Computational Analysis of Patterns of DNA Damage

Sarah Kim Wooller

Thesis submitted for the degree of Doctor of Philosophy

University of Sussex

December 2020

Declaration

I hereby declare that this thesis has not been and will not be, submitted in whole or in part to another University for the award of any other degree.

Signature :

Acknowledgements

Many, many, thanks to my supervisor, Dr Frances Pearl, for her unique combination of support, friendship and challenge. She put her faith in me from the moment we met, when I was very nervous indeed about what the future would hold after 20 years as a career civil servant. I have never felt that faith waiver, and I am grateful for it. Not only does she keep on insisting on biological relevance but she has done a lot to get and keep me organised, and to get me to finish things. Thanks also to her family for their friendship. I am also very grateful to Dr Graeme Benstead-Hume, the best PhD team-mate I could have wanted. Graeme has a boundless energy and interest in all things computational and all things genetic, and I am very pleased to have seen him go on to do a post-doc even though I have missed him a lot over the last year. I will also miss team-mates Xiangrong (Tina) Chen, Fahmida Banani and Adnan Cinar and wish you all the best with your future studies.

I would like to take this opportunity for a quick shout out to my children Catherine and Robert who never complained about my going back to university at the same time that they went, and to my lovely husband David for his help and support. I now understand that there is a certain art form to conversing about cancer and deadly bacteria and my family have managed beautifully, as well as learning when to tell me to just change the subject.

Finally, I would like to take this opportunity to thank the University of Sussex for funding my DPhil studies.

This thesis is dedicated to Henrietta Lacks (1920-1951), the unknowing source of the world's first immortalised cell line - HeLa - and to all the patients since who have been able to provide informed consent to allow their cancer data to be used to help others.

Abstract

Cancer and bacterial infection are major killers across the world. Personalised cancer therapies are needed that respond to individual mixes of DNA damage, and new antibiotics are needed to respond to resistance brought about by genetic mutations in bacteria [1][2].

In this thesis, I analyse the gene sequence patterns next to small mutations in cancer cells, identifying those associated with substitutions and indels. I show that in the exome there is an excess of in-frame indels compared to frameshift mutations; evidence of negative selection.

Next I analyse the associations between the sequences of driver genes and mutational frequency in cancers. I find most driver genes are more frequently mutated in those cancers where there is a good match between the mean mutational fingerprint for that cancer and the fingerprint formed from the mutations found in the driver gene in question.

I then extend existing work on mutational signatures, to identify novel bacterial mutational signatures. By comparing the signatures with those of human cancers and environmental mutagens, I identify alkylation as a driver of bacterial mutagenesis.

Next I review translational drug discovery, highlighting the use of bioinformatics to identify drug targets and biomarkers, assess protein druggability; and predict opportunities for drug repositioning.

Finally, I identify therapeutically actionable mutually exclusive gene pairs within human cancers. I show that the Poisson binomial distribution is better for identifying mutual exclusivity. The predictions are available on the new MexDrugs website, and my python implementation of the Poisson binomial test can be installed via pip.

Abbreviations

A	Adenine
BER	Base Excision Repair
COSMIC	Catalogue of Somatic Mutations in Cancer
C	Cytosine
CNV	Copy Number Variance
COAD	Colon adenocarcinoma
DNA	Deoxyribonucleic Acid
DDR	DNA Damage Response
EJC	Exon Junction Complex
FATHMM	Functional Analysis through Hidden Markov Models
G	Guanine
GRCh	Genome Reference Chromosome
HR	Homologous Recombination
KICH	Kidney Chromophobe
KIRP	Kidney renal papillary cell carcinoma
KIRC	Kidney renal clear cell carcinoma
MexDrugs	Mutually exclusive and Druggable gene pairs
MMF	Mean of the Mutational Fingerprints
MMEJ	Microhomology Mediated End Joining
MMseqs	Many-against-Many sequence searching
MMR	Mismatch Repair
MNU	Methylnitrosourea
NER	Nucleotide Excision Repair

NHEJ	Non-homologous End Join
NMF	Non-Negative Matrix Factorisation
PPI	Protein Protein Interactions
ROS	Reactive Oxygen Species
SNP	Single Nucleotide Polymorphism
T	Thymine
TCGA	The Cancer Genome Atlas
TSa	Tumour Suppressor-associated gene
USS	Unique Silent Substitution
UV	Ultra Violet light

Table of Contents

1 Introduction

DNA is constantly under attack. Damage occurs as a result of natural metabolic and hydrolytic processes as well as via double strand breaks during replication. If this damage is not identified and repaired, or is badly repaired, it can lead to non-replicating cells or mutations. In addition, mutations arise from the presence of mutagenic chemical or physical agents such as smoke, UV light, and asbestos, as well as the incorporation of foreign DNA (most normally viral in human cells) into DNA.

The distribution of nucleotides within the exome is far from random. In fact, it shows sufficiently distinctive structure that changes in the nucleotide distribution can be used to predict gene boundaries with considerable accuracy[3]. Not all nucleotide motifs are equally susceptible to damage so, as a result, the patterns of mutations seen provide a historical record of the damage to the DNA[4].

I have chosen to look at cancer cells and bacterial cells because both are under extreme evolutionary pressure, and both are major killers. Increasingly, methods of tackling them have relied on a more detailed understanding of their evolution.

1.1 Overview of DNA

Each chromosome comprises of a single deoxyribonucleic acid (DNA) molecule wound around histone octamers and supported by scaffolding proteins. The DNA itself consists of two paired chains of polynucleotides. Each of the nucleotides has a sugar-phosphate backbone, and a base, either: adenine (A), cytosine (C), guanine (G) or thymine (T). The

polynucleotide chains are anti-parallel and aligned in the familiar double helix so that bases adenine and thymine, and conversely cytosine and guanine, are paired[5]. The sequence of nucleotides on one of the strands is complementary to that on the other strand.

Only a small percentage of DNA is accounted for by protein or RNA coding genes. The function of much of the remaining DNA is poorly understood, though the term 'junk DNA' is now known to be a misnomer. In particular, each DNA strand includes several specialised regions of DNA. These include:

- a centromere allowing for the attachment and subsequent separation of paired chromosomes during replication[6];
- a large number of DNA replication origins allowing for the attachment of DNA polymerase[7];
- promoters, enhancers and insulators - regions of DNA that govern the expression of specific genes and are normally found in CG islands, i.e. regions of the DNA that have a locally higher proportion of C-G pairs[8].
- telomeres - repeats of the unit GGGTTA at the end of each linear chromosome. These are replenished each time a cell divides thus preventing reduction of the length of the chromosome each time the RNA primer reaches the end of the chromosome during replication. Telomeres are not found in circular chromosomes.

In any given cell, the specific set of genes which can be expressed is tightly controlled via changes to the chromatin packaging of chromosomes and covalent modification of DNA. I am particularly concerned here with patterns of methylation. These are methyl groups which become attached to the cytosine in a CpG dinucleotide, primarily in the CG islands.

Where the CG islands occur in the promoter or enhancer of a gene, large numbers of methylated cytosines can indirectly prevent binding of RNA polymerases, thus preventing transcription and gene expression.

1.2 DNA damage

Although DNA is a highly stable molecule, it is subject to attack from chemicals, both from essential chemical reactions such as metabolism, and via environmental mutagens. In addition, some cells are vulnerable to physical attack and to viral invasion. In addition, the very act of DNA winding and unwinding requires purposive damage.

1.2.1 Chemical damage

Chemical reactions as a result of ordinary cellular processes are the primary causes of DNA damage. The most potent of these are oxidation, alkylation, and hydrolysis. These reactions are exacerbated by environmental mutagens, and physiological conditions.

The most frequent damage is caused by reactive oxygen species (ROS) such as hydrogen peroxide (H2O2). ROS form as a result of metabolism and other biochemical processes and can cause several types of DNA damage including oxidised bases as well as both single and double strand breaks. ROS can be exacerbated by chronic infections that lead to an inflammatory response as well as lifestyle and dietary factors [9].

DNA is also constantly under attack from alkylating agents that deposit methyl or other small alkyl groups on DNA. Endogenous alkylating agents are very common. However, many environmental carcinogens also act as alkylating agents. For example, tobacco smoke

contains N-nitrosodimethylamine (NDMA), 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK), acrylonitrile and ethylene oxide, all of which are alkylating agents. [9],[10]. Adducts can also arise from high concentrations of reactive carbon species (RCS) which can result from conditions such as diabetes. These can lead to the creation of abasic sites and it has been shown that some bulky adducts can lead to fork stalling [11].

Finally, hydrolysis can act on DNA either to remove purines and sometimes pyrimidines leaving an abasic site. It also acts via hydrolytic deamination particularly of 5-Methylcytosine giving rise to uracil.

The most important environmental mutagens are by-products of smoking tobacco [12]. However, other carcinogenic chemicals include: some farming products – particularly chlorinated, organophosphate, and carbamate insecticides and phenoxy acid and triazine herbicides [13]; industrial chemicals such as aromatic amines, vinyl chloride, benzene, and chromium compounds; atmospheric pollutants resulting from incomplete combustion of fossil fuels; contaminants in drinking water produced during water chlorination; and some medications, particularly anticancer drugs, oestrogens, and analgesics [12].

1.2.2 Physical damage

In addition to chemical assault, some cells are subjected to physical assault from radiation such as sunlight, or from the inhalation of particulates such as dust or asbestos. UV light, and particularly UVB light, has the ability to directly attack DNA at two adjacent pyrimidines, particularly TT, forming dimers [14]. Therapeutic or nuclear radiation is of higher energy and is able to directly induce double strand breaks [15]. Physical damage from dust etc. is less

direct. The key importance here is size. Particulate matter which is small enough to enter the cell can give rise to inflammation which in turn leads to the generation of elevated levels of ROS which damage the DNA [16].

1.2.3 Viral attack

DNA from some specific viruses persist in the cell by becoming incorporated within the host genome. [17]. These include the Epstein-Barr virus [18] and the Karposi sarcoma herpes virus[19]. Genetic material from RNA based viruses such as the Rous Sarcoma Virus can also become incorporated [20].

1.2.4 Purposive damage

All activities that require DNA to be unwound quickly lead to increased tension in the DNA known as supercoiling. This includes acts of DNA replication, transcription, recombination, and chromatin remodelling. If left unresolved supercoiling would quickly lead to termination of the activity. In order to reduce the stress, purposive single strand DNA breaks are induced in order to allow the DNA to rotate at the point of the nick. Experiments have also been carried out to introduce topologically knotted structures in DNA. The successful resolution of these implies that that purposive double-stranded breaks can also be initiated and resolved [21][22].

1.2.5 Relative importance of DNA damage

The human mutation rate is approximately 1 mutation/1010 nucleotides/cell division. This equates to the loss of roughly 18,000 purine base a day as well as deamination of around 100 cytosine bases (p268) [23]. Three-quarters of lesions in human cells are single-strand

DNA breaks, which can be repaired using the complementary strand of the DNA as a template. The remaining quarter are double strand DNA breaks. These are more dangerous. A cancer cell may acquire over 1000 mutations as a result of an increasingly unstable genome, and compromised DNA damage repair system (DDR).

For bacterial cells the levels of change are very variable. However, as an example, Escherichia coli is estimated to acquire 8.9×10^{-11} mutations per base-pair per generation[24]. Whilst this may seem comparatively slight bacterial generation times can be as fast as 10 minutes in some species, so where mutations have no impact on fitness the variation due to single nucleotide polymorphisms (SNP)s can quickly mount up[25].

The pattern of unique mutations left in the DNA record of a cell reflect the mutagens to which the cell has been exposed, the efficacy of DNA damage repair pathways, and the cell's ability to survive any permanent damage[26].

1.3 DNA repair pathways

In order to maintain the integrity of the genome, all cells are equipped with many different pathways that enable the cell to identify and repair damage. Some of these DNA damage repair pathways are faithful, whilst others are error-prone, increasing the mutation rate. Most DNA repair pathways are available to humans and bacteria alike, though the exact genes involved vary. However, bacteria have two DNA repair pathways that are not available to animal cells: Direct Reversal and SOS.

1.3.1 Single strand repair

Damage to a single strand of DNA can be repaired directly using direct reversal or the damage can be cut out and repaired using the other DNA strand as a template. Nucleotide excision repair, base excision repair and mismatch repair are all variations on that theme. Direct reversal (DR) is used to repair damage done by the methylation of guanine bases, as well as some types of alkylation of adenine and cytosine bases[27]. A mechanism is also available in bacterial cells to enable direct repair of the pyrimidine dimers created in DNA by UV light [28]. This last mechanism is not available to humans who use nucleotide excision repair instead (see below).

Nucleotide excision repair (NER) is used to repair any bulky lesion affecting one strand of the DNA. The DNA is unwound at the site of the damage, a short section of the damaged strand is removed and DNA polymerase then replaces the damaged section using the opposite DNA strand as a template. This pathway is available to both human cells and to bacterial cells and is highly conserved. NER is used to remove a wide range of small lesions including those caused by sunlight [29].

Base excision repair (BER) is much more specific than NER, but is available for repairing the most common forms of damage including both depurination and depyrimidination. The damaged base is cut from DNA strand and then replaced using the opposite DNA strand as a template.

Mismatch repair (MMR) is a highly conserved pathway used to repair replication errors. A single strand of DNA is excised either side of the mis-match and then a new section of DNA

is resynthesized and ligased, using the other strand of DNA as a template. Defects in MMR are particularly important in cancer giving rise to substitutions and also to singleton indels at the site of mononucleotide repeats [30][31].

1.3.2 Double strand repair

Double strand breaks are a potentially lethal form of damage, with one unrepaired break sufficient to cause cell death or large scale genomic disruption such as chromosome loss [32]. Repairs may be made using homologous recombination or the less faithful non-homologous end joining or microhomology mediated end joining. Homologous recombination (HR) is an accurate double strand break repair mechanism which relies on use of the sister chromatid as a template. Following a double strand break, and in the presence of a homologous section of DNA, a section of double strand is unwound and degraded on either side of the break. The single stranded DNA then invades the sister chromatid at a site of shared homology, temporarily displacing one of the DNA strands. Polymerase is used to continue formation of the single strand. The two ends of single strand then join at a site of good homology, forming a now continuous length of DNA before polymerase is once more used to reform a double strand and ligase used to seal the ends. HR is available to humans during the S-phase when the sister chromatid is readily available. HR is also a highly conserved pathway in bacteria, but in haploid bacterial cells relies on the presence of homologous alien DNA[33].

DNA non-homologous end joining (NHEJ) is an error-prone mechanism for mending double strand breaks. No template is used, so NHEJ is available to bacterial cells and to human cells when no sister chromatid is readily available. The two ends of the break may be directly

religated leaving no error, or the ends may be remodelled and then religated, leaving insertions or deletions at the breakpoint. The role of NHEJ in preserving genomic integrity is two-sided: defects in the pathway lead to genomic instability, but on the other hand because of the error-prone nature of the NHEJ repair mechanism, dysregulation of the pathway is also associated with genomic instability and carcinogenesis. For example, although DSBs are preferentially repaired by HR during the S phase, defects in the HR pathway can lead to NHEJ repairing one-ended DSBs with another distal DSB resulting in translocations [34].

Microhomology-mediated end joining (MMEJ) or alternative nonhomologous end-joining (Alt-NHEJ) is a deletion-inducing mechanism in humans for repairing double-strand breaks in DNA. The use of small matching sequences during the alignment of strands results in deletions flanking the original break. MMEJ is frequently associated with chromosome abnormalities such as deletions, translocations, inversions and other complex rearrangements [35].

1.3.3 Changes to the regulation of DDR pathways

In addition to the pathways identified above, bacteria have an SOS response which is not available to mammals. The SOS response is a DDR pathway in that it regulates the repair of DNA damage, but it does so in a way which leads to a highly elevated mutation rate. As a result, it generates genetic diversity and enables adaptation to stress, which can provide a survival advantage to the bacterial colony. The response is provoked by a wide range of exogenous triggers including some antibiotics, physical stress such as pressure or starvation, or incorporation of new genetic material, or endogenous triggers such as Reactive Oxygen

Species (ROS). These triggers lead to high levels of single strand DNA (ssDNA) that trigger the pathway. Critically the SOS pathway makes use of error-prone DNA polymerases which lead to the increased mutation rate [36][37].

Analogies can be drawn with cancer cells which frequently have dysregulated DDR paths. Mutations to the DDR paths may be under positive selection pressures to provide cells with the ability to tolerate the high levels of double strand breaks during replication that are induced by activated oncogenes[38][39]. When the DDR pathways are defective the specific patterns of mutations depend on the mutagens to which the cell has been exposed: MMR deficiency gives rise to a high number of substitution mutations together with singleton insertions and deletions at the site of homopolymer repeats, whereas deficiencies in double strand break repairs tend to exhibit a more uniform base substitution spectrum, together with small deletions and tandem duplications [40]. Copy number variations can also be brought about by erosion of telomeres at the end of the chromosomes. This may lead to senescence but can give rise to telomere crisis, during which time DNA damage repairs goes wrong. This crisis may lead to cell death but alternatively can lead to gene fusions, amplifications, and deletions during inaccurate DNA damage repair [41].

1.4 Overview of human cancer cells

In 2018 over 18 million new cases of cancer were diagnosed and this is predicted to rise to 29 million by 2040 as populations age[42]. As well as surgery and radiation, there are three pillars of medicinal cancer therapy: cytotoxic chemotherapy, targeted therapy, and immune therapy. The research field is strong in each of these areas. For example, smart drug delivery systems are being developed to enable reduced doses and better targeting of cytotoxic

drugs, so reducing side-effects and toxicity [43]. There is considerable research into inhibitors of known oncoproteins such as tyrosine kinases inhibitors[44][45] and oestrogen antagonists[46] and because MYC is both such an important oncogene and yet also undruggable there is considerable research into ways of reducing its expression by drugging BET bromodomains and CDKs that support MYC in tumour cells [47].

In addition, advances have been made in cancer immunotherapy, enabling the body to fight the cancer particularly in metastatic melanoma where monoclonal antibodies improve T-cell immune function [48]. Yet we are not close to saying that cancer is cured.

Cancer is caused by DNA damage to the cells of the host that is inadequately repaired. The damage is not targeted, and generally either tolerated or leads to cell senescence or apoptosis. Although childhood cancers exist, most tumours develop over the course of many years or decades as a result of successive genetic and epigenetic changes. DNA mutations are unfortunately a normal part of cell aging and, by middle-age, even cells that do not show signs of cancer may nevertheless have a high burden of mutations [49]. Cancer arises much more frequently in some tissue types than in others and the rate has been shown to be strongly correlated to the normal replication rate for the cell [50].

1.4.1 Human DNA

In each of the typical somatic cells in the human body there are 23 pairs of linear chromosomes. 22 non-sex or autosomes present as a matched but not identical pair, and the appropriate X and Y chromosomes, generally XX for women or XY for men, though other possibilities exist. This is dubbed the euploid karyotypic state [51]. In total the human genome has roughly 3.2 billion nucleotide pairs (6.4 billion for a diploid cell)[52].

Within this vast amount of information roughly 1% is accounted for by 30,000 regions of DNA known as genes which provide the template for the generation of protein (21,000), or in 9,000 cases functionally significant RNA molecules.

The chapters of this thesis concerning cancer explore primarily mutations to protein-coding genes, changes to copy number of the chromosomes and the methylation of promoter regions, as this is strongly associated with somatically heritable repression of gene expression[23].

1.4.2 Hallmarks of cancer

Where DNA damage results in changes to gene and protein expression and/or protein function, the DDR pathways can become disrupted and a form of Darwinian evolution can become apparent. Populations of clonal cells emerge which are derived from a single cell slightly more able to grow, replicate and evade death. Further DNA damage within such a population gives rise to new sub-clones which compete for limited nutrients. These sub-clones may die out, exist side by side, or expand further to become the dominant clone.

As successive generations of sub-clones emerge, the more successful sub-clones bear traits that enable them to become tumorigenic, and eventually to spread to other organs. These traits have become known as the hallmarks of cancer, and they provide a way for understanding the biology of cancer. They are: sustaining proliferative signaling; evading growth suppressors; activating invasion and metastasis; enabling replicative immortality; inducing angiogenesis; and resisting cell death. These hallmarks were described originally in 2000 by Hanahan and Weinberg [53] , and were updated in 2011 to include two further

emerging hallmarks: deregulating cellular energetics, and avoiding immune destruction; and two characteristics that enable the development of the other hallmarks: genomic instability and inflammation [54]. These are shown in figure 1.1 and described in more detail below.

