.

In Figure 4.4 there are four melanoma clusters that the model has formed. In this case the clusters are mostly split into them based on the substitution type, but they are also split into these clusters based on their features. Clusters 5 and 12 resembled each other based on their features and 11 and 16 resembled each other. For example, the age distributions were similar to each other in the pairs and dissimilar to the other pair. Other features such as the top principal components of the sample information also showed this same split into pairs. The exact reasons for this split were hard to parse out and I could not form coherent patient profiles or groups out of the clusters.

Similarly, to the cancer types the age and sex affect the number of mutations in the samples and thus affect the mean p_{any} and p_{true} in these same groups. The sex difference was visible in OLS too as the sample being from a male increased the probabilities slightly, but this is probably because there were more male samples to begin with. In Figure 3.3 it is visible that older age groups have higher number of mutations and at least in the case of PCAWG in Figure 4.8a the mean probabilities are also affected by age. In the 25-45 and 45-60 age groups females have more mutations and higher predicted probabilities than males of the same ages. Here we can see the effect of the number of mutations on the probability outputs.

The CMC boxplots in Figure 4.9 are interesting as the hypothesis regarding driver mutation prediction with DMM is that the driver mutations should be harder to predict. In these pictures it is clearly visible that the more pathogenic tiers have lower probabilities than the 'Other' tier that contains mutations that are not known to be pathogenic. Especially the PCAWG results looked promising as the difference between the medians between the pathogenic tiers and the 'Other' tier were large and especially for tier 1, which has the most evidence for being pathogenic. The differences between medians are not as great in

the TCGA model and the predicted probabilities are again quite flat. Interestingly in the TCGA model tier 2 mutations had lower median probability than tier 1 mutations. Similarly in Figure 4.10 the SynMICdb score correlates better with the PCAWG model than with the TCGA model probabilities. With PCAWG the correlation is quite strong too and the interpretation from the scatterplots would be that when the predicted probability increases then the potential functional impact of the synonymous mutation lessens. The same idea behind the driver mutations applies here, mutations that have high functional impact are rare and thus it is hard to predict them. In Figure 4.11 most of the SynMICdb mutations that have a higher score seem to cluster around the wedge shaped regions that have in Figure 4.5 lower values for p_{any} and p_{true} as can also be seen from the scatterplots. These regions in the Figure 4.11 are also not as clearly dominated by a certain substitution or by any one cancer type as seen in Figure 4.2.

Interestingly OncoVar and CADD diverge from this hypothesis that driver mutations are harder to predict. CADD had no significant correlation for either p_{any} or p_{true} in either the PCAWG or the TCGA dataset. OncoVar did not exhibit correlation with the datasets either. Interestingly in Figures 4.13 and 4.14 the differences between the significance tiers are quite strong and especially in the case of OncoVar score.

The hotspot figures 4.15 and 4.16 seem to support the hypothesis again that driver mutations are harder to predict if we consider the hotspot mutations to be likely driver mutations. CADD also behaves as expected in hotspots in Figure 4.17 as the hotspots have a higher median score. In this case I considered only the hotspot positions instead of the exact mutations as there were many mutations that were in the hotspot positions but did not have the exact same alleles. The mutations in known hotspot had overall lower medians for the predicted probabilities than the mutations that are not located in the hotspot loci. In PCAWG the median p_{any} was also lower in the Figure 4.16a for the four cancer genes. However, in the case of p_{true} for the hotspots in Figure 4.16b *TP53* has a higher median for hotpots than for unknown mutations, which could be because In the case of *TP53* the specific hotpot mutations might be so common that the DMM model simply learns these mutations, while for *KRAS* it is hard to say which group has a higher median for p_{true}. The other two genes have lower medians for the hotspot mutations.

In Figure 4.16c only *APC* hotspot mutations have a clearly lower median p_{any} than the unknown mutations do. On the other hand in the case of p_{true}, *TP53* was the only gene that had higher median p_{true} for the hotspot mutations than the unknown mutations as seen in Figure 4.16d. For the other three genes the median was lower for the hotspot

mutations. The flatness of the TCGA model is visible here too as the hotpot and unknown groups' median values became more similar.