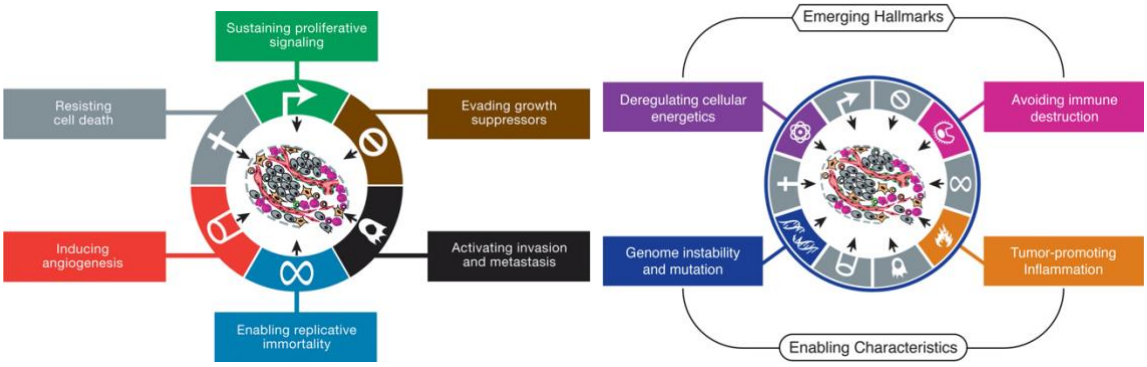


Figure 1.1 – The original and emerging hallmarks of cancer, reproduced from "Hallmarks of Cancer: The Next Generation" [54]

1.4.2.1 Sustaining proliferative signaling

Ordinarily, tissues strictly regulate cell growth and division, using growth factors to enable cross talk between cells. By so doing they maintain homeostasis. By contrast, cancer cells regrow and proliferate in a way which is largely irresponsive to growth factors. They do this through mutations that deregulate the impact of growth factors, for example by enabling constitutive signaling of the MAP and P13K signaling pathways [55] and by suppressing negative feedback of these pathways for example by mutations to PTEN [56].

1.4.2.2 Evading growth suppressors

Before replicating, normal cells require both the positive growth signals mentioned above, and the absence of negative signals from tumour suppressors that regulate the cell cycle. When tumorigenic cells begin to grow and divide more rapidly the cells become stressed,

for example through less access to oxygen and nutrients, and as a result of genomic damage. Normally these conditions would be monitored by tumour suppressing pathways that bring the cell cycle to a halt. However, mutations to key tumour suppressors allow the cell to evade growth suppression. Two key examples of such tumour suppressors are: RB which relays and integrates inhibition signals primarily from outside the cell[57] and TP53 which halts the cell cycle if there are too many signs of stress or abnormality from within the cell [58]. There is however considerable redundancy in the gene network enabling mice without a functioning copy of RB or TP53 to maintain cell homeostasis[59],[60].

1.4.2.3 Activating invasion and metastasis

A malignant tumour must be able to move from the primary tissue to alternative tissues and, once there, adapt to cope with the different environment growing into a new tumour. Three mechanisms are known that enable metastasis. The first is called the epithelial-mesenchymal transition (EMT). During EMT, individual cells activate processes used during formation of embryonic organs and also during wound healing to allow the cell to invade the local tissue before entering and moving through blood and lymphatic vessels, exiting into new tissues and finally growing into new tumours[61]. Two other forms of metastasis are documented. The first is collective invasion, whereby whole cohesive clusters of cancer cells advance into adjacent tissues bypassing the need for the major changes during EMT [62]. Finally, in amoeboid invasion, individual cancer cells change shape, elongating to move through gaps in the extracellular matrix rather than first clearing a path[63]. Each step of migration and metastasis puts further evolutionary pressures on the cells. As a result, the genetic changes to metastatic cells may be very different to those of the originating tumorigenic cells.

1.4.2.4 Enabling replicative immortality

Normal cell lineages grow and divide a limited number of times before either entering senescence (which prevents further division) or undergoing a crisis leading to cell death. This crisis is brought about by the erosion of the telomeres during replication. When the telomeres reach a critically short length, they can no longer protect the chromosome from lethal fusions which generally lead to cell death. In most cancer cells, expression of telomerase bypasses cell death by extending the telomeres. However, some cancer cells with deficiencies in the apoptosis pathway may experience and survive telomere crisis before telomerase activation, resulting in cancer-promoting chromosomal fusions [64].

1.4.2.5 Resisting cell death

Non-tumorous cells that are suffering from the levels of stress seen in tumours would either go into cell cycle arrest and senescence, enter the process of programmed cell death (apoptosis), or break down cell organelles (autophagy). Tumour cells resist apoptosis, either by avoiding monitoring stress or during execution. The monitoring includes sensing the levels of stress (including by TP53), responding to insufficient survival factor signaling (via BH3) and responding to high levels of proliferation (via MYC). Tumour cells can resist apoptosis in a variety of different ways, for example through TP53 loss[65] [60], upregulation of antiapoptotic regulators and survival signals or downregulation of proapoptotic factors. Alternatively, they may short-circuit the apoptosis pathway[66]. Similarly, tumour cells may resist autophagy. However, autophagy may also be cancer enabling by allowing cancer cells to enter a form of dormancy during treatment which can later be reversed[67][68]. Necrosis, whereby cells die in a less controlled fashion, is similarly

a two-edge sword for cancer cells. The centre of a rapidly growing tumour becomes nutrient and oxygen deficient which can lead to necrotic cell death. During necrosis the cell contents explode into the surrounding tissue. These contents can then recruit inflammatory cells which induce angiogenesis, proliferation, and invasion[69].

1.4.2.6 Inducing angiogenesis

By the time a human has reached adulthood the blood system is largely complete with new blood vessels (angiogenesis) only being formed after wounding and during the menstrual cycle and pregnancy. However, new blood vessels are constantly needed for tumour growth to prevent death from hypoxia and nutrient deficiency, and the angiogenetic switch is typically switched on early in tumour progression, either directly by the upregulation of oncogenes such as MYC or indirectly via immune inflammatory cells[70].

1.4.2.7 Deregulating cellular energetics

Metabolism in cancer cells close to the centre of tumours is dominated by the need to cope with the hypoxic conditions found at the centre of tumours. In response, some cancer cells largely reprogram their glucose metabolism to glycolysis, even though this is much less efficient. However, glycolytic metabolism occurs even in oxygen rich tumours. It is believed that this may be because the alternative metabolic route also provides the cell with intermediary chemicals in order to biosynthesise new amino acids etc. Both KRAS activation and hypoxia lead to more glycolysis [71].

1.4.2.8 Tumour-promoting inflammation

Chronic inflammation enables tumour development through a number of different mechanisms. As previously mentioned, inflammatory cells release ROS, thereby accelerating mutagenesis of nearby cells, as well as aberrations in methylation. This impact has been associated with cancers arising as the result of viral or bacterial infections or other inflammatory conditions. Inflammatory immune cells also supply cancer with growth factors, survival factors, proangiogenic factors, and enzymes that modify the extracellular matrix so facilitating angiogenesis, invasion, and metastasis[72].

1.4.2.9 Genome instability and Mutation

Mutations to genes responsible for genome surveillance, maintenance and repair, together with loss of telomeric DNA, greatly accelerate the rate of DNA evolution. This evolution, whilst leading to the death of many tumour cells also enables the tumour to acquire mutations that lead to new clones with more of the hallmarks of a fully-fledged cancer[73].

1.4.2.10 Avoiding immune destruction

During the tumour's early stages, cytotoxic immune cells eliminate many immunogenic cancer cells. However, macrophages - innate immune cells- may either eliminate cancer cells or be pro-tumorigenic contributing to angiogenesis, remodelling of the extracellular matrix, and suppressing anti-tumour effector cells. Cancers can be classified by the expression profile of immune pathways in ways that largely cut across traditional cancer classifications. Such classifications reflect the different composition of immune cells present[74] [53][54].

1.4.3 Types of DNA damage in cancer

DNA damage results from three main sources: small scale mutations to the DNA, changes to the chromosomal copy number and heritable changes to the epigenome. Very few such changes have an important impact on carcinogenesis because most affect so-called 'junk' DNA or genes which play no direct role in cancer. However, a few hundred genes play a pivotal role in protein pathways affecting the cancer hallmarks identified above. These are dubbed 'driver genes' because disruption to their function can drive carcinogenesis. Often, one of the impacts of early mutations to driver genes is widespread epigenetic dysregulation. Heritable changes in the patterns of methylation as well as chromatin remodelling and histone modifications lead to stable, pathogenic, alteration in profiles of gene expression. A complex interplay between gene expression and both genetic and epigenetic damage then ensues, leading to further damage and genetic instability. Most cancers share this basic pattern, though some cancers develop much more quickly, either because a vital genetic mutation has been inherited and thus occurs in all the cells, or because of a single event of huge replicative stress leads to massive chromosomal rearrangement and simultaneous inactivation of many separate pathways [75].

In the same way that cancer is an umbrella term covering many different diseases of altered cell replication and metastasis, the damage at a genetic and epigenetic level is highly heterogenous.

At a statistical level, cancers at different sites can be classified by characteristic combinations of aneuploidy, somatic mutations and methylation patterns. For example, clear cell renal cell carcinomas frequently lose chromosome arm 3p and gain a copy of 5q, a pattern not seen in other organs [76]. However, two patients presenting with the same

pathology are likely to have very different profiles of DNA damage, and these differences translate into differences in prognosis and appropriate treatment. For example, breast cancer is partially classified by the status of receptors for oestrogen, progesterone, and human epidermal growth factor (ER, PR, and HER2). Women suffering from cancers with different biomarker status have varying responses to different therapies, and have different prognoses. In short, these biomarkers can be used to identify different diseases, not apparent from phenomenological symptoms [77].

The biomarkers and indeed driver genes may be either genes that are mutated or dysregulated in cancers found throughout the body or, as in the example above, genes which are important in only one or a small group of cancers [78][79].

1.4.3.1 Substitutions

Most cancers are characterised by many small mutations. The vast majority of these are the substitution of a single nucleotide (SNP), while most of the remainder are the insertion or deletion of a single nucleotide or a few nucleotides (indels).

Mutations outside the protein-coding region generally have no or little impact on phenotypic impact, and some cells are able to withstand many hundreds of mutations within the exome without becoming cancerous or dying. The main reasons for this are that: many mutations are silent, having no impact on protein production; many areas of the protein, particularly the disordered regions, can withstand considerable change without impacting on protein function, and within most cells there are considerable numbers of genes which are non-essential. This may be because they are not expressed in the tissue of

interest [80], or because a cell can continue to grow and reproduce without the protein in question – albeit with impaired function[81].

Silent mutations are nucleotide substitutions that result in no change of amino acid. There is considerable redundancy in the code used to translate RNA to amino acids. For example, all codons of the form 'ACN' code for threonine, where N is any of A,C,G,T. As a result, many mutations are silent, having no impact on the amino acid sequence. At the other end of the scale , where a substitution results in the formation of a premature stop codon TAA, TGA, TAG the resulting protein is foreshortened or, more usually, does not form at all as a result of RNA surveillance mechanisms. This is because exon junction complex (EJC) proteins are deposited on the RNA fragments at the site of exon-exon junctions. Where there is a premature stop codon ahead of an exon-exon junction, the presence of a subsequent EJC protein triggers nonsense mediated decay [82]. Substitutions that are neither silent nor nonsense mutations are classified as missense mutations. These substitutions give rise to the change of a single amino acid. Generally, the impact of these changes is very low but where the missense mutation takes place in a driver gene affecting one of the active domains of the resulting protein the impact can be profound. One of the most well studied driver genes, BRAF, is strongly selected for mutations at the codon 600 site. The substitution of V>E at that site causes the BRAF signalling pathway to become constitutively active, that is, it is not sensitive to feedback mechanisms. The modified protein is a strong driver in melanomas [83].

1.4.3.2 Indels

Around 4% of all small mutations are either insertions or deletions, collectively known as indels. Most of these are the insertion or deletion of a single nucleotide which has the impact of changing the codon frame from the point of the indel onwards. As a result, most indels lead to a profound change in the amino acids from that point onwards, as well as the introduction of a new stop codon which may lead to nonsense mediated decay. Throughout this thesis I therefore assume that frameshift indels give rise to loss of function of the protein. Inframe indels, consisting of multiples of three nucleotides lead to the addition or loss of a number of amino acids and may thus be less pathogenic.

1.4.3.3 Copy Number Variation

In a typical, non-tumorigenic cell, there are 22 matched pairs of non-sex or autosome chromosomes as well as an XX or XY pair. In some cancer types this ordered state is often lost, with frequent loss or gain of whole chromosome arms or large chunks of DNA. Analysis of the impact of changes in copy number is complicated by gene dosage compensation. That is copy number gains do not necessarily lead to increased expression. However, the loss of a part of both chromosomes leads to non-expression of any genes contained therein. It is also common to have copy neutral areas where DNA damage has nevertheless led to a loss of heterozygosity. In such areas mutations on tumour suppressors such as TP53, BRCA1, BRCA2 and PTEN are often copied over to the sister chromatid during double strand break repairs [84].

1.4.3.4 Gene Expression

At the time of writing the largest databases of cancer 'omics do not include protein expression. The closest, and inadequate, substitute for it is the RNA seq data provided either as fragments per kilobase million or as raw counts. Analysis of the RNA seq data can be used to compare expression between two groups. However, matched tumour/normal pairs of RNA seq data are not available, and moreover the RNA seq provides a single snap-shot during an ever-moving parade, quite unlike the historical record of DNA mutations and copy number and the stable changes to methylation. For many proteins there is a poor link between gene expression and protein expression. As a result, I have not included gene expression analysis within this thesis.

1.4.3.5 Methylation

Tumorigenesis progresses not only via permanent changes to the genome, but also through interactions between genomic mutations and abnormal gene expression. Whilst gene expression is intrinsically reactive, the silencing or activation of specific genes caused by epigenetic changes can be very stable as they are linked to mutations in epigenetic writers, readers, and erasers, thus offering opportunities for cancer therapies[85]. These epigenetic changes include: histone modifications; the reorganisation of chromatin causing heritable silencing or activation of particular genes; and methylation of CpG nucleotides. It can be extensive [86]. However, as with other modifications, the epigenetic changes are remarkably specific to particular cancers. For example there are large differences between the methylation patterns found in childhood MLL-r and other common childhood leukaemias [87].

1.4.4 Driver genes

At the time of writing around 1% of all genes have been implicated in the development of cancer via mutations, most usually in a specific tissue type [88]. These are referred to throughout this thesis as either driver genes or cancer-associated genes in contexts where the link between the gene and cancer in question has not been established. These driver genes are commonly dubbed tumour suppressor genes and proto-oncogenes.

1.4.4.1 Tumour Suppressor Genes

Typically, tumour suppressor genes are guardians of the normal functioning of cell replication, so named because when they function normally, they suppress tumour growth. Tumour suppressor genes have a wide range of functions. The first group is involved in the DNA damage response pathways identified above and so ensure stability of the genome. For example ATM, which is implicated in leukaemia, is a critical component of the DDR signalling pathway [89][90]. A second group regulates the cell cycle, preventing excessive growth, and initiating senescence or causing apoptosis in cells that have accumulated too much damage. For example, RB1 inhibits cell cycle progression until the cell is ready to divide and is commonly implicated in cervical tumours [59]. Tumour suppressors and oncogenes can be locked in regulatory pathways, whereby the tumour suppressor antagonises the activity of the oncogene. For example, PTEN, which is implicated in breast cancer, reduces the activity of the oncogenes PIP3 and AKT [91].

In each case it is normally a loss of function mutation which drives the cancer, and commonly the loss of function is recessive: it must occur in both alleles to be effective. This is rarely through two separate mutations. Instead the disabling mutation often occurs to a

gene before loss of heterozygosity means that the disabling mutation is copied to the sister chromatid. Not all mutations will inactivate a tumour suppressor. However, the range of effective mutations is quite extensive and includes deletions, frameshift indels, and nonsense substitutions as well as some pathogenic missense substitutions[92].

1.4.4.2 Oncogenes

In contrast to tumour suppressors, proto-oncogenes are genes involved in cell differentiation and proliferation of cells that, when overactivated, support tumorigenesis. The process of over-activation may happen in a number of different ways. There may be overexpression either as a result of amplification of copy number (such as MYC [93]), change in methylation (e.g. HOX11 [94]). Occasionally active domains from two separate genes become fused, again as a result of chromosomal translocation, giving the new gene a distinct function, or enabling the resulting protein to signal in the absence of ligand-binding. A common example of this is the BCR-ABL gene which is found in Chronic Myeloid Leukaemia. BCR-ABL results from a fusion of BCR on chromosome 22 with ABL on chromosome 9 leading to production of a tyrosine kinase which is always on[95]. More commonly, a relatively small alteration to the structure of the resulting protein as the result of a missense mutation or inframe indel enables the protein to become substantively active. One example of this is HER2 which can dimerize with a number of related proteins, and is implicated in breast cancers[96]. Such mutations need not affect both alleles, and generally they are very specific missense mutations local to a few highly clustered mutation sites [51][97].

This distinction between the mutations that affect the two types of driver genes led to the 20/20 rule – Vogelstein's highly effective rule of thumb that any commonly mutated gene where more than 20% of mutations were inactivating was a tumour suppressor and conversely any commonly mutated gene where more than 20% were recurrent missense mutations was a proto-oncogene [98]. A drawback to this approach is that a few genes – such as TP53 – display both oncogenic and tumour suppressor like features [99].

1.4.5 Mutational signatures

One of the most promising avenues for making sense of the array of differences between cancers is the work emerging over the last six years on mutational signatures. Statistical clustering of mutational frequencies carried out using COSMIC and other large cancer mutation data sets [100]–[103] has enabled the identification of specific profiles of patterns in substitutions, indels and large scale rearrangements. For single nucleotide substitutions, these signatures are derived from the mutational fingerprints of each sample/cell-line. These fingerprints are frequency counts of each the twelve possible substitutions, within the context of the nearest nucleotide neighbour of either side i.e. AAA>ACA …TTT>TGT. If no further assumptions are made this gives rise to a 196-dimensional vector. However, a common assumption is that substitutions are equally likely on either DNA strand. Thus, the mutation AGT>AAT on the leading strand is treated as the same as the mutation TCA>TTA on the lagging strand. Conventionally all mutation frequencies are converted to have the wild-type C or T, reducing dimensionality to 96. Mutational signatures are then formed by clustering the fingerprints using Non-negative matrix factorisation (see section 1.9.2). A typical pictorial representation is shown in figure 1.2 below.

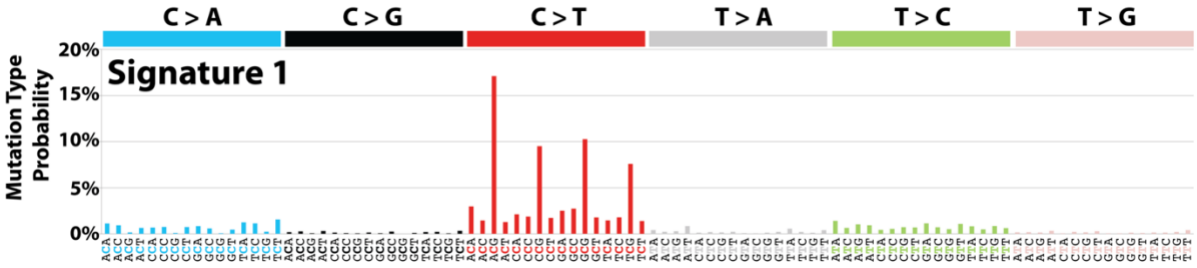


Figure 1.2: Mutational cancer signature 1 – The signature shows the relative frequencies of the six substitution types that have a C or T wild type, within the context of different neighbouring nucleotides. This signature found in most cancers and is associated with aging. It was described in “Clock-like mutational processes in human somatic cells” [104] and reproduced from the COSMIC mutational signatures[105].

1.4.6 Genetic Interactions in cancer cells

Cancer therapies which aim to reduce the impact of oncogene activation typically work by using small molecules to directly inhibit the active domains of key oncogenes in signaling networks. However, these inhibitors do not provide a sufficient range of therapies. Not all oncogenes are druggable (for example, neither RAS nor MYC have so far proved druggable), responses are often partial and cancers often acquire resistance to therapies. It is therefore important to find alternative therapies that allow the targeting of mutated tumour suppressor genes. In order to do this, it is necessary to make use of genetic interactions.

When a double mutation produces a phenotype that is surprising in light of the effect of each individual mutation the phenomenon is said to be a genetic interaction. A double mutant which results in a more extreme outcome than expected is considered a synthetic

interaction, whereas if the impact of the double-mutant phenotype is less severe than expected, it may be considered to be an alleviating interaction [106].