Finally, I performed OncodriveFML analysis of the different genes for the four cancer types. Figure 4.18 shows the results for PCAWG melanomas. In the PCAWG Skin-Melanoma genes it was most striking how there is p-value inflation in the p_{any} based score analysis as all of the gene points start to rise and hover over the line of expected values as seen in Figure 4.18a. By Contrast the p_{true} based score works much better as most of the genes end up being only slightly above the line of expected values. *BRAF* also shows up in the p_{true} figure with a very significant p-value as it reached the smallest possible value as seen in Figure 4.18b. The CADD values seem to work better than the DMM based scores as they identify both *BRAF* and *TP53* as significantly mutated genes with very low p-values as seen in Figures 4.18d and 4.18c. In this case the DMM scores likely need some kind of adjustment by either training the model again or there might need to be more samples as the number of observed mutations for the PCAWG melanoma were quite low as shown in Table 3.1. The TCGA melanoma dataset had many more mutations within the regions and overall worked better in OncodriveFML so more samples might be what is needed.

On the other hand, the opposite is seen in TCGA skin cutaneous melanoma where the CADD scores seem to rate all genes as significant even though only four of them are supposed to be cancer genes as seen in Figures 4.21d and 4.21c. The p_{any} based score does not seem to work all that well according to subfigure 4.21a here either as most of the p-values the cancer genes get are not that significant and *BRAF* is not significant at all even though it would be expected for it to be as it is very commonly mutated in skin cancers. It does become significant when observing Figure 4.21b where p_{true} score was used and it achieves very good p-values. Figure 4.21b is also interesting as it is the only one where the OncodriveFML has managed to separate all of the four known cancer genes from the other genes even if it does not give them very significant p-values. Only *BRAF* and perhaps *KRAS* seem like they would have significance. Interestingly here the TCGA model seems to perform better than the PCAWG model even though in most cases the TCGA model seems to have had trouble differentiating between more pathogenic and benign mutations. This, however, could also be due to the TCGA dataset having more observed mutations available for melanomas than the PCAWG dataset does.

In Figure 4.19 we can see that for the MSI samples with PCAWG p_{any} score that the score does not work at all as all of the genes are around the line for expected values. The p_{true} score does not work all that great either while the CADD scores are able to separate out

the four cancer genes. Even CADD does not seem to work all that well in significance of the genes with the major exception being *KRAS*. In this case the analysis likely suffered from only having very few observable mutations within these genes.

Similarly, TCGA hepatocellular carcinoma samples only had very few mutations which likely caused issues in the analysis. In Figure 4.20 it is observable that the DMM based scores did not work but at least the genes are around the line for expected values rather than suffering from p-value inflation. Even the CADD scores seem to have had some issues with the analysis as the only cancer gene to get very significant p-values is *TP53*, while they identified other genes as likely significant most of them do not seem to have very high p-values.

Overall, it seems that DMM score can be used as a score in OncodriveFML but the exact way the score is constructed starting from the DMM model training is going to need some work. Larger mutational datasets are also likely to work better for the analysis as a lot of the genes had very few observable mutations.

5.2 Research objective fulfillment

Currently, DMM features are only based on the sample and sequence information. The basic features of DMM indicate that the models might be quite dependent on the number of example mutations used during training. The indicators for this are the substitutions that have the highest predicted probabilities are also the more common ones in the datasets and as these same datasets are used for training the distribution of training examples are likely very similar. Additional support for this is that the more mutations there are in a cancer type the higher the predicted probabilities are for it and this applies to other features too such as the sex of the patient from which the sample is from. The fact that the sequences which contain an extra mutation have higher probabilities also supports this conclusion as the DMM model is more likely to see these regions often during the training process and it might learn the features in these regions better. It is also possible that the sequence length used affects the outcomes as the model for PCAWG data used sequences of length 2048 base pairs while the TCGA model used 256 base pairs. The longer sequences in PCAWG model may contribute to the higher predicted values of the sequences with extra mutations as the longer sequences mean that the training regions of the different mutations may result in greater overlap between different the training examples.

The dimensionality reduction of the DMM outputs shows that the predicted mutations

and cancer types tend to cluster towards similar data points near them. This results in the same type of substitutions being close to each other. As the spectrum of substitutions varies between different cancer types the different cancer types also tend to cluster around each other and they also cluster based on the substitution type. For example, many of the cancer types form multiple clusters in Figures 4.3 and melanomas form the clearest separate clusters. These clusters consist almost completely of melanoma mutations. The main differences between these clusters seems to be the substitution type, but there are also some other unknown features that separate them. The effect is that these clusters tend to have different age distributions and other features. Since the UMAP reductions are based on the mutation probabilities it is expected that the probabilities p_{any}, p_{true} and p_{pred} tend to cluster around mutations that have similar probabilities and where C→T substitutions also tend to cluster.