The most extreme case, synthetic lethality, is defined as being where loss of either of two genes individually has little effect on cell viability but inactivation of both genes simultaneously leads to cell death. This concept is useful therapeutically because it implies that if a tumour suppressor is inactivated, then inhibiting the protein products of its synthetically lethal partner would lead to death of cancer cells sparing those cells where the tumour suppressor is not inactivated [107]. This is shown schematically in figure 1.3.

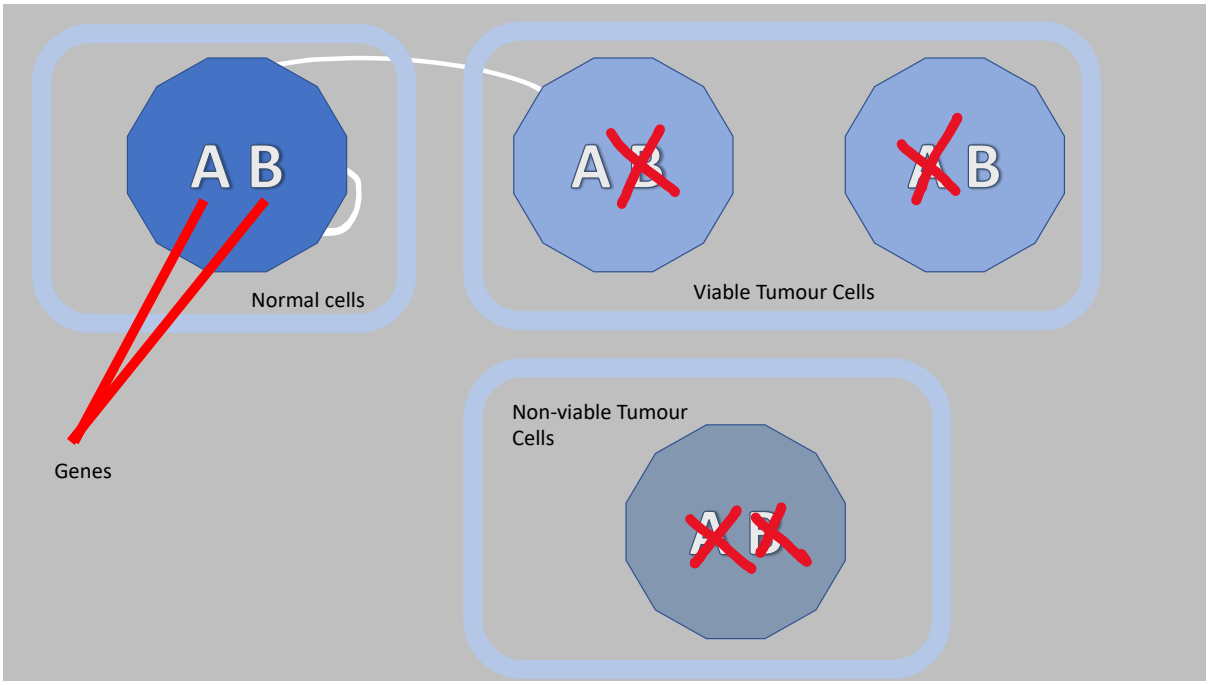


Figure 1.3 legend – Synthetic Lethality schematic. Two genes, A and B may be considered synthetically lethal if a tumour cell where either gene is inactivated remains viable, but becomes unviable if both genes are inactivated.

Application of the concept of synthetic lethality to drug discovery has not progressed far. Only one synthetic lethal pair, that between BRCA and PARP, has resulted in the much-awaited personalised therapies [108]. This is because the gene pairs that appear synthetically lethal in model organisms such as yeast and fly, have not translated well to

human cancers. Moreover, even within cancers, synthetic lethality depends on the specific cell under consideration.

1.5 Overview of bacteria

As well as mutations in cancer cells this thesis looks at patterns of mutations in bacterial strains. Although several thousand bacterial species exist in the human gut alone, and bacteria are essential for all animal life, there are around one hundred that are responsible for human disease: either through tuberculosis and pneumonia, via food poisoning from bacteria such as *Salmonella* or *Listeria* or infections. The most deadly, *Mycobacterium tuberculosis*, alone causes some 1.8 million deaths per year [109].

Antibiotic resistance is a major and growing concern across the globe. In 2017 the World Health Organisation (WHO) agreed a priority list of bacteria other than mycobacterium species for which new treatments are urgently needed [110]. Eight of the bacteria that I include in the analysis are on that list. See table 1.2 below. The bacteria studied also have more than 200 published exomes in Ensembl[111].

Phyla	**Class**	**Species**	**WHO Research category**
Actinobacteria	Actinobacteria	*Mycobacterium tuberculosis*	Globally established priority
		Mycobacterium abscessus	Globally established priority
Firmicutes	Bacilli		
		Bacillus cereus	
		Enterococcus faecalis	
		Enterococcus faecium	High
		Listeria monocytogenes	
		Streptococcus pneumoniae	Medium
	Clostridia		
		Clostridioides difficile	
Proteobacteria	Beta proteobacteria		
		Neisseria meningitidis	
		Neisseria gonorrhoeae	High
		Burkholderia pseudomallei	
	Gamma proteobacteria		

	Acinetobacter baumannii	Critical
	Escherichia coli	Critical
	Klebsiella pneumoniae	Critical
	Pseudomonas aeruginosa	Critical
	Salmonella enterica	Critical

Table 1.2. Importance of research into bacterial species to meet World Health Organisation goals [110].

Despite their ability to cause disease, in many cases these bacteria are found in healthy human micro-flora, particularly in the lower intestines, the upper airways or skin. I include here several bacteria that can be found in varying proportions in the lower intestine, where they are frequently asymptomatic. These include *Clostridioides difficile, Escherichia coli*, *Listeria monocytogenes, Klebsiella pneumoniae*, *Salmonella enterica* as well as *Enterococcus faecalis*, and *Enterococcus faecium.* In the case of *Streptococcus pneumoniae, and Neisseria meningitidis* the bacteria can occur without causing disease in the airways. These bacteria are opportunistic, becoming pathogenic when the constituent parts of the microflora are badly perturbed, in persons with compromised or undeveloped immune systems, or when introduced to normally sterile parts of the body [112][113][114][115][116]. I also consider two obligate pathogens, Mycobacterium tuberculosis and Neisseria gonorrhoea. These normally reside in the human body but must cause symptoms in order to be transmitted [117] [118].

In some species, such as *Bacillus cereus*, *Pseudomonas aeruginosa*, *Burkholderia Pseudomallei* and *Mycobacterium abscessus*, the bacteria are primarily found in the environment, most particularly in soil and water. These bacteria then become pathogenic when introduced to the body in sufficient quantity via ingestion (*B. cereus*, *L. monocytogenes*) , infection (*M. abscessus*), inhalation (*P. aeruginosa*). In some cases more than one route is available (e.g. *B. pseudomallei*) [119]–[123] [124]. However, despite these apparently clear distinctions most pathogenic bacterial species, can survive for months on dry surfaces [125].

1.5.1 Bacterial DNA

Bacteria generally have a single circular chromosome, though variations on this theme exist. Some, such as Burkholderia species, may have up to three or even more chromosomes containing essential genes and some, such as *Neisseria gonorrhoeae*, are polyploidy. Bacterial DNA also includes smaller circular sections of DNA known as plasmids. These do not contain essential genes and vary from strain to strain. They are readily picked up from other bacterial cells or the environment via conjugation, and lost again during replication. The plasmids are copied independently of the chromosome and numbers thus vary over the life of a single cell [126].

A typical bacterial genome includes around 5000 protein coding genes from within a genome of around 5 mbps (see table 1.1). However, this figure is deceptive as the range is considerable, both between bacterial species and within them. For example, a comparison of more than 2000 *Escherichia coli* genomes found that while there were just 3100 genes, or slight variants thereof, that were present in all of the strains, in total there were roughly 89,000 genes that were present in some but not all of the strains.

The size of bacterial genomes ranges from around 15 mbp with 12,000 genes (a *Sorangium cellulosum* strain) to 112 kbps with 137 proteins (*Nasuia deltocephalinicola* strain) with the larger genomes tending to be found in bacteria that inhabit more complex environments. There tends to be roughly 1 gene per thousand bps. The bacterial species that I study here include both bacteria with low genetic variation e.g. *Mycobacterium tuberculosis,* as well as *Burkholderia pseudomallei* where the variation in genome size can be more than 1 mbps [127].

This variability can mean that determining species can be difficult. A number of different approaches has been taken: the sequence identity of conserved RNA gene 16S is used to define new species with 97% taken as the cut-off, and phylogenetic profiling is possible using groups of conserved genes. The more genes that are considered the less sensitivity to horizontal gene transfer (see below) [127][128].

Bacterial name	**Genome Length** [129]	**Protein coding Genes** [129]	**Chromosomes**
Acinetobacter baumannii	3,844,542	3,475	Single circular [130]
Bacillus cereus	5,462,435	5,283	Single circular [131]
Burkholderia pseudomallei	7,161,665	5,571	Two chromosomes [132][133]

Clostridioides difficile	4,268,322	3,815	Single circular [134]
Enterococcus faecalis	3,106,425	2,734	Single circular [135]
Enterococcus faecium	2,807,001	2,728	Singular circular [136]
Escherichia coli	5,443,340	5,494	Single circular [137]
Klebsiella pneumoniae	5,781,501	5,514	Single circular [138]
Listeria monocytogenes	2,776,517	3,131	Single circular [139]
Mycobacterium abscessus	5,158,669	5,131	Single circular [140]
Mycobacterium tuberculosis	4,327,834	4,040	Single circular [141]
Neisseria gonorrhoeae	2,151,002	2,209	Single circular but polyploidy [118]
Neisseria meningitidis	2,188,020	2,461	Single circular [142]

Pseudomonas aeruginosa	6,341,502	5,718	Single circular [143]
Salmonella enterica	4,786,542	4,471	Single circular [144]
Streptococcus pneumoniae	2,188,259	2,080	Singular circular [145]

Table 1.1 Bacterial chromosomal statistics for reference genomes

1.5.2 Overview of mutations within bacteria

As with human cancer cells, mutations via both indels and point mutations are possible. However, there are a couple of important distinctions. Firstly, most bacterial genes are haploid and thus, without a second intact copy of the gene, the distinction between mutations in dominant and recessive genes is lost. All mutations can have a faster phenotypic impact.

Secondly, large chunks of DNA can be taken up into the cell from the environment through horizontal gene transfer. Bacterial cells excrete DNA either when they die, or during their life and this extracellular DNA may be persistent for minutes or even hours. During this time the DNA will come into contact with living cells, some small percentage of which may be 'competent', i.e. able to take up new DNA, converting it to single stranded DNA as it does so. DNA may then reform self-replicating plasmids or, if there is sufficient similarity between the host cell and new DNA, be incorporated into the main chromosome through homologous recombination.

This thesis looks just at DNA substitutions and I have thus looked at the genes shared across all the strains within a given cluster, and unique mutations within them. This should avoid problems introduced by horizontal gene transfer and the distribution of different alleles within strains.

1.6 Sources of data

Gene essentiality data was downloaded from the Achilles depmap portal [81] at *https://depmap.org/portal/achilles/.* The Achilles depmap project catalogues gene essentiality across hundreds of genomically characterized cancer cell lines. Gene essentiality is determined using the viral introduction of short hairpin RNA or CRISPR/Cas9 libraries to silence or knock out individual genes in hundreds of cells where the other features of the cancer cell line are known. By doing so it enables the identification of those genes that affect the survival of the cell, and association of them with specific cancers.

Methylation links with gene expression was downloaded from Broad Institute TCGA Genome Data Analysis Center [146] at *http://firebrowse.org/?cohort=ACC# etc.* The Broad Institute pipeline provides access to analysis of TCGA data for thirty-eight TCGA projects. This analysis spans mutation, copy number analysis, gene expression and methylation and includes the Spearman correlations between beta levels of methylation at CpG sites and mRNA expression levels.

Cancer mutations, the cancer gene census and probabilities of different mutational signatures was downloaded from the Catalogue of somatic mutations in cancer

(COSMIC)[147] at https://cancer.sanger.ac.uk/cosmic/ . COSMIC enables downloads of somatic mutations and larger genetic aberrations such as CNVs and fusion information, together with some data on gene expression and methylation levels. COSMIC also curates the mutational signatures and the cancer gene census, as well as providing several visualisation tools. File transfer using the FTP protocol is possible using a generated password.

A database of drugs that target specific genes was downloaded from the Drug gene interaction databank DGIdb [148] at https://www.dgidb.org/search_interactions DGIdb collates information on drug-gene interactions and druggable genes from forty-one sources allowing TSV data downloads.

Nucleotide and amino acid sequences, annotation, and gene and protein identification information was downloaded from Ensembl including Ensembl bacteria[129] at https://bacteria.ensembl.org/index.html and Ensembl Biomart[149] which is available via https://www.ensembl.org. Ensembl Bacteria enables the download of genome sequence and annotations of protein-coding and non-coding genes from bacterial strains. An ftp site is available for bulk download. Biomart provides the gene and protein annotations, including alternative names, positions and sequences for 77 animals, and the genome database enables the download of the entire genome GRCh38.

TCGA methylation data was downloaded the Genomic Data Commons (GDC)[150] at https://portal.gdc.cancer.gov/. The GDC provides sequencing information, gene expression, CNV, and methylation data for over 84,000 cancer samples together with some limited

clinical information about the case, statistics and visualisation tools. Tools are available to enable the download of bulk information.

Protein clusters were downloaded from the STRING database [151] at *https://string-db.org/*. String-db collates and presents information about physical and functional protein-protein interactions (PPI), and provides computational predictions about protein clusters, together with functional enrichment analysis on the basis of gene ontology. Functional PPIs are provided down to roughly the granularity of a KEGG pathway map.

1.7 Machine Learning, statistical and predictive techniques

Unsupervised machine learning methods are used throughout this work to cluster results to provide meaningful insights. The most commonly used examples are explained below.

1.7.1 FATHMM

The COSMIC database of somatic mutations in human cancer samples is used extensively throughout this thesis. Most of these mutations are missense substitutions. The majority of these will have little or no impact on the function of the gene's protein product. However, some will be pathogenic. To distinguish between these two fates, I use the FATHMM score provided as part of the COSMIC database.

FATHMM is a pathogenicity prediction model. It works by using a Hidden Markov Model to identify homologous sequences. Such models assume that there is some hidden stochastic process which is not directly observable, but about which inferences can be made on the basis of the observations. On the basis of the preceding sequence of amino acids, the HMM

predicts the amino acid of interest. If the probability of the mutated amino acid is less than the probability of the wild-type amino acid, the substitution is assumed to be deleterious, otherwise not. Prediction thresholds in FATHMM were set to maximise sensitivity and specificity against a known set of disease associated amino acid substitutions.

1.7.2 Non-negative matrix factorization

I make use of the work pioneered by Alexandrov et al. to identify mutational signatures from missense mutations in cancers and extend this to look at those that occur during bacterial evolution[152]. In bioinformatics terms, mutational signatures are clusters of often many hundreds of mutational fingerprints. These mutational fingerprints are matrices in S by 96 dimensions, where S is the number of samples. Dimensionality reduction is then performed by using non-negative matrix factorisation (NMF) which enables the splitting of one matrix V of high dimensionality into 2 non-negative matrices W and H of lower dimensions plus an error term (see figure 1.4).

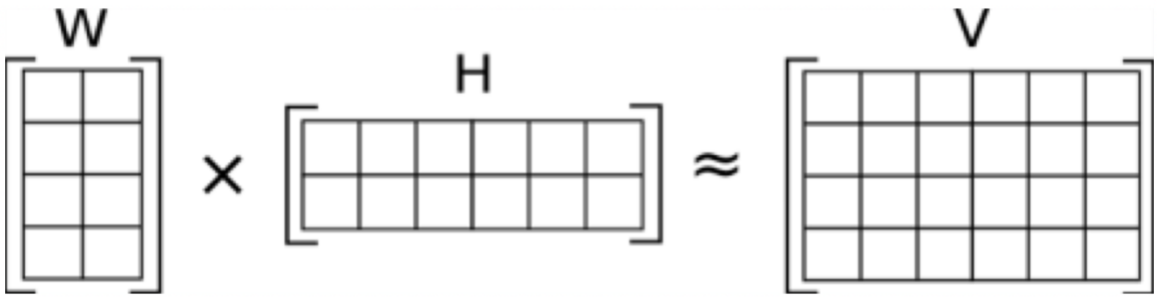


Figure 1.4: Schematic of non-negative matrix factorisation.

For mutational signatures the matrix V would be the mutational fingerprint matrix formed by S samples, each of 96 mutational frequencies. V is approximated by the multiplication of two matrices, W and H, of much smaller dimension. W is an n by 96 -dimensional matrix where n is the number of mutational signatures. Each column of mutational frequencies describes one of the signatures. H is an S by n -dimensional matrix. Each column provides a breakdown of a single mutational fingerprint into its constituent signatures. This has the benefit of clear interpretation: each mutational fingerprint can be described as a linear superposition of signatures. Thus, mutational signatures form an approximate basis for the vector space. The scipy package scipy.decomposition.nmf is used to calculate H and W.

1.7.3 MMseqs2

In the chapter on "Mutational signatures in bacteria", I make extensive use of MMseqs2 to find the orthologs of genes across hundreds of bacterial genomes at a time. MMseqs2 is a very sensitive and very fast 'BLAST-style' alignment algorithm for aligning large numbers of amino acid sequences [153]. The alignment itself is carried out by a Smith-Waterman local alignment, but in order to speed things up considerably MMseqs2 has a filtering stage which looks at the distribution of similar words each made up of 7 amino-acids (similar-k-mers). It finds all the target sequences that have consecutive similar-k-mers on the same diagonal. The Smith-Waterman alignment is then carried out on the filtered sequences. The increased speed of alignment allowed me to make extensive use of MMseqs2 to identify orthologous genes in different bacterial strains, clustering amino acid sequences from the genomes of 200 bacterial strains at a time. Genes were first translated into amino acids then clustered using MMseqs2 before using clustalo to align the orthologous gene families. I also used MMseqs2 to identify DDR genes for each of the bacterial species.

1.7.4 Clustalo

In order to identify the missense mutations in orthologous genes I did many thousands of global nucleotide alignments on the genes from up to 200 strains at a time. This was possible using the stand-alone Clustalo package. Clustal omega aligns sequences in progressively larger and larger subalignments, using a guide tree to determine the order. The process is accelerated by the use of reference sequences: the guide tree is based on the distance from each sequence to each of the reference sequences [154].

1.7.5 Hierarchical clustering

In the chapter on "Using mutational signatures in cancer to explore tissue specificity of driver genes" tissue-types are clustered according to the prevalence of mutations in cancer-associated genes, and also according to the mutational signatures present. In chapter 3 bacterial strains are clustered according to the similarities of sequence identity in 100 orthologous gene families. In these cases, I use the python scipy package to perform hierarchical clustering using the Ward's method. Ward's algorithm minimises the variance (i.e. the squared Euclidean distance) between different members of the same cluster. It starts with clusters of size 1 and successively joins the two clusters that minimise the variance.

1.7.6 Statistical methods

A number of statistical methods, distributions and tests are used throughout this thesis. A few of the most important ones are described here:

Cosine similarity- The cosine similarity of two positive vectors gives the cosine of the angle between two vectors as a measure of the extent to which two vectors are aligned. The measurement is bound between 0 (perpendicular vectors) and 1 (parallel vectors).

Hypergeometric distribution- Given a population of N objects where K have a particular feature, the hypergeometric distribution describes the probability that if n objects are drawn without replacement, randomly and independently, then k of the objects will have the desired feature. This distribution is often used to describe the distribution of mutations in pairs of genes where there is no genetic interaction. However, it relies on the assumption that a genetic mutation is equally likely in any of the samples.

Binomial distribution - Given a series of N Bernoulli trials where each probability of success does not vary from trial to trial then the poisson binomial describes the probability that there will be k such successes.

Poisson binomial distribution- Given a series of N Bernoulli trials where each probability of success varies then the poisson binomial describes the probability that there will be k such successes. This distribution can be used to describe the distribution of mutations in pairs of genes where there is no genetic interaction and does not rely on the assumption that a genetic mutation is equally likely in any of the samples.

Normal distribution – is the familiar non-skewed bell-shaped distribution. It is primarily of interest because of the central limit theorem. The CLT states that where sufficient

independent samples are drawn from a large population then the sample mean will approximate a normal distribution.

As such it is a good distribution for describing error terms, where the error terms may be assumed to result from a large number of independent errors.

Chi square distribution (and test) – If two or more independent variables Z_i have a normal distribution then the sum of the squares will be distributed according to the chi square distribution. Thus, if measurement errors associated with two variables are assumed to be normally distributed then the goodness of fit test associated with the chi square distribution allows us to test the hypothesis that the two variables are independently distributed.

1.8 Project aims

In this thesis I explore a range of ways in which patterns of mutations and other genetic and epigenetic alterations are shaped by DNA, and by the other alterations taking place within the cell. My primary focus is on exploring the patterns that underlie the heterogeneity of cancer cells, both at the level of individual indels and in terms of mutual exclusivity and cooccurrence of genes. Here I am particularly driven by the need to find new synthetically lethal gene pairs that are susceptible to translational drug therapies. I also developed novel mutational signatures for bacterial species. All code is written in python and is available via bitbucket at https://bitbucket.org/bioinformatics_lab_sussex/workspace/projects/PAT

*1.8.1 **Chapter** Error! Reference source not found.*

DNA mutations in cancer samples occur in a way that is governed both by the need for the cell to survive and ultimately become a dominant clone, and also by the different proclivities

of different nucleotide patterns to mutate. In this chapter, I explore the separate contributions that these two pressures put on the distribution of substitutions and indels, in the COSMIC database. I analyse the patterns of mutations in the context of the preceding six nucleotide, identifying the distribution of indel lengths and exploring reasons for the discrepancy between numbers of inframe indels and frameshift indels. I look both at the evidence that the discrepancy is caused by availability of homology hooks needed for replication slippage and NHEJ. I also explore the possibility that the discrepancy reflects the inability of cells to tolerate frameshift indels. To do this, I compare the prevalence of di- and tri- nucleotide indels in the coding region, where the indels are subject to evolutionary pressure, with those in the non-coding region where they are not. To remove the bias introduced by replication slippage I look just at those indels that are not at a mono-nucleotide repeat.

1.8.2 Chapter 3

Different tissue types give rise to different numbers and types of mutations and to alteration of different driver genes. There is good reason to assume that the two might to be connected. The mutations that arise reflect not only DNA damage but also the pathways available for repair, so mutations in DDR genes could affect the mutational profile seen. On the other hand, not all mutations can make the alterations needed to inactivate tumour suppressors or activate pseudo-oncogenes, so the mutational profile could impact on the driver genes brought into play. Here, I look at both whether the numbers and types of mutations influence the driver genes that are altered, and also at whether the driver genes altered change the mutational profile.

I find limited evidence that the mutational status of the genes is predictive of mutational profiles, but much clearer evidence that around two thirds of the cancer-associated genes appear to be partly opportunistic. I also see in TP53 a clear distinction between those sites where nonsense mutations are preferred and others where nonsense mutations are under-represented suggesting potential oncogenic action.

1.8.3 Chapter Error! Reference source not found.

I then extend the mutational signature techniques used in previous chapters, using them not for cancer cells but for SNVs in sixteen bacterial species. I build up mutational fingerprints using just those mutations that are both unique to a solitary bacterial strain and silent substitutions. I normalise the fingerprints by the expected distribution of mutations given the nucleotide make-up of the bacteria. Signatures are found, as before, using NMF. I find that there are commonalities between some of the bacterial mutational signatures and some of those seen in cancer cells, or as a result of the treatment of stem cells with mutagens. In particular, some of the signatures are similar to those caused by incorrectly repaired alkylation.

1.8.4 Chapter 5

Finally, I look for potential ways of using existing drugs by finding mutually exclusive gene pairs between genes which are likely to be tumour suppressors and genes where the protein products are known to be druggable. I extend the omic data sets commonly used to include gene inactivation through methylation. I then simplify use of the Poisson binomial statistical test to deal with large data sets. I show that its use is superior to the use of the hypergeometric test by simulating realistic gene/sample sets. I then use both the Poisson

binomial test and the hypergeometric test to find mutually exclusive or co-occurring gene pairs and associate the druggable gene with existing drugs by reference to the Drug Gene Interaction databank (DGIdb). I use my new website *MexDrugs* to display the results and allow downloads.

2 Deciphering the influence of the exome on mutations

2.1 Abstract

Improving our understanding of DNA structures and patterns, and their impact on mutations, is allowing us to better understand which areas of the genome are most at risk from different types of mutations, and the mechanism of damage.

In this chapter, I analyse the small mutations from 27,000 exome-wide sequences in the COSMIC database. I show that the nucleotide patterns found at the site of short insertion and deletion mutations (indels) have more influence on mutational patterns than those at the site of substitution mutations. I find that somatic indels are predominantly single nucleotide deletions, and less frequently insertions, arising at the site of mononucleotide repeats. The risk of other small indels occurring was also increased when there were duplicates of that sequence pattern in the underlying genomic sequence. In order to assess the impact of selection pressure on frameshift indels I calculate the number of indels that are not next to repeated motifs (and thus cannot be explained by replication slippage) and find that there is an excess of trinucleotide indels (i.e. in-frame indels) compared to dinucleotide indels (i.e. frameshift indels) in protein coding regions where the DNA mutations exert a high selective pressure, compared to non-protein-coding regions where selective pressure is largely absent. I believe that this is evidence of the cells' reduced ability

to tolerate frameshift indels and is evidence of negative selective pressure in cancer evolution. Finally, few samples have lots of indels and a predominance of insertions and I find evidence that this is associated with the loss of both mismatch repair and homologous recombination.

2.2 Introduction

DNA is continually subjected to mutations as a result of both exogenous and endogenous processes, each process leaving a distinctive pattern of scarring of the DNA [155], [156]. Where the cell dies as a result of the mutation, the mutational record is lost, whereas the mutations that drive cancer will occur more often than expected by chance. Analyses of cancer samples show clear evidence of positive selection for the mutations that drive cancer [98]. However, the vast majority of mutations in a cancer confer no selective growth advantage to the tumour [157]. The accumulation of many hundreds of passenger mutations reduce the fitness of the cell [158], but analysis of the ratio of non-synonymous mutations to synonymous mutations suggests that cancer cells are surprisingly good at tolerating passenger mutations.

The 2017 Bakhoum and Landau's analysis of substitution mutations showed only limited evidence of negative selection within cancer [159]. Thus the totality of the genetic mutations within cancer cells provides an accurate record of the most common mutagenic processes to which it has been subjected [160]. This means that the frequency and distribution of different types of mutations can cast light both on the type of mutational processes at play in any individual cancer but also on the DNA features that increase the risk of mutation [102], [161]. For example, UV light preferentially induces transition mutations at dipyrimidine sequences, whilst many tobacco by-products damage DNA bases by adding

alkyl or bulky chemical groups to the DNA molecule [162]. DNA is also under attack from endogenous process such as the generation of Reactive Oxygen Species (ROS). These arise as part of the normal function of the cell, but cause damage when levels are elevated during times of oxidative stress caused by factors such as smoke or inflammation. ROS most commonly lead to substitutions of GC base pairs and tandem CC>TT substitutions [92], [163].

The COSMIC database includes exome-wide sequencing from 27,000 tumours of which over 20,000 include whole genome variants [147]. Previous analysis of these exomes has highlighted that several key mutational patterns are observed across all cancers. Single point substitutions are the most frequently observed type of mutation, with the most common type being C>T, G>A, G>T substitutions [164], [165]. More deletions are observed than insertions, and single nucleotides indels are much more common than longer indels [164].

Valuable information is also provided by looking at substitutions in the context of the nucleotides immediately to either side of the mutation for example G[C>T]G. Different mutagens, DNA mutational processes and defects in DNA damage repair processes give rise to distinctly different patterns in the frequency of these substitution motifs [101], [166]. Patterns derived from the frequencies of these mutations are termed mutational signatures and have since been used in a wide range of applications, such as characterising cancer sub-types in breast cancer [165], [167] and in gastric cancer [168] and identifying signatures associated with tobacco smoking [169].

A number of different mechanisms evolved to detect and repair damage to the DNA are error prone. As a result, the pattern of mutations seen in any one cancer sample depends not only on the cell type and initial environmental stresses, but also on the health of the DNA replication and repair systems. For example, indels commonly happen as a result of DNA strand slippage during replication [170] and the repair of double strand breaks via Microhomology Mediated End Joining (MMEJ) [155]. Replication slippage occurs when the DNA polymerase disassociates and then re-engages at the wrong nucleotide in the sequence, resulting in the addition or deletion of a nucleotide on the newly-synthesized strand. The MMEJ mechanism lines up microhomologies between the opposite strands of sister chromatids. Where there is no sequence duplication the MMEJ should proceed in a trouble-free fashion. However, where a sequence is duplicated it is possible for the chromatids to mis-align giving rise to characteristic insertions and deletions [155], [170], [171].

The frameshift indels present in the COSMIC database are of particular interest as they shift the translational reading frame. Although this means that they have the potential to dramatically alter the downstream sequence and alter the position at which the first stop codon is encountered, translation is largely prevented if the frameshift indel is encountered before the last exon. In such cases mRNA is removed via nonsense-mediated decay [172].

Previous work in Frances Pearl's laboratory (unpublished) as well as work by Helleday et al. [173] suggested that indels were over-represented at sites of mononucleotide runs. In addition, it showed that the nomenclature protocol for indels in COSMIC, is that when an

indel is observed in a run of identical bases, the position of the indel is generally assigned to the end of that run.

In this chapter, I analysed the relationship between the frequency and distribution of different types of mutations, and nucleotide motifs in the DNA background. In order to better understand the extent of different types of DNA damage mechanisms, I started by investigating the frequency with which the different sextuplets of nucleotides are found before both indels and substitutions in order to improve understanding of how the make-up of the DNA changes the risk of particular types of mutations. I then considered the extent to which indel frequencies can be explained by recurring repeats of the indel-motif in the DNA background. I postulated that the damage caused by frameshift indels is more pronounced than that caused by substitution mutations and might therefore be more likely to result in cell death, leading to a reduced number of frameshift indels. To test this hypothesis, I looked for signs that frameshift indels are negatively selected for in cancers, when compared with inframe indels. To do this I focused on just those indels that are not next to motif repeats and are therefore not associated with opportunities for replication slippage and MMEJ.

2.3 Materials and Methods

I analysed all the point substitutions and short insertion and deletion mutations from 27,000 whole exomes from the COSMIC database v88 together with the non-protein-coding variants from over 20,000 samples [147]. Table 2.1 shows how many samples were derived from each cancer type. To clean the data, duplicate entries were removed. In addition,

when the mutations were mapped to more than one transcript, only mutations on the major transcript were included.

Tissue type	Sample number
adrenal gland	350
autonomic ganglia	530
biliary tract	472
bone	406
breast	2284
central nervous system	1822
cervix	185
endometrium	337
eye	38
fallopian tube	2
gastrointestinal tract (site indeterminate)	1
genital tract	67
haematopoietic and lymphoid tissue	3830
kidney	1774
large intestine	2088
liver	1846
lung	1629
meninges	65
oesophagus	1125

ovary	701
pancreas	1347
parathyroid	28
peritoneum	11
pituitary	55
pleura	154
prostate	1630
salivary gland	78
skin	1011
small intestine	50
soft tissue	245
stomach	633
testis	17
thymus	27
thyroid	810
upper aerodigestive tract	668
urinary tract	645

Table 2.1: samples from each of the primary cancer sites in the COSMIC database.

To identify if there were sequence dependencies in the local environment, I identified the sextuplet of nucleotides that occurred before the mutation (on both strands) and then analysed the frequency with which each of the sextuplets occurs for each type of mutation. For indels I then counted how often the indel motifs of different lengths are repeated at the

site of the insertion or deletion and compared the results with the frequency with which motifs of different lengths are repeated in the exome, and the non-protein-coding region.

2.3.1 Baseline Frequency of Sextuplets in the Human Exome

The fasta sequence for every exon in the human genome was downloaded from the Ensembl database version GRCh 38.p12 [174] and aggregated to form a baseline human exome. This exome was used to calculate the background frequency for every sextuplet of nucleotides (e.g. AACTTG).

2.3.2 Analysis of sextuplets next to mutations occurring in the COSMIC database

Where an indel occurs next to a repeat, the possibility of ambiguity arises. For example ACACAC-AC is indistinguishable from ACAC-AC-AC. I found that in COSMIC such indels are put at the end of the repeated motif. So, to capture these repeats, for each mutation in COSMIC I identified the sextuplet of nucleotides that occurred before the mutation, looking at both the forward and reversed strand. For example, the mutation c.22_23delGA in transcript ENST00000313386 changes the wild-type sequence TGCGTG-GA-GAGAGA to TGCGTG-GAGAGA. The sextuplet before the deletion GA is TGCGTG whilst on the reversed strand the complementary sequence is TCTCTC -TC-CACGCA and the sextuplet before the mutation is TCTCTC. Mutations were categorised as insertions, deletions and substitutions. Indels were then sub-divided into in-frame and frameshift indels, whilst substitutions were subdivided into missense, nonsense and silent substitutions. Missense mutations were further segregated dependent on the wild type and mutant bases. For each type of mutation, the log frequency distribution of the sextuplet immediately preceding the mutation was calculated and normalised with respect to our baseline human exome.

For the missense mutations, the sextuplets were grouped together by the triplet closest to the mutation. So, for example ATAGGG and ATTGGG were grouped together. Box and whiskers plots were then created showing the log fold enrichment of each triplet compared to the baseline exome frequency.

For each indel I identified how many repeats of the indel motif were present at the mutation site. These results were plotted as heatmaps. For each sequence of length L, I also identified the frequency with which that sequence motif is repeated N times in the exome. By subtracting these two logs I was able to compare log frequency heatmaps, showing the fold enrichment for each indel length L found next to N repeats. This analysis was then repeated for the non-protein-coding genome in each of the 22 chromosomes, splitting each chromosome into 100000 base-pair sections for computational efficiency.

2.3.3 Identification of Mutational Signatures

To identify the mutational signature associated with specific samples, I identified the nucleotides closest to each substitution to associate a quadruplet of nucleotides with each substitution e.g. A (C>A) T. I then counted the number of each quadruplet occurring in the sample. In common with previously published work on mutational signatures [166], I assumed that there was no mutational bias towards the leading or lagging strand and thus reduced the dimensionality of quadruplets from 192 to 96 by enforcing that all the wild type of each substitution is either C or T. The published mutational frequencies for each of the 30 COSMIC Mutational Signatures (version 2) was then used to disaggregate the main

mutational signatures present in each of the COSMIC data samples using non-negative matrix factorisation [167][161][173][101].

2.3.4 Indel sequences

To identify the percentage of indels that have a high sequence identity with the nucleotides on either side I used Needleman and Wunsch alignments. The Needleman and Wunsch algorithm was imported from biopython [175] and used with a global alignment scoring 1 for each identical nucleotide with no gap penalty. The cut-off for a 'good' alignment score was set at 0.8 (that is 80% matching nucleotides). The probability of achieving such a score decreases with length of indel. 0.8 was determined ensuring that any insertions longer than six nucleotides have a less than 5% probability of being ranked as good by chance. In order to determine the probability that our overall results could have been generated by chance, I found the percentage of 'good' alignments for 10,000 strings drawn at random from the exome.

2.3.5 Potential pathogenicity of frameshift indels

Exon positions for each transcript were downloaded from the Ensembl database [174]. For each frameshift indel I identified the corresponding gene, all possible transcripts associated with that gene, and the exon number associated with the indel for each of the transcripts. For each indel I identified whether or not there were any potential transcripts that could be generated where the indel was in the last exon and could therefore escape nonsense-mediated decay.

2.3.6 Links between mismatch repair and indel frequency.

The list of DNA damage repair genes and corresponding pathways was downloaded from the supplementary material from Pearl *et al.* [73]. I then identified all samples within the COSMIC database that have colon cancer and split the samples into two groups – those that have a damaging mutation in any of the genes associated with mismatch repair and those that do not. A damaging mutation was considered to be missense substitution identified by FATHMM as pathogenic, and any frameshift or nonsense substitution in tumour suppressors.

2.3.7 Statistics

Statistical tests were generated using python's scipy statistics library [176].

2.4 Results

2.4.1 The frequency of sextuplets in protein coding regions of the genome

In order to calculate the relative frequency of sextuplets before each mutation I first calculated the frequency of each of the sextuplets in the human exome. The baseline frequency of the 4096 nucleotide sextuplets in the human exome follows a noticeably skewed distribution. The sextuplets that were least frequently observed were predominantly those with embedded stop codons such as TAACGC, TTAGCG, CGTTAG, TAGCGT. It is unclear why there were any such codons, though they could potentially be in non-coding transcripts.

A few sextuplets were highly over-represented forming a long tail to the distribution. The very highest frequencies include many repeated codon pairs resulting in the amino acid

pairs: LL RR GG QQ EE. Whilst codons coding for L R G Q and E are common in the background exome this is still unexpected. If C_a^j is the jth codon coding for amino acid a then the frequency of sextuplets coding for most amino acid pairs is closely approximated by assuming that there is no correlation between occurrence of amino acid:

$$\sum_{i,j} freq(C_{a'}^i, C_a^j) \sim \sum_{i,j} freq(C_{a'}^i) \times freq(C_a^j)$$

However, repeated amino acids are observed more frequently than this formula predicts (see figure 2.1). This suggests that evolutionary pressures have led to a disproportionate number of repeated amino acids in coding regions of the genome.

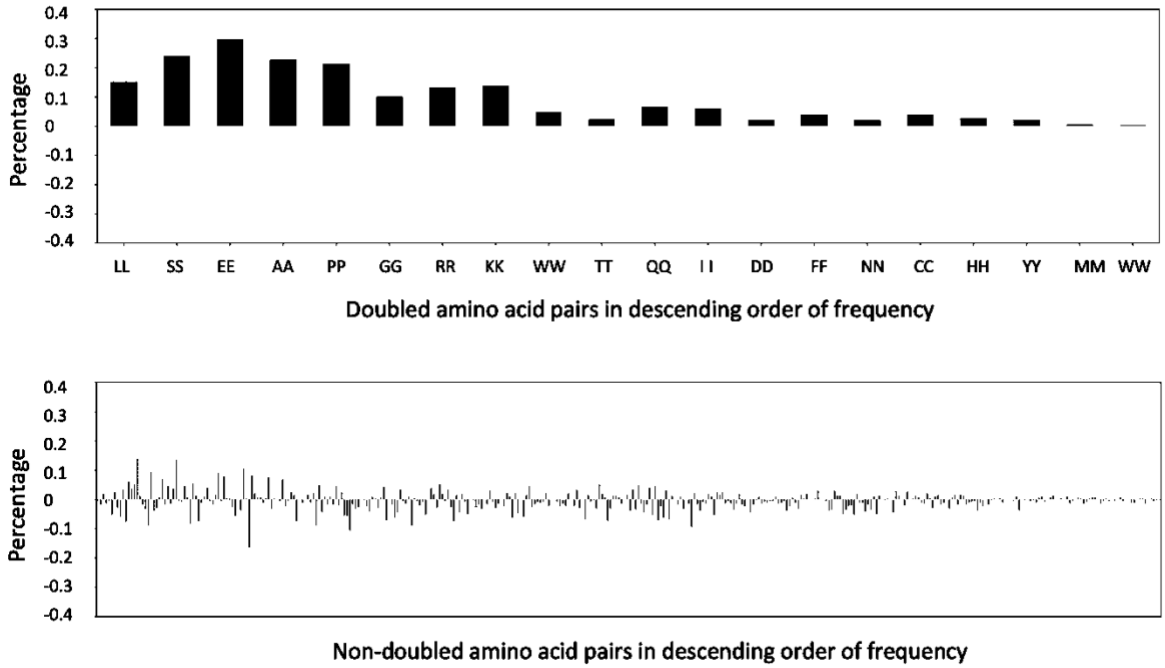


Figure 2.1: The graphs show the difference between observed frequency % and expected frequency % for each of the sextuplets in the exome. The top row shows the difference for

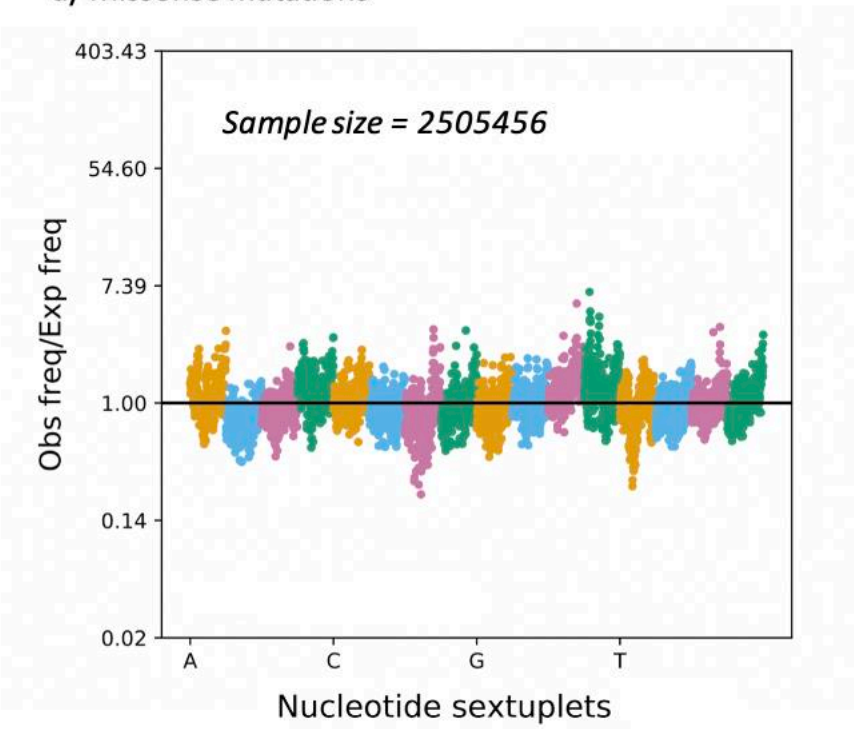
sextuplets which would be translated into repeated amino acids; the bottom row shows the difference for all sextuplets.