Since current ways of scoring mutations' functional impact and pathogenicity are mostly the same for every cancer type and patient DMM could provide a method for making more customisable scores. As the DMM scores depend on the sample features this means that it can consider the cancer type, sex, age, and other features from the tumour as something that may influence the probability of the mutation. CMC significance tiers, CADD, and SynMICdb all have the same scores for every cancer type which means that they may not fit all cancer types or even individual tumours equally well. OncoVar has cancer type specific data available that was not analysed here. There are also other limitations on these scores such as the fact that CMC significance tiers assign all synonymous mutations as not significant and SynMICdb only works for synonymous mutations.

DMM interacted as expected with the CMC significance tiers and with SynMICdb. As DMM should be worse at predicting driver mutations it was expected that it should have lower values for the higher CMC significance tiers, though the effect was quite weak for the TCGA mutations. Similarly, the expected negative correlation existed between SynMICdb score and the DMM probabilities as the more pathogenic synonymous mutations should have lower DMM probabilities while they at the same time have high SynMICdb score. The mutations that had higher SynMICdb scores also clustered in the same regions as lower p_{any} and p_{true} values in the UMAP reductions as can be seen in Figures 4.5 and 4.11. DMM thus, should be able to find pathogenic synonymous mutations and not just non-synonymous mutations. CADD and OncoVar on the other hand did not work as expected. CADD should also exhibit negative correlation with the DMM probabilities, but it does not. The same can be observed with the OncoVar score and driver levels. When

using the CADD PHRED or OncoVar scores for the mutations in the CMC significance tiers they do behave as expected in that the tiers ranked as having some evidence for being drivers have higher medians than the 'Other' tier as can be observed from Figures 4.13 and 4.14. The issues with OncoVar may have been technical if the mutations did not combine together correctly. The hotspot mutation analysis worked as expected with DMM the mutations that are in hotspots have lower predicted probabilities than the ones that are not in known hotspots. CADD works as expected too in these hotspots too, as it has higher values in the hotspots than in the unknown ones. In general, DMM worked as expected except for CADD and OncoVar scores. It does not currently seem like DMM would be competitive as the other scores seem to predict the pathogenicity better, but DMM could be improved by adjusting the training and by adding epigenetic information to its modules. The training might need to be adjusted to consider different mutation clusters, recurrencies of driver mutations, and hypermutational processes present in large number of tumours. Epigenetics would likely improve the model as epigenetic features can explain large parts of the mutation density in different genomic regions.

DMM overall performed differently in the four different cancer datasets in the OncodriveFML analysis. Overall, p_{true} seems to work better than p_{any} in separating out the true cancer genes. CADD however consistently outperformed DMM in this. Thus, while DMM probabilities can be used in OncodriveFML in its current form it does not have the necessary power to separate the cancer genes from other genes. DMM's predictive power could be improved by either altering the DMM model itself or by altering the way the score file was created as now the mutations that had a zero probability were assigned the lowest non-zero probability found from the same sample and gene. If the probabilities were kept as close to zero as possible it might improve the analysis.

6 Conclusion

Cancer is one of the leading causes of death in the world and as its incidence is expected to increase it is critical to understand what causes it. Understanding cancer better will lead to better outcomes for the patients. As our understanding of cancer has grown it has led to understanding that cancer is not a single disease, but a group of disease caused by similar mechanisms. The cause of cancer is thought to be driver mutations that lead to the uncontrolled expansion of a somatic cell growth that ends up forming a tumour. The driver mutations can differ between cancers, so it is crucial to understand the driver mutation profiles of different cancer types. However, even then there is still the level of individual tumours and their mutation profiles. For the best results in cancer treatment understanding individual tumours would be ideal.

Driver mutations can be classified into non-synonymous, synonymous, mini, and non-coding driver mutations depending on what kind of mutation it is. Non-synonymous mutations are the most studied out of these groups. Most of the driver discovery methods and scores have been developed for the non-synonymous mutations. Most of these methods have focused on searching recurrent missense mutations. These same methods have often considered synonymous mutations as neutral events even though there is evidence that this is not always so. This focus on the coding regions is however somewhat limiting as most of the human genome is non-coding and most mutations are bound to happen there. Therefore, it is important to develop methods for driver discovery that can also function in the non-coding genome and can also work on both synonymous and non-synonymous mutations. Deep Mutation Modelling, DMM, is a deep neural network based methods for predicting the probability of mutations. As it is a deep learning based method it can work on any region of the genome as long as sufficient examples are available for the training.