2.4.2 Fold enrichment of mutations by sextuplet

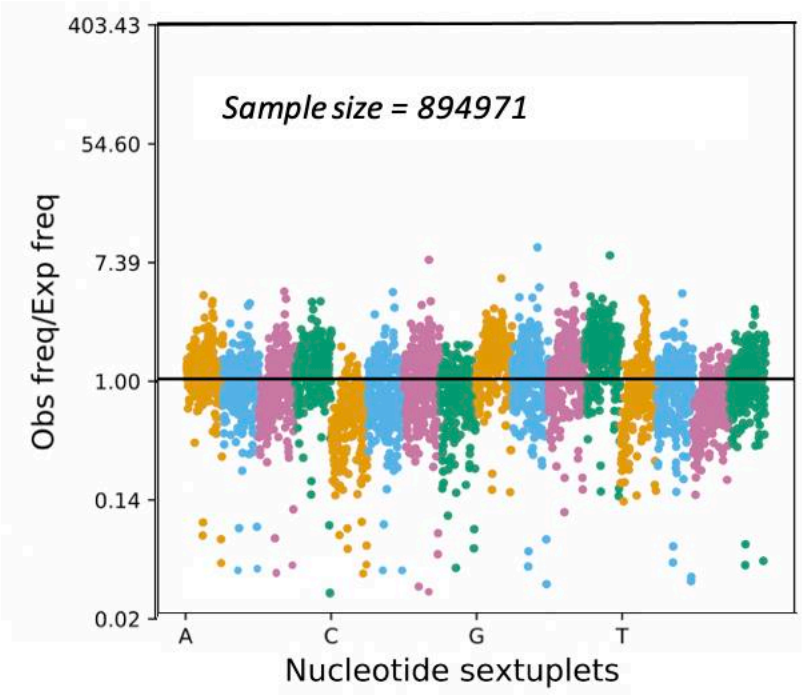
I split the mutational data into the seven major types – substitutions (missense, silent and nonsense), deletions (frameshift and in-frame) and insertions (frameshift and in-frame). For each mutation type I calculated the log ratio of the frequencies of each of the 4096 possible sextuplets before the mutation, in the direction that the gene is read. The frequencies were normalised with respect to the background distribution of sextuplets within the exome.

Initial exploration suggested that indels were more likely to occur next to a mono-nucleotide repeat. There is inherent ambiguity in how such indels are represented, because AAA (insert AA) AAA is indistinguishable from AAA(insert AA). In the COSMIC database such indels are normally represented as occurring *after* the mononucleotide run. I therefore chose to identify the nucleotide sextuplets that occur before each mutation. These frequencies were then normalised by the frequency of the sextuplets in the exome found in the section above. The results are shown in Figure 2.2.

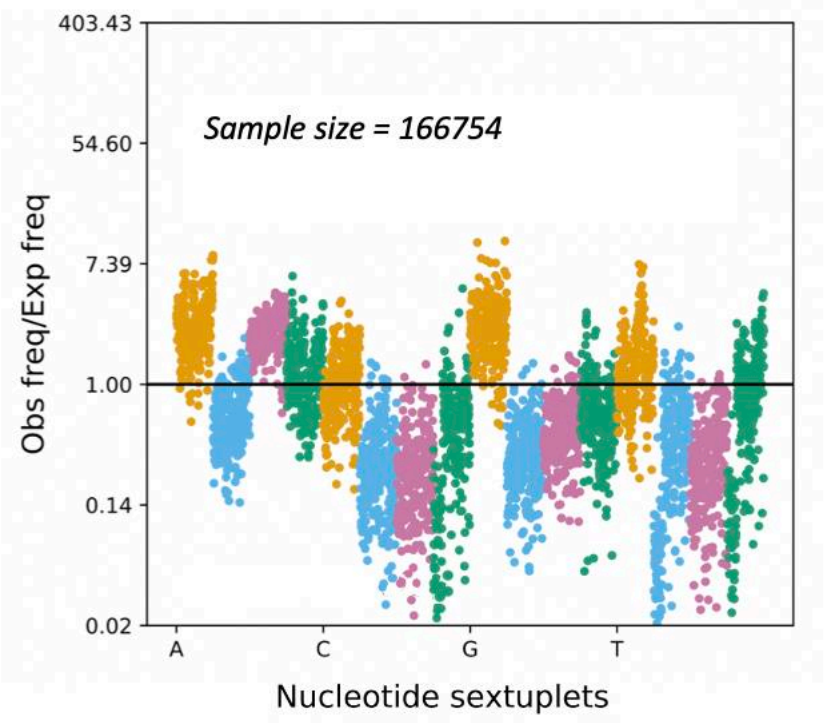
a) Missense mutations



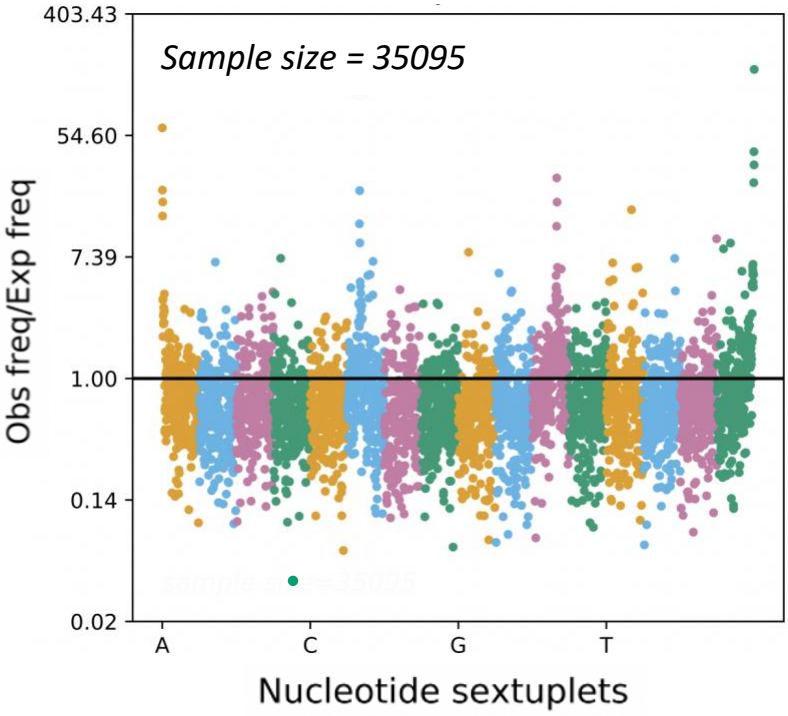
b) Silent mutations



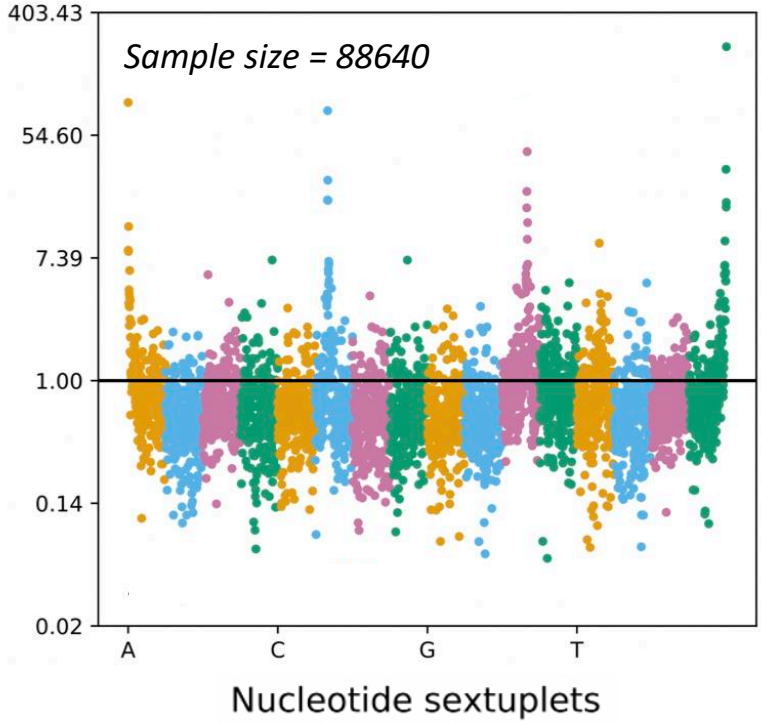
c) Nonsense mutations



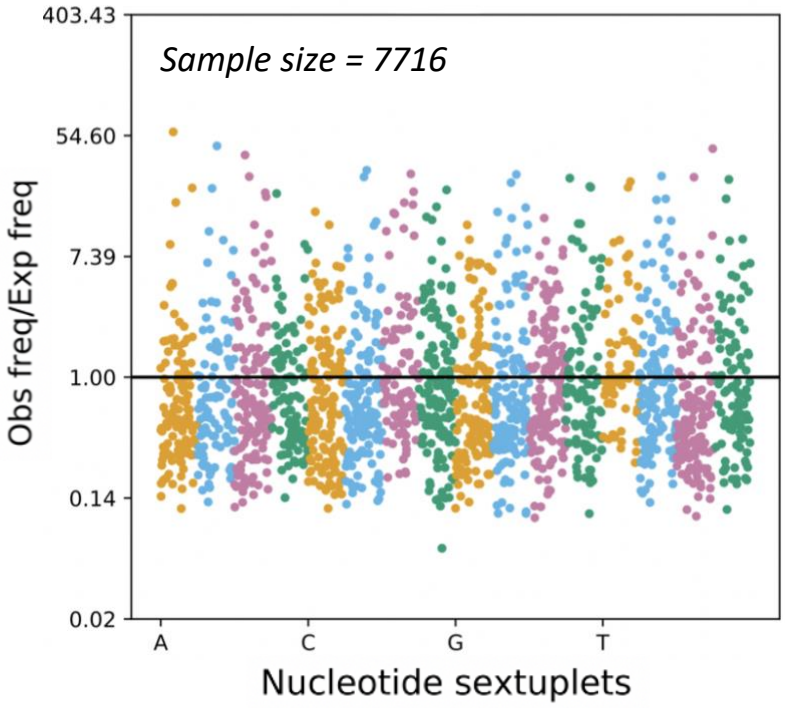
d) Frameshift insertions



e) Frameshift deletions



e) Inframe insertions



f) Inframe deletions

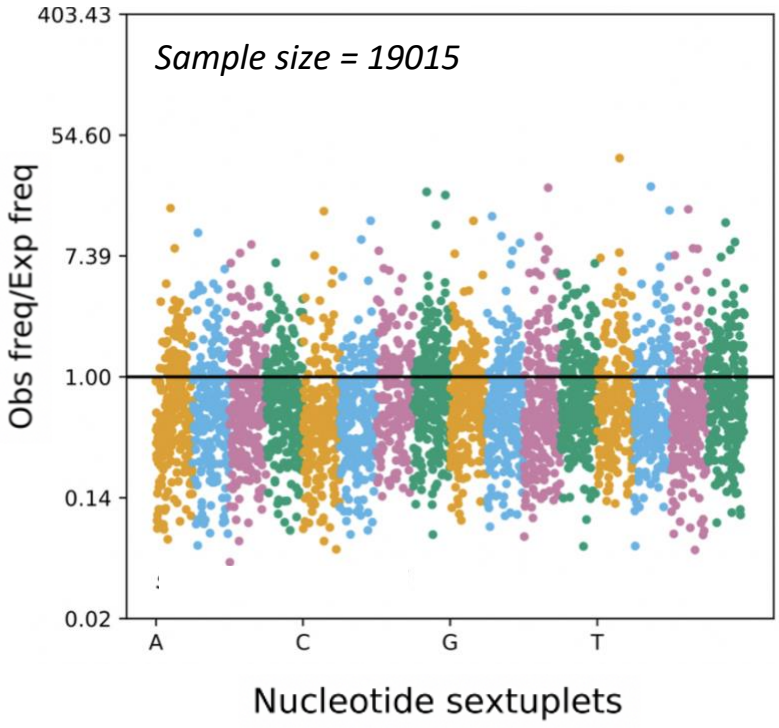


Figure 2.2: The graphs show the ratio of the observed frequency of sextuplet to the expected frequency for each of a) missense mutations, b) silent mutations, c) nonsense mutations, d) frameshift insertions, e) frameshift deletions, f) inframe insertions and g) inframe deletions.

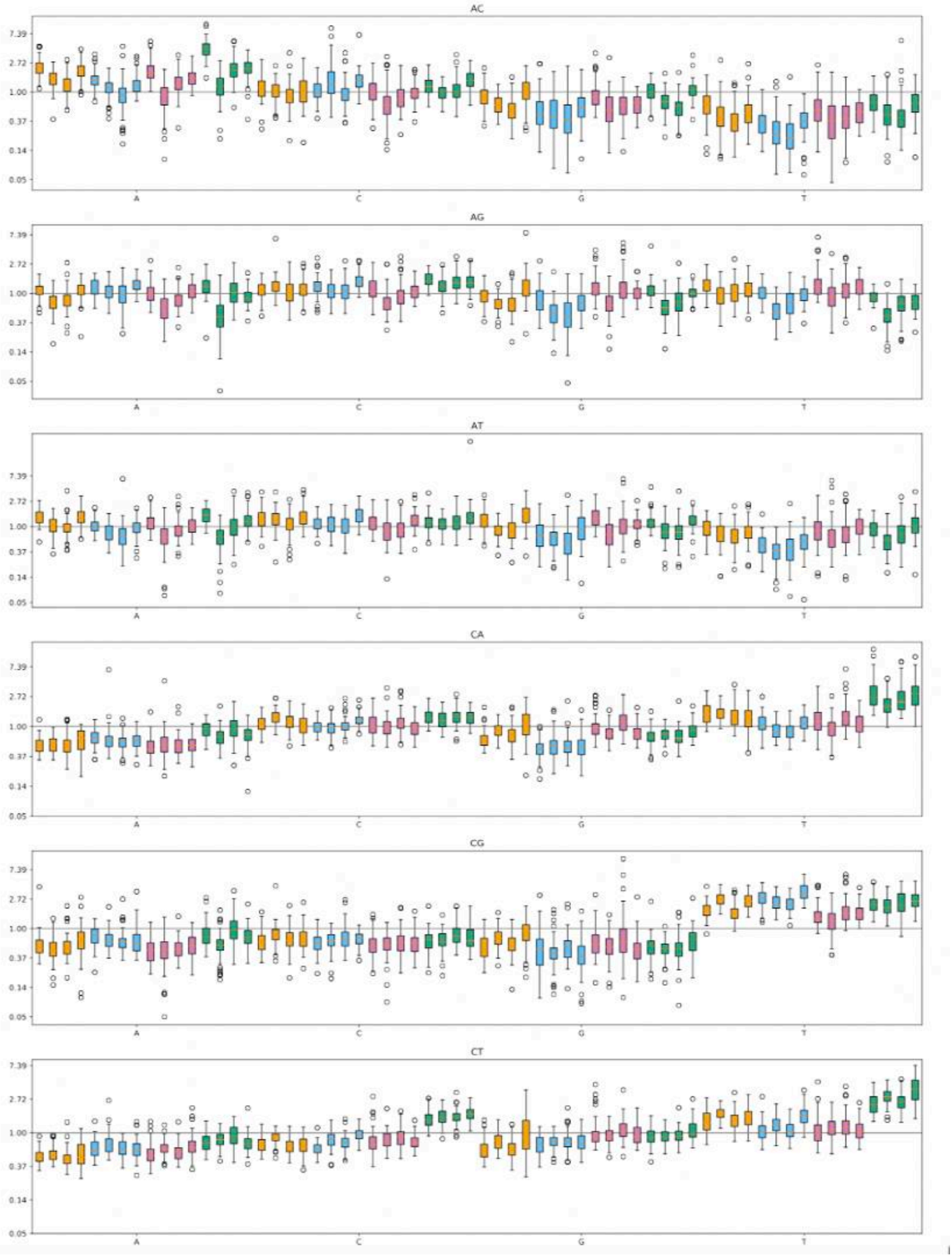
My assumption was that the nucleotides closest to the mutation were the most important. Therefore, in order to group sextuplets such as TCAAAA and GCAAAA together, for each graph the sextuplets were reversed before being ordered alphabetically. The values for each of the 4096 reversed sextuplets from AAAAAA to TTTTTT is shown. The nucleotide closest to the mutation is shown in the x axis and preceding nucleotide is given by the colour: orange A; blue C; pink G; and green T. Values are shown on a log scale.

2.4.3 Substitution Mutations

In total, I analysed over 5 million substitution mutations. Of these 70% resulted in missense substitutions, 25% silent and 5% in nonsense mutations. The sextuplet fold enrichments are seen in figure 2.2. (2.2a -missense substitutions, 2.2b - silent substitutions and 2.2c - nonsense substitutions.

For missense substitutions it is known from existing work on mutational signatures [165], that different mechanisms of damage give rise to different substitution profiles. These mutational signatures look at the nucleotides immediately to either side of substitution mutations segregated by their wild and mutant type, e.g. ACT>AGT. No pattern is discernible from looking at the sextuplets of all the missense substitutions together (see figure 2.2a). I therefore split the substitutions by substitution type (e.g. C>T) and for each group

calculated the log ratio of observed frequency/ expected frequency for each sextuplet, as before. I found that the three nucleotides closest to the mutation have more impact that those more remote, so I grouped together the first three nucleotides, before plotting the resulting values. The results are show as boxplots in figure 2.3. There are two distinct patterns that are outliers: A>T substitutions showed a fold enrichment of 29 after the nucleotide pattern GATTTC. T>G substitutions showed a fold enrichment of 37 after the nucleotide pattern GGCGGG. More generally, I see that cytosine substitutions are more likely after sextuplets that end in mononucleotide runs of thymine.


AC
AG
AT
CA
CG
CT
7.39
2.72
1.00
0.37
0.14
0.05
A
C
G
T

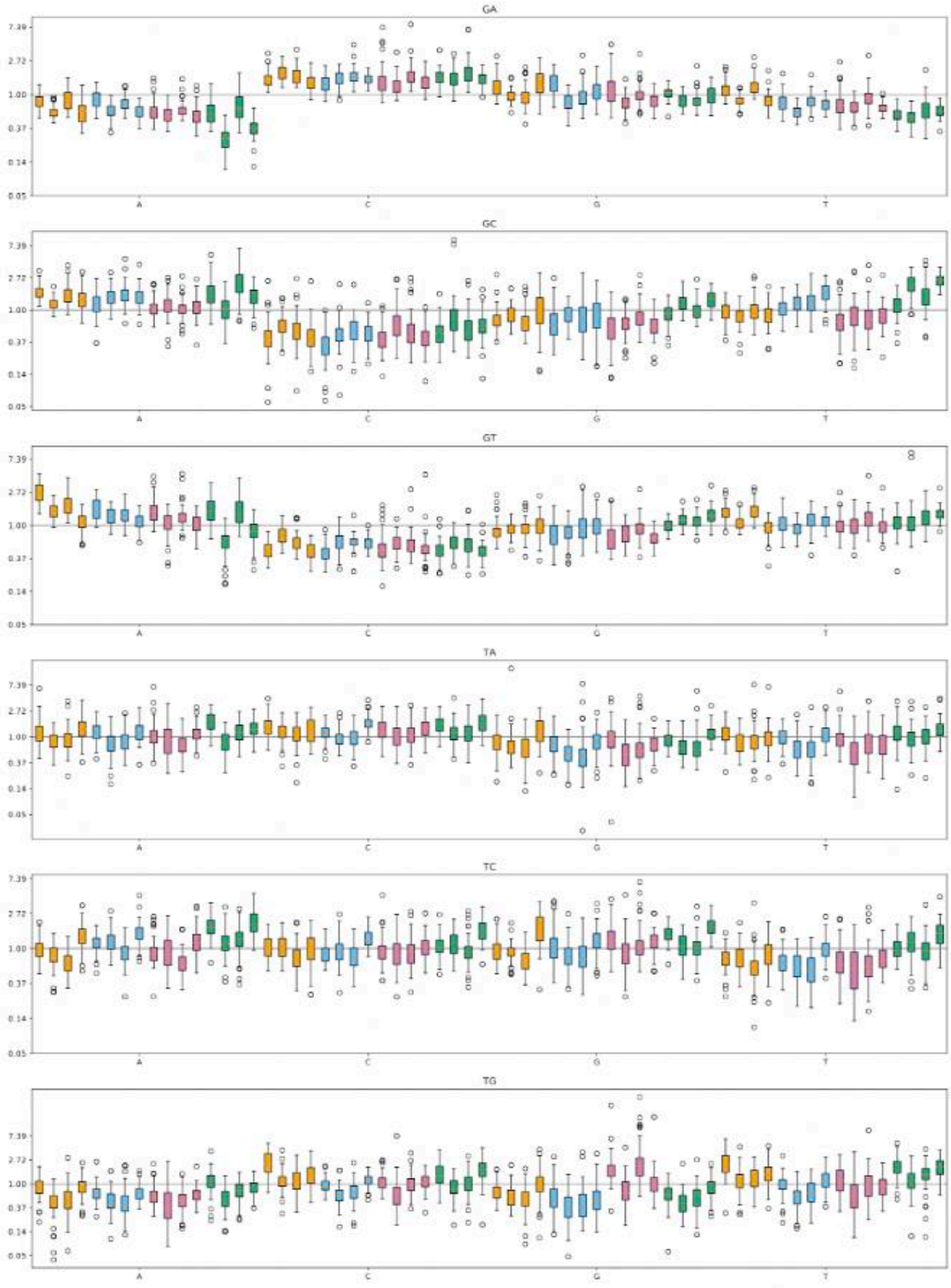

GA
GC
GT
TA
TC
TG
7.39
2.72
1.00
0.37
0.14
0.05
A
C
G
T

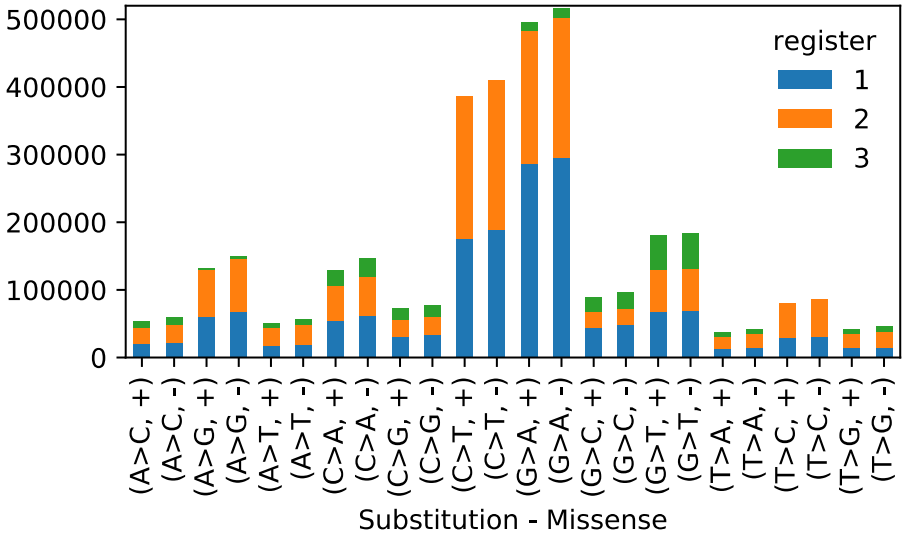
Figure 2.3: Ratio of observed frequency to expected frequency (shown on a log scale) for the each of the 64 triplets that can precede a substitution mutation. Each box plot shows the range of fold enrichments for all the relevant sextuplets. As with figure 2.2 the nucleotide closest to the substitution is shown in the x axis and preceding nucleotide is given by the colour: orange A; blue C; pink G; and green T.

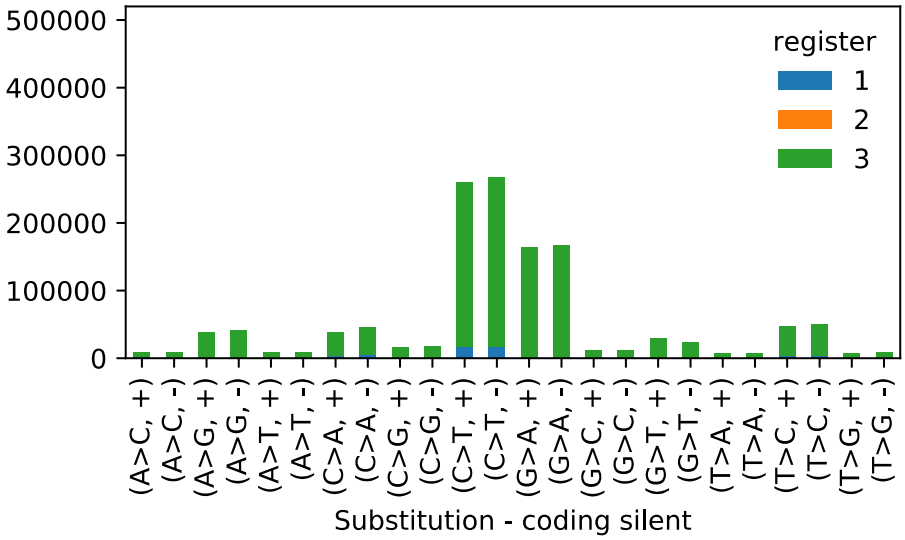
The patterns of fold enrichments observed in the DNA sequence patterns for silent mutations in figure 2.2b) can be largely explained by the redundancy of the amino acid code in that there are only a limited number of substitution mutations that can result in a silent mutation. For example, there are fifteen amino acids (alanine, valine, serine, asparagine, threonine, isoleucine, arginine, histidine, proline, leucine, cysteine, tyrosine, phenylalanine, glycine and aspartic acid) whose codons end in either a cytosine or a thymine thus allowing the possibility of a silent substitution of C>T in position 3 of a codon.

Similarly, the mutation fold enrichments observed in the sequence patterns observed in nonsense mutations are also driven primarily by the makeup of stop codons (See figure 2.2c). In particular, at the site of nonsense substitutions a disproportionate number of the preceding sextuplets end with either AA or AG. Nearly half of all nonsense substitutions (43%) involve a C>T substitution resulting in the beginning of TAA, TAG and TGA (created roughly in the ratio 3:4:5).

Substitution mutations were most commonly reported in the first nucleotide of a codon (37%, 32%, 31% registers 1, 2, and 3 respectively). However, the impact of each substitution depends strongly on the register in which it occurs as shown in figure 2.4. C>T changes in

the third register are likely to lead to silent mutations whilst stop codons predominantly arise from C>T changes in the first register.


500000
400000
300000
200000
100000
0
register
1
2
3
(A>C, +)
(A>C, -)
(A>G, +)
(A>G, -)
(A>T, +)
(A>T, -)
(C>A, +)
(C>A, -)
(C>G, +)
(C>G, -)
(C>T, +)
(C>T, -)
(G>A, +)
(G>A, -)
(G>C, +)
(G>C, -)
(G>T, +)
(G>T, -)
(T>A, +)
(T>A, -)
(T>C, +)
(T>C, -)
(T>G, +)
(T>G, -)
Substitution - Missense


500000
400000
300000
200000
100000
0
register
1
2
3
(A>C, +)
(A>C, -)
(A>G, +)
(A>G, -)
(A>T, +)
(A>T, -)
(C>A, +)
(C>A, -)
(C>G, +)
(C>G, -)
(C>T, +)
(C>T, -)
(G>A, +)
(G>A, -)
(G>C, +)
(G>C, -)
(G>T, +)
(G>T, -)
(T>A, +)
(T>A, -)
(T>C, +)
(T>C, -)
(T>G, +)
(T>G, -)
Substitution - coding silent

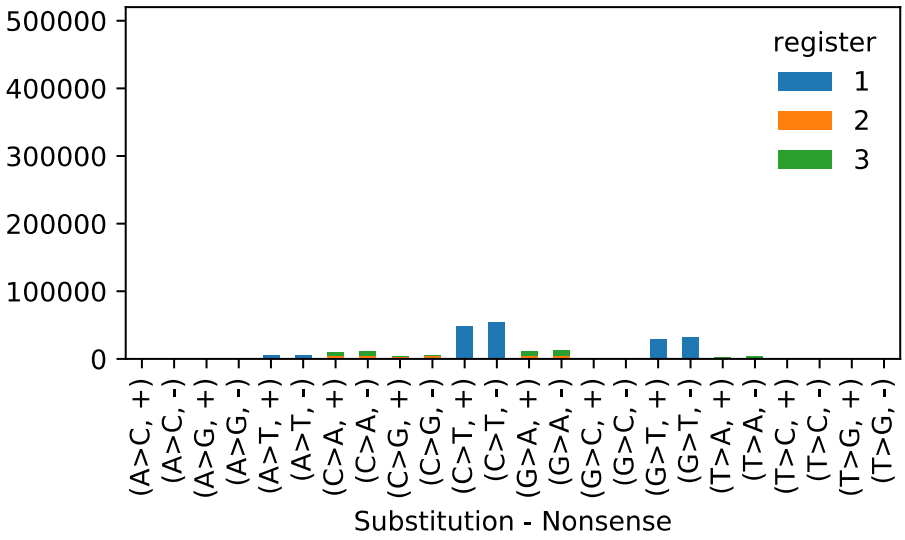

500000
400000
300000
200000
100000
0
register
1
2
3
(A>C, +)
(A>C, -)
(A>G, +)
(A>G, -)
(A>T, +)
(A>T, -)
(C>A, +)
(C>A, -)
(C>G, +)
(C>G, -)
(C>T, +)
(C>T, -)
(G>A, +)
(G>A, -)
(G>C, +)
(G>C, -)
(G>T, +)
(G>T, -)
(T>A, +)
(T>A, -)
(T>C, +)
(T>C, -)
(T>G, +)
(T>G, -)
Substitution - Nonsense

Figure 2.4: Analysis of the register for each substitution mutation showing whether the mutation results in a missense mutation (top), silent mutation (middle), or nonsense mutation (bottom).

2.4.4 Indels

Although the most frequently observed cancer mutation is the single nucleotide substitution, approximately half of all cancer samples contain between 1 and 35 indels. A further 2% of samples contain between 36 and 1500 indels. In total, there were 50,000 insertion mutations and 121,000 deletion mutations in our data set where the length of the indel was between 1 and 60 bases. 100,000 indels were of a single nucleotide 24,000 indels were longer frameshift mutations and 27,000 indels were in-frame. The frequency distribution of indels of different lengths is shown in figure 2.5 below. Subfigures 2.5a) and 2.5b) show the frequencies of deletions and insertions respectively in the protein coding region. Subfigure 2.5c) and 2.5d) shows the length distribution of nucleotide motif repeats within the exome.

As before, I calculated the sextuplet fold enrichment that occur before frameshift insertions and deletions (figures 2.2d and 2.2e), as well as in frameshift insertions and deletions. Approximately two thirds of frameshift indels are singleton indels and there are four peaks in the fold enrichments, at the sextuplets ‘AAAAAA’, ‘CCCCCC’, ‘GGGGGG’ and ‘TTTTTT’ that are predominantly caused by the mutations. Frameshift indels were both found on average 120 times more likely to occur next to a mononucleotide run than would be expected if the sextuplets at the site of indels had the same frequency distribution as the rest of the genome.

Fold enrichment peaks near mononucleotide runs are not apparent for the inframe indels. However, as with singleton indels, inframe indels are also more likely to occur at sites where an indel motif is repeated. Under a random distribution of fold enrichments, the expected number of repeated triplets within the top 64 fold-enrichments would be 1. However, for in-frame deletions, 17 of the top 20 fold-enrichments of sextuplets are repeated triplets, and for in-frame insertions 6 of the top 20 fold-enrichments of sextuplets are repeated triplets.

For indels of lengths greater than one, the frequency of indels of different lengths in the COSMIC database can be modelled as a power law. This is true for insertions and deletions both in the coding region and the non-protein-coding region. Specifically, using a linear regression on the logged insertion frequencies (length 5-30) in the coding region gives: $freq(length) \sim 2206 \times 0.87^{length}$ where the multiplicative standard error of the base is 1.027. The observed frequency of all of the inframe insertions is higher than predicted by this model. using a linear regression on the logged deletion frequencies (length 5-30) in the coding region gives $freq(length) \sim 3668 \times 0.89^{length}$ where the multiplicative standard error of the base is 1.006. The observed frequency of all of the inframe deletions is also higher than this model predicts. When the length distribution of indels in the non-coding area was found (see subfigures 2.5 c) and 2.5 d)) no frequency preference for inframe indels is seen.

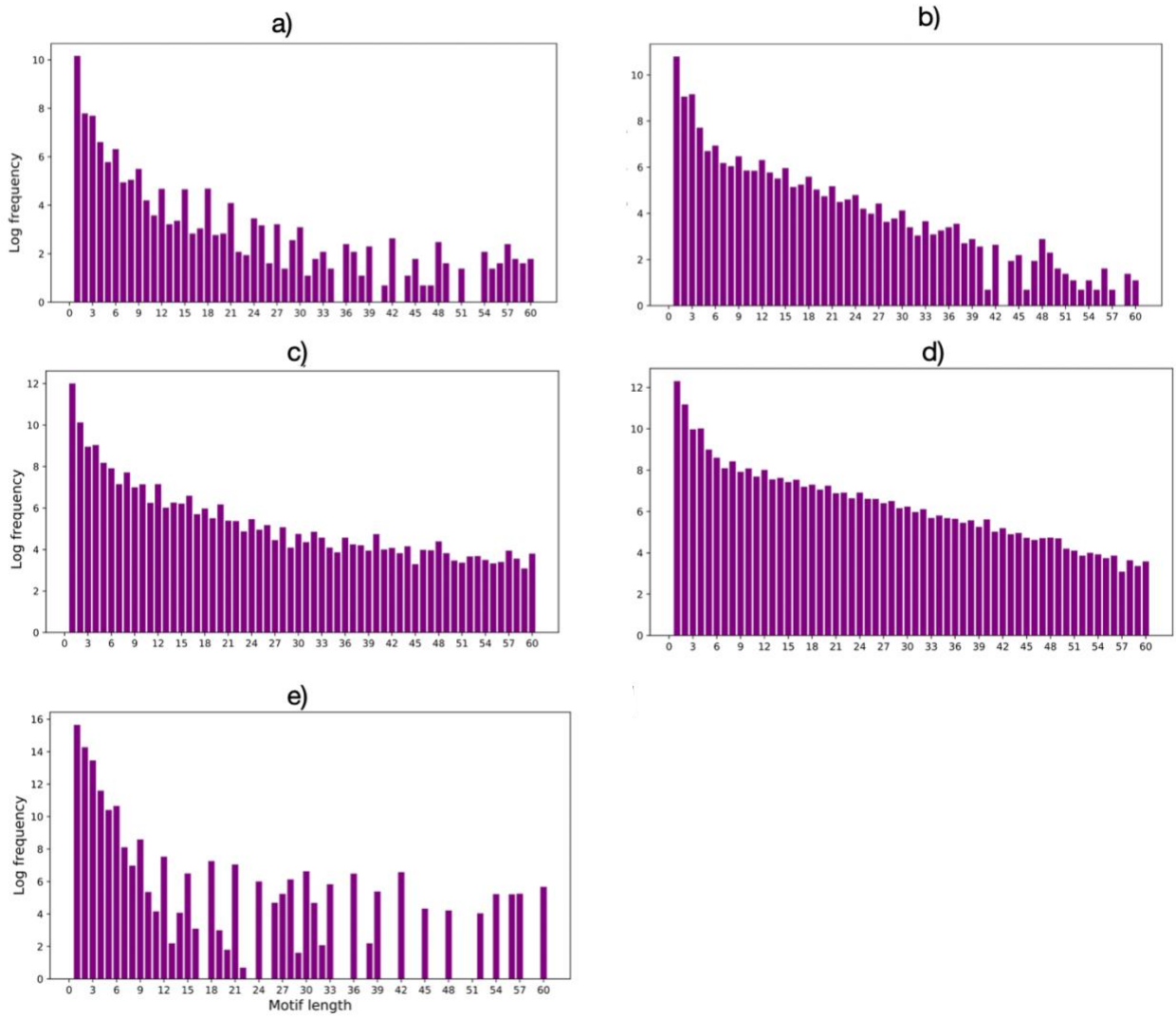


Figure 2.5: Frequency of indels of different lengths shown on a natural log scale. Subfigures 2.5 a) and 2.5 b) show the number of deletions and insertions of different lengths in the coding region. Subfigures 2.5 c) and 2.5 d) show the number of insertions of different length in the non-protein-coding region. Subfigure 2.5 e) shows the number of repeated motifs of different lengths found within the exome.

2.4.4.1 Reliance of insertions and deletions on 'homology hooks'

I have thus far shown that there is an excess of inframe indels in the coding region, and that both inframe indels and singleton indels are over-represented at sites where the indel motif is repeated. I next calculated the percentage of indels that were next to exact or similar motifs for indels with less than 20 bases. To do this I used the Needleman Wunsch algorithm. For small indels I used exact matches only but for indels of six base pairs and longer, I included in this analysis the percentage of indels that have a close but not perfect match with the sequence on one side. The probability that by chance an indel will share a given sequence identity with the sequence to either side depends on the length of the indel. For motifs longer than 5 base pairs I calculated that the probability of a relative Needleman Wunsch alignment score >= 80% sequence identity is less than 5% (and falls rapidly as the length increases).

80% of singleton insertions and between 30-60% of slightly longer insertions (2-20 bp) occur adjacent to a site where the indel motif is already present (see figure 2.6 a)). These figures increase substantially when close approximations to the indel motif are included. In total, approximately 60% of mid-length (4-20 bp) insertions occur at sites that have the indel motif or a close copy already. Deletions shows a different pattern. For very short deletions (1-3 bp) between 50-80% occur at the site where the deletion motif is repeated (see figure 2. 6 b)). However, for longer deletions (4-20 bp) this pattern changes. Only 25% of in-frame deletions occur at sites of motif repeat or near repeat, with much lower percentages for frameshift deletions. In total, around 80% of these longer (4-20 bp) deletions occur at sites which do not have the repeated or near-repeated motif. However, for almost all lengths (1-

20 bp) far more deletions occur at the site of repeated or near repeated motifs than would be expected by chance given the distribution of repeats in the exome.