Deep learning methods can learn directly from the data so human aid in discovering features is not needed. DMM thus should be able to learn what kind of mutation contexts surround driver mutations and what kind of contexts surround passenger mutations. As driver mutations are rarer than passengers its ability to predict drivers should be worse. This was demonstrated by the fact that DMM probabilities were worse for mutations that have been identified by Cancer Mutation Census to be pathogenic, or by a hotspot analysis to be in hotspots. The same effect was seen with synonymous mutations where more

pathogenic synonymous mutations have lower DMM probabilities. This effect, however, was not observed with all compared pathogenicity scores. DMM still shows promise in this area but it requires improvements to increase its predictive power. Besides just predicting possible mutations the DMM output can be used as a score in other methods. OncodriveFML is one such method where DMM can be used. However, just as with the driver mutation probabilities DMM does not have enough predictive power to work reliably with OncodriveFML yet.

Overall, DMM is a potentially useful method for scoring mutations anywhere in the genome for multiple different mutation types. As it works with individual tumour samples it can bring studying cancer mutations one step closer to personalised medicine.

7 Acknowledgements

The results shown here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga

Bibliography

[1] The ENCODE Project Consortium. Overall coordination (data analysis coordination). Dunham I. *et al.* "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489 (2012), pp. 57–74. DOI: https://doi.org/10.1038/nature11247.

[2] Steve Agajanian, Odeyemi Oluyemi, and Gennady M. Verkhivker. "Integration of Random Forest Classifiers and Deep Convolutional Neural Networks for Classification and Biomolecular Modeling of Cancer Driver Mutations". In: *Frontiers in Molecular Biosciences* 6 (2019). DOI: https://doi.org/10.3389/fmolb.2019.00044.

[3] Christof Angermueller et al. "Deep learning for computational biology". In: *Molecular Systems Biology* 12 (2016). DOI: https://doi.org/10.15252/msb.20156651.

[4] Laura Bennett et al. "Mutation pattern analysis reveals polygenic mini-drivers associated with relapse after surgery in lung adenocarcinoma". In: *Scientific Reports* 8 (2018). DOI: https://doi.org/10.1038/s41598-018-33276-3.

[5] Anna-Leigh Brown et al. "Finding driver mutations in cancer: Elucidating the role of background mutational processes". In: *PLOS Computational Biology* (2019). DOI: https://doi.org/10.1371/journal.pcbi.1006981.

[6] Francesc Castro-Giner, Peter Ratcliffe, and Ian Tomlinson. "The mini-driver model of polygenic cancer evolution". In: *Nature Reviews Cancer* 15 (2015), pp. 680–685. DOI: https://doi.org/10.1038/nrc3999.

[7] Matthew T. Chang et al. "Accelerating Discovery of Functional Mutant Alleles in Cancer". In: *Cancer Discovery* 8 (2018), pp. 174–183. DOI: https://doi.org/10.1158/2159-8290.CD-17-0321.

[8] Esa Pitkänen *et al. Modeling somatic mutagenesis with deep convolutional networks, unpublished manuscript.*

[9] Jianjiong Gao et al. "3D clusters of somatic mutaions in cancer reveal numerous rare mutations as functional targets". In: *Genome Medicine* 9 (2017). DOI: https://doi.org/10.1186/s13073-016-0393-x.

[10] Levi A. Garraway and Eric S. Lander. "Lessons from the Cancer Genome". In: *Cell* 153 (2013), pp. 17–37. DOI: https://doi.org/10.1016/j.cell.2013.03.002.

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

[12] Alan Hodgkinson and Adam Eyre-Walker. "Variation in the mutation rate across mammalian genomes". In: *Nature Reviews Genetics* 12 (2011), pp. 756–766. DOI: https://doi.org/10.1038/nrg3098.

[13] Eran Hodis et al. "A Landscape of Driver Mutations in Melanoma". In: *Cell* 150 (2012), pp. 251–263. DOI: https://doi.org/10.1016/j.cell.2012.06.024.

[14] Jaime Iranzo, Iñigo Martincorera, and Eugene V. Koonin. "Cancer-mutation network and the number and specificity of driver mutations". In: *PNAS* 115 (2018), pp. 6010–6019. DOI: https://doi.org/10.1073/pnas.1803155115.

[15] Riku Katainen et al. "CTCF/cohesin-binding sites are frequently mutated in cancer". In: *Nature Genetics* 47 (2015), pp. 818–821. DOI: https://doi.org/10.1038/ng.3335.