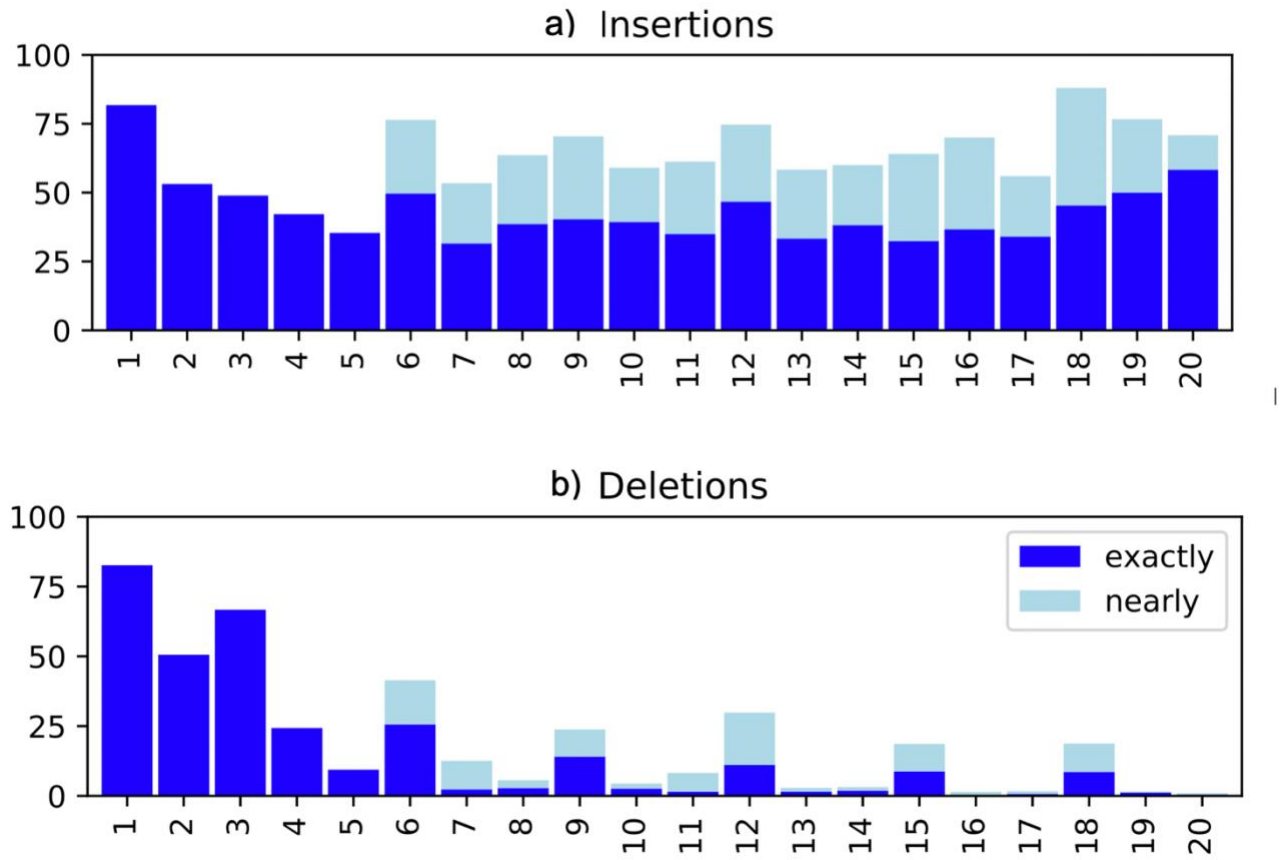


Figure 2.6: Bar charts showing the percentage of indels of length 1-20 base pairs which are either an exact repeat or a near repeat of at least one of the neighbouring nucleotide motifs. A Needleman-Wunsch score of 0.8 was used as a cut-off point for judging 'nearness' as less than 5% of random sequences >5bp exhibit this degree of nearness.

2.4.4.2 Mechanisms of acquisition of insertions and deletions

Indels often occur due to the infidelity of mechanisms for repairing double strand breaks (DSBs) [155] . There are several mechanisms for repairing double strand breaks: non-homologous end joining (NHEJ), homologous recombination (HR), Micro-Homology Mediated End Joining (MMEJ), and Single Strand Annealing (SSA) [155]. Of these HR is the most faithful. NHEJ often results in small indels – typically less than 5 nucleotides long and relies on micro-homologies, whilst MMEJ relies on micro-homologies and results in longer deletions (and occasionally insertions) typically of 5-25 nucleotides with some inserted nucleotides [155]. Alternatively, a failure to remove intermediates during MMR can result in replication slippage at the site of mononucleotide repeats. The distinct difference in homology frequency between insertions and deletions suggests that the mix of DDR mechanisms causing the two types of indels is quite different and needs further exploration.

2.4.4.3 Sequence repeats in the exome and the risk of insertion and deletion.

One distinction between replication slippage and DDR mechanisms that require micro-homology hooks, is that replication slippage is commonly found in micro-satellite sequences where a di- or tri-nucleotide is repeated not once but several times. To identify whether replication slippage contributes to the frequency distribution of slightly longer indels, I calculated the frequency of indels by both indel length and motif repeat, normalising against the frequency of the motif repeats in the background. Figure 2.7 and figure 2.8 show log indel frequency heatmaps in the coding region and non-protein-coding region.

For indels in both the coding region and non-protein-coding region, the dominance of singleton indels after mononucleotide runs of 5-10 can clearly be seen. However, for longer

indels the heatmaps representing the exome and those representing non-protein-coding regions are different. In the exome, there are excess indels at lengths 3bp, 6bp and 9bp, both for indels and these are reflected to some extent by motif repeats in the exome of the same length. In the non-protein-coding region this pattern is replaced by excess indels of lengths 4bp, 6bp, 8bp. This suggests that replication slippage is a relevant mechanism for creating indels longer than the di and tri-nucleotides, but particularly for inframe indels.

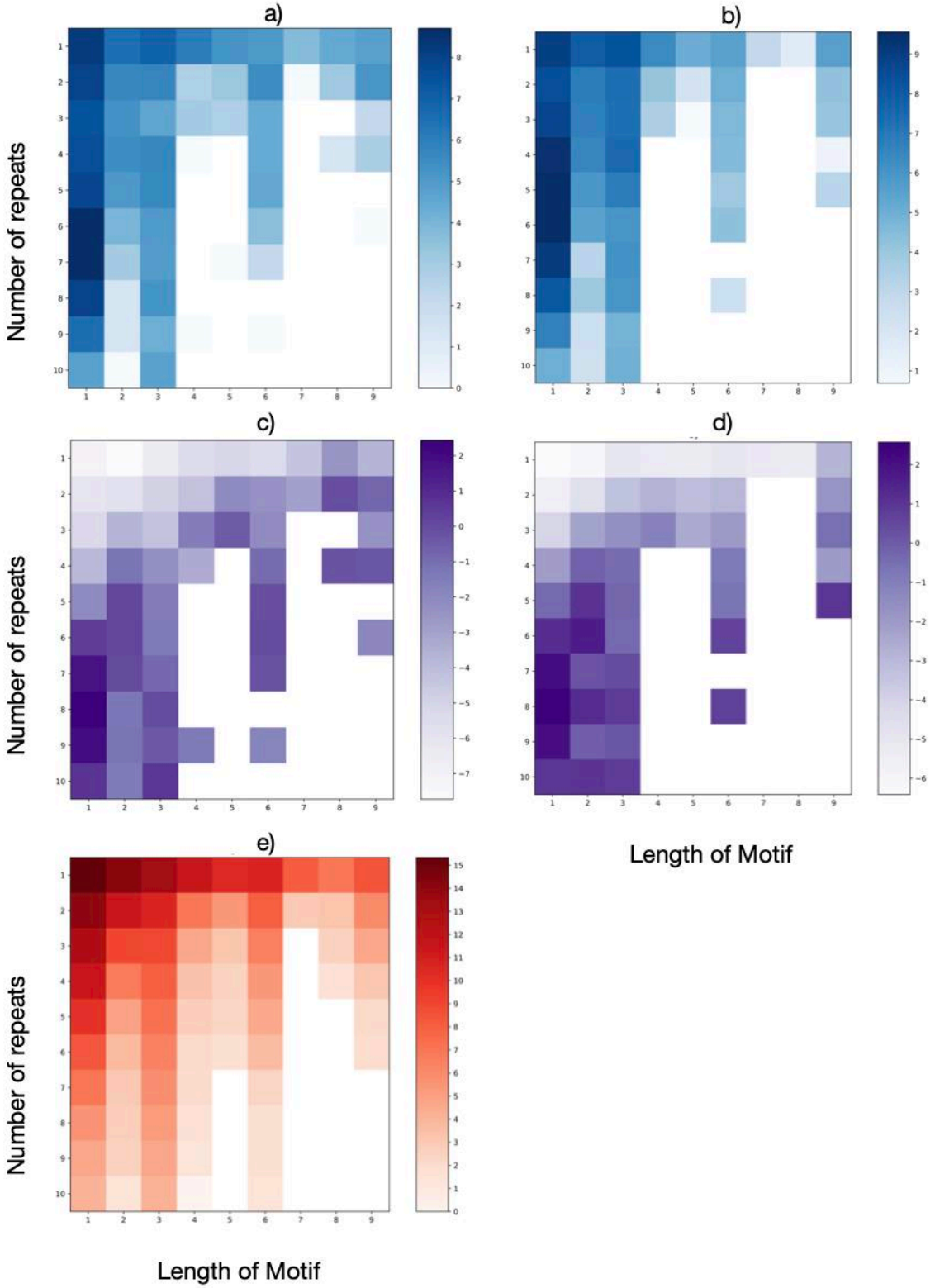

a)
b)
c)
d)
e)
Number of repeats
Number of repeats
Number of repeats
Length of Motif
Length of Motif

Figure 2.7: The heatmaps show the log frequency of indels in the exome as a function of both the indel length (the x axis) and the number of times the motif is repeated at the site of the indel (the y axis). Values with no data are shown white. Subplots a and b show the frequencies non-normalised for insertions and deletions whilst c and d show the same data normalised against the log frequencies of motif repeats of different lengths in the exome. The heatmaps show clearly both the preference for singleton and inframe indels (note in particular the comparatively dark stripes, indicating fold enrichment at 1,3,6,9 and significant lack of data at 4,5,7,8).

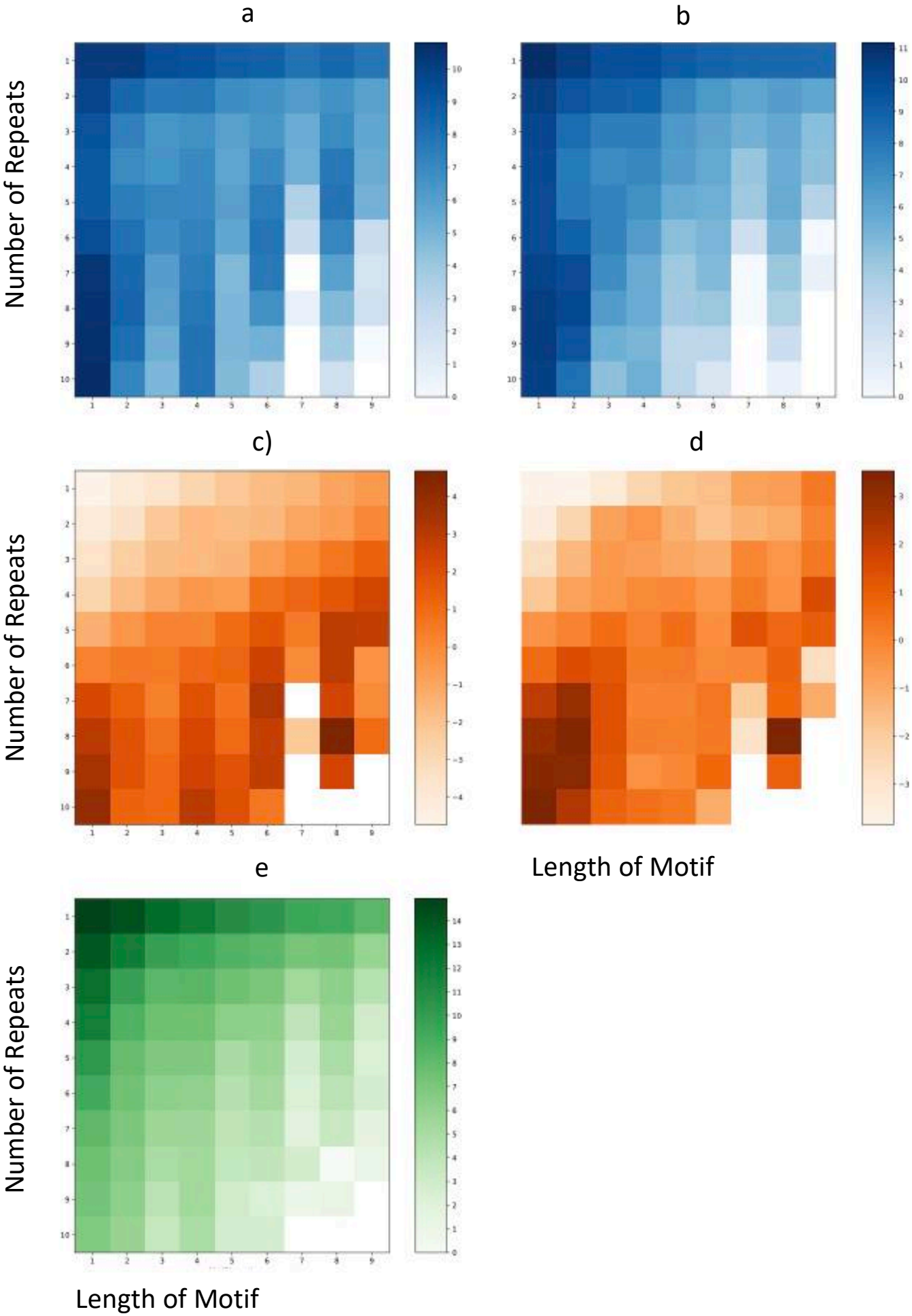

a
b
c)
d
e
Number of Repeats
Number of Repeats
Number of Repeats
Length of Motif
Length of Motif

Figure 2.8: The heatmaps show the log frequency of indels in the non-protein-coding region as a function of both the indel length (the x axis) and the number of times the motif is repeated at the site of the indel (the y axis). Subplots a and b show the frequencies non-normalised for insertions and deletions respectively, whilst c and d show the same data normalised against the log frequencies of motif repeats of different lengths in the exome (shown in e). The heatmaps show excess of singleton indels and fold enrichment for indels of lengths 4,6,8.

2.4.4.4 Impact of frameshift mutations

Nonsense-mediated mRNA decay (NMD) selectively degrades mRNAs that have premature termination codons. However, frameshift mutations which occur in the last exon of a protein may escape NMD [177]. By looking at alternative transcripts for all of the genes with indels I found that for both insertions and deletions 24% of the indels are found in at the last exon of at least one of the potential transcripts and so could escape nonsense mediated decay. Not all of these transcripts are protein coding, so these data suggest that more than 76% of frameshift mutations cause loss of protein product in the cell. Less than 24% may lead to alternative transcripts and protein products.

2.4.4.5 Enrichment of in-frame indels not next to motif repeats

I have shown that there is an excess of inframe indels in the exome. However, as frameshift mutations are normally more damaging that inframe indels I postulated that the excess of inframe indels might be caused in part by a reduced number of frameshift indels where the more damaging indel had led to cell death.

As I have shown that comparatively more inframe indels (particularly inframe deletions) occur at the site of a homology hook or repeated motif, and given that repeated codons are more common in the exome than expected, it is necessary to remove this source of bias. Thus, in order to identify whether or not negative selection has any role to play in the number of overall indels, I look now just at those indels that are not next to a motif repeat.

Let D/T be the number of dinucleotides divided by number of trinucleotides. If there is no negative selection acting on frameshift indels then I would expect the ratio of dinucleotide indels to trinucleotide indels to be roughly the same in the exome as in the non-protein-coding region. That is, $D/T_{exome} \sim D/T_{non\ coding\ region}$. However, if negative selection is apparent, I would expect that the impact of selection in the exome would outweigh that in the non-protein-coding region. That is $D/T_{exome} < D/T_{non\ coding\ region}$.

By looking just at indels not next to a motif repeat I have taken out the bias due to replication slippage and forms of repair that rely on a homology hook. See table 2.2 for numbers of indels. For insertions I find $D/T_{exome} = 1.07$ while $D/T_{non\ coding\ region} = 2.20$, chi statistic = 334, pvalue = 3 e-71 chi-squared contingency test For deletions I find $D/T_{exome} = 1.46$ while $D/T_{non\ coding\ region} = 2.10$, chi statistic = 236, pvalue = 3 e-53 chi-squared contingency test.

This provides evidence that frameshift indels are subject to negative selection pressure.

Insertions: chi-statistic = 334, p-value = 3 e-71

	di-nucleotides	tri-nucleotides
Coding	1416	1318
Non-coding	27691	12608

Deletions: chi-statistic = 236, p-value = 3 e-53

	di-nucleotides	tri-nucleotides
Coding	5202	3566
Non-coding	33967	16189

Table 2.2: Numbers of di and tri nucleotide indels in the exome and non-coding regions. This includes just those indels that are not next to repeats of the indel motif to remove bias due to replication slippage and DDR mechanisms that rely on homology hooks.

2.4.4.6 Frequency of Insertions compared to Deletions

Analysis by tissue type has not been attempted as the comparatively low number of indels reduces statistical significance for individual tissue types. However, in all primary sites, tumours typically had between one and four times as many deletions as insertions within protein coding regions (mean = 2.7). Almost all tissue types had a broadly similar frequency distribution of indel lengths: that is most indels were predominantly of a single nucleotide, and the frequency dropped off rapidly with indel length.

In a small number of cancers, more insertions than deletions are observed and I identified 96 samples in COSMIC that had both more than 10 deletions, and more insertions than deletions. The majority had a frequency distribution of indel length, similar to the other cancers and analysis of their mutational signatures[101][166] suggested that these samples were enriched in mutations characteristic of ageing and/or defective mismatch repair.

In 22 of these samples, all taken from a single study of head and neck cell-lines [178], the pattern of observed indels differed dramatically from the norm. These samples also showed a remarkably similar distribution of mutations to each other both in the number and type of substitutions and the presence of a large numbers of both insertions and deletions greater than 3 base pairs (bp) long. Further analysis showed that these samples were all enriched in mutational signature 3, which is characteristic of cancers defective in homologous recombination (HR) [179]. 18 samples also contained a pathogenic mutation in MSH3 suggesting they were probably also deficient in mismatch repair (MMR). Our analysis agrees with previous studies showing that defective MMR is associated with the presence of more indels in general [180]. The existence of a small number of unusual head and neck samples suggests that damage to both MMR and HR may be linked in some way to the preponderance of longer insertions [181].

2.5 Discussion

In this chapter I have analysed the local sequence patterns occurring next to cancer mutations to try and evaluate how the local sequence patterns influence mutational risk. I found that for substitution mutations there was little influence on the sequence pattern of the proceeding sextuplet except in a few cases. For example, an increase of A>T

substitutions were observed after the nucleotide pattern GATTTC and an increase in T>G substitutions observed after the nucleotide pattern GGCGGG. These sequence patterns have not previously been reported in the literature and it is unclear why they would result in an increased mutation rate. I speculate that they could potentially be binding sites for specific regulatory or DNA damage repair proteins, or that they are preferential binding sites for a particular cytotoxic molecule. Substitutions in cytosine are also marginally more likely after a run of thymines. Again, a reason for this is not known.

There is a stronger association between nucleotide neighbourhoods and the likelihood of an indel. The most frequently observed indel is the addition or deletion of a single base and these generally occur in the presence of mono-nucleotide runs in particular of 5-7bp in length in the genomic sequence. These indels are thought to be formed by replication slippage. I also observed that the risk of other small indels was increased when there were duplicates of that sequence pattern in the underlying genomic sequence. I found this in both protein coding and non-coding regions, however in non-protein coding regions these duplicate patterns tended to be multiples of 2, whereas in protein coding regions, these repeating patterns tended to be duplicates of 3. I postulate that these insertions may also arise due to replication slippage and the different frequencies may be influenced by the underlying genomic sequence patterns; there are far more duplicates of di-nucleotides in non-protein-coding regions and more multiples of 3s in coding regions.

I investigated whether there was any evidence of selective pressure on the presence of indels in cancer samples. I analysed the frequencies of di-nucleotide indels (frameshift) and tri-nucleotide indels (in-frame) not adjacent to a repeat of the indel motif. The results

suggested that protein coding regions were less tolerant of frameshift mutations than inframe mutations.

Finally, I found that most cancer samples contain more deletion than insertion mutations. The few samples with large numbers of insertions (more than 100 indels per sample), and more insertions than deletions come predominantly from head and neck cancers that have both mutated MSH3, impacting formation of the MutSβ and a clear mutational signature 3 suggesting damage to the homologous recombination (HR) pathway as well as mismatch repair (MMR). The heterodimer MutSβ comprises MSH2 and MSH3 recognises larger loop mis-pairs and is good at mending longer deletions [182][183]. It has been reported that mutations in MSH3 normally lead to deletions in repetitive DNA tracts due to damage of the functions on MutSβ.

3 Using mutational signatures in cancer to explore tissue specificity of driver genes

3.1 Abstract

Cancers in different tissue types show remarkable heterogeneity, both in the profile of mutations they exhibit and in the driver genes mutated. In this chapter I examine the links between the two.

Using mutation data from 6,430 samples in The Cancer Genome Atlas, I assess whether mutations in major cancer-associated genes are associated with different mutational signatures. Initially I stratify the samples according to their mutational status for each of the

major cancer-associated genes, and calculate the mutational signature breakdown for each group, identifying those genes associated with statistically significant changes. Mutations in fourteen genes are associated with a change in mutational signature breakdown in at least one tissue type. In five genes (PIK3CA, CHEK2, IDH1, TP53 and BRAF) mutations at a specific site are associated with a change in mutational signature breakdown.