[16] Riku Katainen et al. "Discovery of potential causative mutations in human coding and noncoding genome with the interactive software BasePlayer". In: *Nature Protocols* 13 (2018), pp. 2580–2600. DOI: https://doi.org/10.1038/s41596-018-0052-3.

[17] Sunkyu Kim et al. "Mut2Vec: distributed representation of cancerous mutations". In: *BMC Medical Genomics* 11 (2018). DOI: https://doi.org/10.1186/s12920-018-0349-7.

[18] Johanna Kondelin et al. "Comprehensive evaluation of coding region point mutations in microsatellite-unstable colorectal cancer". In: *EMBO Molecular Medicine* 10 (2018). DOI: https://doi.org/10.15252/emmm.201708552.

[19] Runjun D Kumar, S Joshua Swamidass, and Ron Bose. "Unsupervised detection of cancer driver mutations with parsimony-guided learning". In: *Nature Genetics* 48 (2016), pp. 1288–1294. DOI: https://doi.org/10.1038/ng.3658.

[20] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep Learning". In: *Nature* 521 (2015), pp. 436–444. DOI: https://doi.org/10.1038/nature14539.

[21] Jia Li et al. "Mining the coding and non-coding genome for cancer drivers". In: *Cancer Letters* 369 (2015), pp. 307–315. DOI: http://dx.doi.org/10.1016/j.canlet.2015.09.015.

[22] Eric Minwei Lie et al. “Identification of Cancer Drivers at CTCF Insulators in 1,962 Whole Genomes”. In: *Cell Systems* 8 (2019), pp. 446–455. DOI: https://doi.org/10.1016/j.cels.2019.04.001.

[23] Lucas Lochovsky et al. “LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations”. In: *Nucleic Acids Research* 43 (2015), pp. 8123–8134. DOI: https://doi.org/10.1093/nar/gkv803.

[24] Ping Luo et al. “deepDriver: Predicting Cancer Driver Genes Based on Somatic Mutations Using Deep Convolutional Neural Networks”. In: *Frontiers in Genetics* 10 (2019). DOI: https://doi.org/10.3389/fgene.2019.00013.

[25] Laura E. MacConaill and Levi A. Garraway. “Clinical Implications of the Cancer Genome”. In: *Journal of Clinical Oncology* 28 (2010), pp. 5219–5228. DOI: https://doi.org/10.1200/JCO.2009.27.4944.

[26] Iñigo Martincorena and Peter J. Campbell. “Somatic mutation in cancer and normal cell”. In: *Science* 349 (2015), pp. 1483–1489. DOI: https://doi.org/10.1126/science.aab4082.

[27] Iñigo Martincorena et al. “Universal Patterns of Selection in Cancer and Somatic Tissues”. In: *Cell* 171 (2017), pp. 1029–1041. DOI: https://doi.org/10.1016/j.cell.2017.09.042.

[28] Iñigo Martincorera et al. “High burden and pervasive positive selection of somatic mutations in normal human skin”. In: *Science* 348 (2015), pp. 880–886. DOI: https://doi.org/10.1126/science.aaa6806.

[29] Francisco Martínez-Jiménez et al. “A compendium of mutational cancer driver genes”. In: *Nature Reviews Cancer* 20 (2020), pp. 555–572. DOI: https://doi.org/10.1038/s41568-020-0290-x.

[30] Leland McInnes, John Healy, and James Melville. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” in: *ArXiv e-prints* (2018). URL: https://arxiv.org/abs/1802.03426.

[31] William McLaren et al. “The Ensembl Variant Effect Predictor”. In: *Genome Biology* 17 (2016). DOI: https://doi.org/10.1186/s13059-016-0974-4.

[32] Loris Mularoni et al. “OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations”. In: *Genome Biology* 17 (2016). DOI: https://doi.org/10.1186/s13059-016-0994-0.

[33] The Cancer Genome Atlas Research Network, Genome Characterization Center, and Chang K. *et al.* "The Cancer Genome Atlas Pan-Cancer analysis project". In: *Nature Genetics* 45 (2013), pp. 1113–1120. DOI: https://doi.org/10.1038/ng.2764.

[34] Paul A. Northcott et al. "Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma". In: *Nature* 511 (2014), pp. 428–434. DOI: https://doi.org/10.1038/nature13379.

[35] Ruth Nussinov and Chung-Jung Tsai. "'Latent drivers' expand the cancer mutational landscape". In: *Current Opinion in Structural Biology* 32 (2015), pp. 25–32. DOI: https://doi.org/10.1016/j.sbi.2015.01.004.

[36] S. W. Piraino and S. J. Furney. "Beyond the exome: the role of non-coding somatic mutations in cancer". In: *Annals of Oncology* 27 (2016), pp. 240–248. DOI: https://doi.org/10.1093/annonc/mdv561.

[37] *Pitkäniemi J, Malila N, Virtanen A, Degerlund H, Heikkinen S, Seppä K. Syöpä 2018. Tilastoraportti Suomen syöpätilanteesta. Suomen Syöpäyhdistyksen julkaisuja nro 93. Suomen Syöpäyhdistys, Helsinki 2020.*

[38] Paz Polak et al. "Cell-of-origin chromatin organization shapes the mutational landscape of cancer". In: *Nature* 518 (2015), pp. 360–364. DOI: https://doi.org/10.1038/nature14221.

[39] Eduard Porta-Pardo et al. "Comparison of algorithms for the detection of cancer drivers at subgene resolution". In: *Nature Methods* 14 (2017), pp. 782–788. DOI: https://doi.org/10.1038/nmeth.4364.

[40] Sunniyat Rahman and Marc R. Mansour. "The role of noncoding mutations in blood cancers". In: *Disease Models & Mechanisms* 12 (2019). DOI: https://doi.org/10.1242/dmm.041988.

[41] Philipp Rentzsch et al. "CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores". In: *Genome Medicine* 13 (2021). DOI: https://doi.org/10.1186/s13073-021-00835-9.

[42] Philipp Rentzsch et al. "CADD: predicting the deleteriousness of variants throughout the human genome". In: *Nucleic Acids Research* 47 (2019), pp. 886–894. DOI: https://doi.org/10.1093/nar/gky1016.

[43] Esther Rheinbay et al. "Analyses of non-coding somatic drivers in 2,658 cancer whole genomes". In: *Nature* 578 (2020), pp. 102–111. DOI: https://doi.org/10.1038/s41586-020-1965-x.

[44] Noa Rivlin et al. “Mutations in the p53 Tumor Suppressor Gene: Important Milestones at the Various Steps of Tumorigenesis”. In: *Genes & Cancer* 2 (2011), pp. 466–474. DOI: https://doi.org/10.1177/1947601911408889.

[45] Benjamin Schuster-Böckler and Ben Lehner. “Chromatin organization is a major influence on regional mutation rates in human cancer cells”. In: *Nature* 488 (2012), pp. 504–507. DOI: https://doi.org/10.1038/nature11273.

[46] Yogita Sharma et al. “A pan-cancer analysis of synonymous mutations”. In: *Nature Communications* 10 (2019). DOI: https://doi.org/10.1038/s41467-019-10489-2.

[47] Zbyslaw Sondka et al. “The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers”. In: *Nature Reviews Cancer* 18 (2018), pp. 696–705. DOI: https://doi.org/10.1038/s41568-018-0060-1.

[48] Tom Strachan and Andrew Read. *Human Molecular Genetics*. New York : Garland Science, 2018.

[49] Hyuna Sung et al. “Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries”. In: *CA: A Cancer Journal for Clinicians* 71 (2021), pp. 209–249. DOI: https://doi.org/10.3322/caac.21660.

[50] Fran Supek et al. “Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers”. In: *Cell* 156 (2014), pp. 1324–1335. DOI: https://dx.doi.org/10.1016/j.cell.2014.01.051.

[51] David Tamborero, Abel Gonzalez-Perez, and Nuria Lopez-Bigas. “OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes”. In: *Bioinformatics* 29 (2013), pp. 2238–2244. DOI: https://doi.org/10.1093/bioinformatics/btt395.

[52] Tao Wang et al. “OncoVar: an integrated database and analysis platform for oncogenic driver variants in cancers”. In: *Nucleic Acids Research* 49 (2021), pp. 1289–1301. DOI: https://doi.org/10.1093/nar/gkaa1033.

[53] The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, Campbell P.J, and Getz G. *et al.* “Pan-cancer analysis of whole genomes”. In: *Nature* 578 (2020), pp. 82–93. DOI: https://doi.org/10.1038/s41586-020-1969-6.

[54] Ziheng Yang and Joseph P. Bielawski. “Statistical methods for detecting molecular adaptation”. In: *Trends Ecology & Evolution* 15 (2000), pp. 496–503. DOI: https://doi.org/10.1016/S0169-5347(00)01994-7.