For each cancer-associated gene I then model the comparative frequency of mutated samples in each tissue type by aligning the mutational fingerprint of the samples with an ideal fingerprint for causing mutations in that gene. I found that around two thirds of the cancer-associated genes appear to be partly opportunistic, mutating more frequently in those cancers where the mutational profiles align well with the mutations able to cause pathogenic mutations.

In the twenty most common cancer missense hotspots, the model can be used to predict a frequency distribution of amino acid substitutions. By identifying mutations that occur more or less frequently that expected it is possible to identify amino acid residues under selective pressure. Interestingly in TP53 there are sites where the nonsense mutations observed are less than expected. It is likely that pathogenic mutations at these sites lead to modification of function rather than loss of function, fitting with the view that TP53 has a dual nature with some oncogenic aspects[99] .

3.2 Introduction

Over the last ten years great strides have been made in our understanding of the molecular basis of cancer [39][184][185]. Vast amounts of multi-omic genetic data have been, and

continue to be, generated and deposited in public repositories (e.g. The Cancer Genome Atlas [186]) allowing multiple groups to interrogate these data (for example see [187][188]). Analyses of these data is making it much clearer which types of genetic and epigenetic damage commonly drive tumorigenesis [88]. They have also improved our understanding of the classes of drugs that provide effective chemotherapies [189].

Emerging cancer therapies often rely on inhibiting onco-proteins to kill cancer cells while leaving surrounding cells undamaged. So, improving the understanding of genetic biomarkers that can be used to stratify patients in particular cancers, and reasons for varying numbers of mutations in different tissue types is an important tool in drug development [190].

Almost all cancers have genetic mutations, but the number varies from disease to disease and, as the disease progresses [78], more aggressive sub-clones may emerge with different characteristic mutations [191]. For example, Acute Myeloid Leukaemia has a mean of just 0.28 mutations per megabase, whilst at the other end of the spectrum Lung Squamous Cell Carcinoma has on average 8.15 mutations per megabase [84]. Most of the genes mutated are so called ‘passenger genes. Mutations in these genes tend to occur at low levels and generally have little impact on the fitness of the cell. However, the Cancer Gene Census [105] has annotated 315 tumour suppressor genes and 315 oncogenes. These genes are positively selected for mutation and facilitate tumorigenesis in some cancers. Again, there is variation. A driver gene in one disease may not drive tumorigenesis in another disease [88].

Some genes can be thought of as generic cancer drivers and are mutated in many different types of cancer. The most familiar of these is TP53 [192]. On the other hand some genes are drivers in a small number of specific cancer types only, for example in VHL somatic mutations are predominantly seen in clear cell renal carcinomas [193]. Some genes have different profiles of mutations in different cancers. For example, in skin cancers BRAF V600E mutations predominate, a form of BRAF in which the monomer is activated. However, in lung cancers alternative oncogenic mutations in BRAF are often present whereby the BRAF continues to dimerize, and the diseases shows resistance to BRAF inhibitors [194].

The likelihood of seeing a mutation in a particular gene in a particular tissue depends on both the probability that the cell will acquire the mutation and the likelihood that the cell with the mutation will go on to be present in any future tumour. The probability that a cell acquires a mutation in a specific gene depends on a number of factors, including the nature of the mutational processes occurring in the cell, the gene length and the position of the gene in the genome; Genes that occur in late replicating regions of heterochromatin are at more risk of mutation than those occurring in euchromatin and regions of earlier replication [195]. Against this model of background risk, mutations that contribute to tumorigenesis are seen at higher frequencies than those that do not [98].

The distribution of mutations and large DNA arrangements that occur in any particular cancer depends strongly on the endogenous and exogenous stresses under which the cells are placed. One of the most promising avenues for making sense of the array of differences between cancers is the work emerging over the last six years on mutational signatures. Statistical clustering of mutational frequencies carried out using COSMIC and other large

cancer mutation data sets [103], [104], [179], [185], [196]–[198] has enabled the identification of specific profiles of patterns in substitutions, indels and large scale rearrangements. These are called mutational signatures. The vast majority of mutations are within passenger genes, so it can be assumed that these mutations convey little or no benefit for cell survival. In some, but not all cases, there is statistical evidence linking these signatures with: age; deficiencies in DNA damage repair and replication; and exposure to various genotoxins. More recent work has involved the systematic exposure of a human induced pluripotent stem cell to a wide range of environmental mutagens as well as radiation under highly controlled conditions, followed by clustering analysis of the resulting mutations[197].

Less research has been done to explain why mutations that occur frequently within one tissue type are almost unknown in another. Here, I discuss the concept of tissue specificity of genetic alterations in cancer and provide general hypotheses to help explain this biological phenomenon. I look at whether the mutational status of cancer-associated genes impacts on the overall mutational signatures found in the cell.

For a gene to act as a driver in a particular tissue it must both be subject to mutations, and also contribute to the overall survival rate of the cell. The majority of cancer-associated driver genes show large variability in the percentage of samples mutated in different tissue types[187]. I posited that this could be purely because the survival advantage varies between tissue types, but alternatively it could be that there is far more opportunity for mutation in some tissue types than in others. In order to explore this second possibility, I looked at how well the mutational fingerprints in different tissue types align with the

fingerprint formed from pathogenic mutations for a particular gene. For those genes where the extent of this alignment correlates well with the percentage of samples mutated, I consider that the variation in mutational frequency may be at least in part opportunistic. Finally, I analysed whether the different frequencies of mutated amino acids at commonly mutated positions reflect the mutational frequencies of the samples with mutations at those positions, or are also driven by selective pressure.

3.3 Methods

3.3.1 Mutational fingerprints

Mutational data for 6,430 whole exome screens for TCGA patients were downloaded from the COSMIC database[147]. The split of samples between the different tissue type studies is shown in Table 3.1 below.

Tissue type	Number of Samples	Percentage of Cohort
Breast	960	14.93
Central nervous system	796	12.38
Cervix	179	2.78
Endometrium	248	3.86
Haematopoietic and lymphoid tissue	188	2.92
Kidney	601	9.35
Large intestine	620	9.64
Liver	188	2.92
Lung	484	7.53

Ovary	474	7.37
Prostate	256	3.98
Skin	341	5.3
Stomach	288	4.48
Thyroid	403	6.27
Urinary tract	405	6.3

Table 3.1: TCGA samples analysed for the different tissue types

Duplicated entries at the same genomic position were removed, retaining the mutation were mapped to the lowest transcript number/variant. In common with previously published work on mutational signatures[199], I characterised each substitution mutation in the TCGA database by a nucleotide quadruplet giving the nucleotides either side of the mutation and the substitution itself, using the genome position to identify the flanking nucleotides [166].

To reduce dimensionality, and again in common with previous work [166] I assumed no transcription bias in mutations. This means that mutations can be read on either DNA strand and I can assume the substitution of either a cytosine or thymine. So, for example TP53, 734G>T is found on the negative strand. The nucleotides at the site on the positive strand are GCC. Thus using C>A, the complement of G>T, my quadruplet is G,C,A,C [103]. I then calculated the 96-dimensional quadruplet distribution for each sample. I refer to these as

mutational fingerprints. This database is available online at https://users.sussex.ac.uk/~skw24/TCGA_mutational_fingerprints.csv .

A list of all the genes identified as either tumour suppressor-associated genes or oncogene-associated genes was downloaded from the Cancer Gene Census [200]. This resulted in a data set of 244 tumour suppressor-associated genes, 243 oncogene-associated genes and 71 genes associated with both tumour suppressor and oncogenic action. Of this gene I analysed those genes that had a pathogenic mutation in at least 4% of samples in at least one tumour. This resulted in consideration of a set of 160 tumour suppressor-associated genes, 129 oncogene-associated genes and 50 genes associated with both tumour suppressor and oncogenic action. The genes are set out in Appendix 2.

3.3.2 Assessing the association between strength of cancer-associated genes and mutational profiles

I considered it possible that part of the reason that driver genes are specific to a few tissue types is that the mutational stresses within those tissues may be better suited to generating pathogenic mutations in the gene in question. To examine this question further I compared the *driver strength* of gene *g* in particular tissue types *t* with a measure of the opportunity for pathogenic mutation of *g* in *t*.

For each driver gene *g* and tissue type *t*, I use the percentage of samples with a pathogenic mutation in *g* as a proxy for the driver gene strength. I used this percentage as a proxy for driver strength and used the term *driver strength(g,t)* to describe it in the text. A cut-off of 4% was used to identify frequently mutated genes. I assume that mutations are pathogenic

if they lead to loss of function for tumour suppressor associated genes or gain of function for oncogene associated genes. For the tumour suppressor-associated genes this included missense mutations that are identified by the FATHMM program [201] in COSMIC as being pathogenic, as well as nonsense mutations and frameshifting indel mutations. For oncogenic genes I included missense mutations that are identified by the FATHMM program in COSMIC as being pathogenic. For genes which have been classified in the Cancer Gene Census as having both tumour suppressor and oncogenic properties the process associated with TSAs was followed.

3.3.2.1 Calculating alignment based on an idealised fingerprint

To approximate mutational opportunity, I calculate an idealised fingerprint and then measure the *alignment(g,t)* between the idealised fingerprint and the observed mutational fingerprint.

The idealised fingerprint is a mutational fingerprint which is optimised for causing the pathogenic mutations seen in the gene. For example, in skin cancer, BRAF mutations occur at L245F, G469A, N581H, L597Q, V600E, K601E in varying proportions. Each of these changes in amino acid is caused by a potentially different nucleotide quadruplet mutation (e.g. L245F is caused by T A>C G which is equivalent to C T>G A). By adding the frequencies of these quadruplets together for every sample in the tissue with a BRAF mutation I can form an idealised mutational fingerprint. The alignment between the idealised and observed fingerprint is compared using the cosine similarity test.

Finally, for each gene *g* I calculated the correlation between *driver strength(g)* and *alignment(g)* across all the tissue types. The number of tissue types is too low to use the

built-in correlation probability so instead I use a permutation test. In order to assess whether the correlations are statistically significant I performed 100,000 permutation tests. For each test I shuffle the *driver strength(g)* whilst keeping the values of the vector *alignment(g)* held and then finding the correlation between them.

3.3.2.2 Assessing the impact of driver mutations on mutational signatures

For each cancer-associated gene I identified the cancer primary sites where the driver gene was pathogenically mutated in more than 4% of samples. For each of these site in turn, I then identified the relevant TCGA samples and their mutational fingerprints as above.

Fingerprints were clustered using non-negative matrix factorisation (NMF). NMF can be used to reduce the dimensionality of a non-negative matrix V into two matrices W and H such that: $V = WH + error$. Here H is the basis for the decomposition and W the basis for the weights. NMF has been extensively used on COSMIC data to form a curated census of mutational signatures, and I use the probabilities for the COSMIC mutational signatures as my basis H. Version 2 is used as this is the version for which most analysis has been done and it is based on predominantly on exomic mutations [166].

The corresponding weights W were identified using python's sklearn.decomposition package, holding H constant. The weights were then were split into two matrices W_{mut} and W_{not_mut} corresponding to samples with a pathogenic mutation in the target gene and those without.

For each matrix I identified the median values and then compared these vectors using a cosine similarity test. In order to assess the significance of the values. I performed a

permutation test 10,000 times, repeatedly shuffling W before splitting it into random matrices of the same size as W_{mut} and W_{not_mut} and then carrying out the cosine similarity test. Finally, results were ranked and evaluated using the Benjamini-Hochberg procedure.

3.3.2.3 Mutational patterns in hotspots

I also investigated signature changes at the site of commonly occurring mutational driver hotspots where more than one amino acid mutation is observed.

TCGA mutations were filtered for mutational sites occurring in more than 100 samples that had available mutational fingerprints. Disease sites were filtered to ensure that each mutation considered occurred in at least 5 samples in each of the diseases considered.

For each tissue type and each mutational position, the frequency with which the wild type amino acid was mutated to different amino acids was counted (e.g. V600E, V600D etc). This gives us the observed frequency for each mutant amino acid. For each hotspot position (e.g. BRAF V600) samples containing a mutation were grouped by the tissue type of the cancer and the mutational fingerprint determined.

To identify the baseline mutation frequencies I assumed that for all the samples with a mutation in gene *g,* tissue type *t,* the mutation to the target amino acid could have occurred equally at each of the three nucleotide registers in the codon, and that the frequency distribution for mutations to A,C,G, or T would be given by the summed mutational fingerprints of samples with the mutation in tissue type *t* and gene *g*. Silent mutations were removed from the vector before normalising, as no selective pressure will be seen for these mutations. This gave us a vector of expected values for each tissue type. The mean vector

for all the tissue types was then found, weighted by the sample size. I used this 'expectation vector' in two ways. Firstly, I identified how well it matched the observations by comparing its cosine similarity with the observation vector against the distribution of random cosine similarities. Secondly, I identified which of the preferential selective pressure for specific amino acid substitutions were statistically significant. I did this by using the expectation vector to generate *S* random mutations 1,000 times where *S* is the sample size. I then compared the frequency of mutant amino acids with that observed.

3.3.2.4 Correlations between mutational probabilities and mutational load

For each tissue type I ordered samples by the overall number of genes which were mutated in any way. I then calculated the average mutational load, smoothing over the nearest 50 samples. Likewise, for every frequently mutated gene the probability that it had a pathogenic mutation was calculated, smoothing over the nearest 50 samples. Python's scipy.stats.pearsonr was used to calculate the correlation: the pvalue was approximated by shuffling the samples 1,000 times and then taking the correlation in the same way. Multiple testing is corrected for using the Benjamini-Hochberg procedure [202].

3.4 Results and Discussion

3.4.1.1 Mutational processes vary between tissue types

Mutational fingerprints capture the mutational processes that have occurred in the development of a cancer. These vary between cancers within different tissue types due to the different exogenous and endogenous pressures within the cell. Considerable work has been done to identify signatures within specific samples (see Chapter 1, section 5.9) [100]–[103].

Although, the individual signatures that occur in each cancer has been described it is unclear which cancers have the most similar mutational finger prints. Here I cluster the median fingerprint for each tissue type to reveal links between the different tissue types. These mean mutational fingerprints are shown in figure 3.1 and explored below.

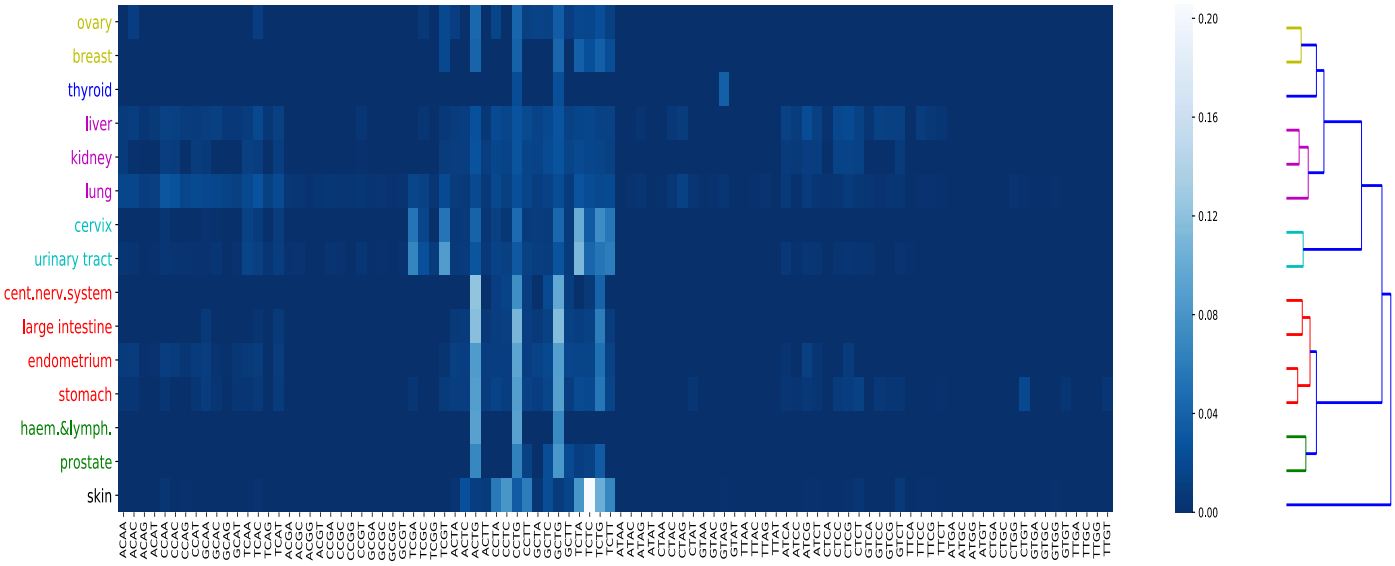


Figure 3.1 - Heatmap showing the median substitution fingerprints for TCGA samples from different cancers. It reveals distinctive patterns of mutations reproduces known COSMIC signatures 1,2, and 7 associated with ageing, AID/APOBEC and UV damage respectively. It also clearly shows the prevalence of GT>AG transversions in thyroid cancer. Clustering demonstrates that whilst all cancers have distinctive the patterns of substitutions are often similar in related tissue types.

The mean mutational fingerprints cluster into seven distinct cancer clusters.

Cancer cluster 1 includes both ovarian and breast cancers. These show slightly elevated signs of NC>GT, where N is any one of A,C,G,T. This pattern is the dominant feature of Cosmic Signature 1, thought to be caused by the deamination of 5-methylcytosine at CpG dinucleotides, and associated with ageing. Ovarian and breast cancers also show slightly elevated signals for TC>TN. Signatures 1, 3, and 5 are found in both ovary and breast cancers.

Cancer cluster 2 comprises thyroid cancers alone. It has distinct fingerprints, which shows slightly elevated signs of the GT>AG mutations that can result in the BRAF V600E mutation, as well as MC>GT where M is either C or G. Signatures 1, 2 , 5 and 13 are found in thyroid cancers.

Cancer cluster 3 includes liver, kidney and lung cancers. These show slightly elevated signs of PC>GT, where P is any one of A,C,G. More generally levels of C>T, T>C and C>A are raised above the background levels. Signatures 1, 5 and 6 are found in liver, kidney and lung cancers.

Cancer cluster 4 includes cervical and urinary tract cancers. These show elevated rates of TC>GN mutations where N is either A or T and in TC>TN mutations where N is any one of A,C,G,T. This pattern of mutation is characteristic of cytosine deamination caused by action of AID/APOBEC enzymes described by Cosmic Signature 2. Elevated levels of mutations of NC>TG are also visible where N is A,C,G or T.

Cancer cluster 5 includes the cancers of the central nervous system, the large intestine, the endometrium and the stomach. These also show the highly elevated levels of mutations of NC>TG associated with Cosmic Signature 1.

Cancer cluster 6 comprises cancers of the haematopoietic and lymphoid tissue and prostate cancers. These also show the highly elevated levels of mutations of NC>TG associated with Cosmic Signature 1. In this they are similar to the mutational fingerprints of cancer cluster 5, but they have slightly less elevated levels of other mutations.

Finally, skin cancers have very distinctive mutational fingerprints, and form cancer cluster 7 showing highly elevated levels of TC>TN and elevated levels of CC>TN. These mutations are characteristic of COSMIC signature 7 which is associated with ultra-violet light exposure that cause C>T substitutions at dipyrimidines sites [203].

3.4.1.2 Most driver genes are recurrently mutated in one or two tissue types

The bioinformatic identification of driver genes depends on identifying not only on frequency of mutations in specific genes, and probable consequence of those specific mutations but also on identifying the expected mutational frequency due to the gene length and timing of replication[204][185]. 558 genes have been described as oncogene, tumour suppressor or both, in the Cancer Gene Census as of December 2019 [88]. Whilst these genes all have some impact on the pathways underpinning carcinogenesis, the extent to which they are mutated in different cancers or drive these cancers is very variable. I refer to these genes as oncogene-associated or tumour-suppressor associated to emphasise that

whilst I am exploring the extent to which they are mutated, I am not assuming that these genes drive cancer in all of the tissue types considered.

I defined a gene as frequently mutated if pathogenic mutations were described in 4% or more samples in a cancer of a particular tissue type (see methods). Of the cancer-associated genes analysed, 105 were frequently mutated in just one primary tissue type with a further 106 frequently mutated in two or more tissue types. TP53 was unique in that it was frequently mutated in almost all (13) of the tissue types. Other driver-associated genes frequently mutated in more than half the tissue types were: the tumour suppressors - LRP1B (9 tissue types), NF1 (9), KMT2C (9), and CSMD3 (10); the oncogenes - PIK3CA (10), and MTOR (8); and genes with both oncogenic and tumour suppressor activity - KMT2D (9). The histogram of frequently mutated genes is shown in figure 3.2.

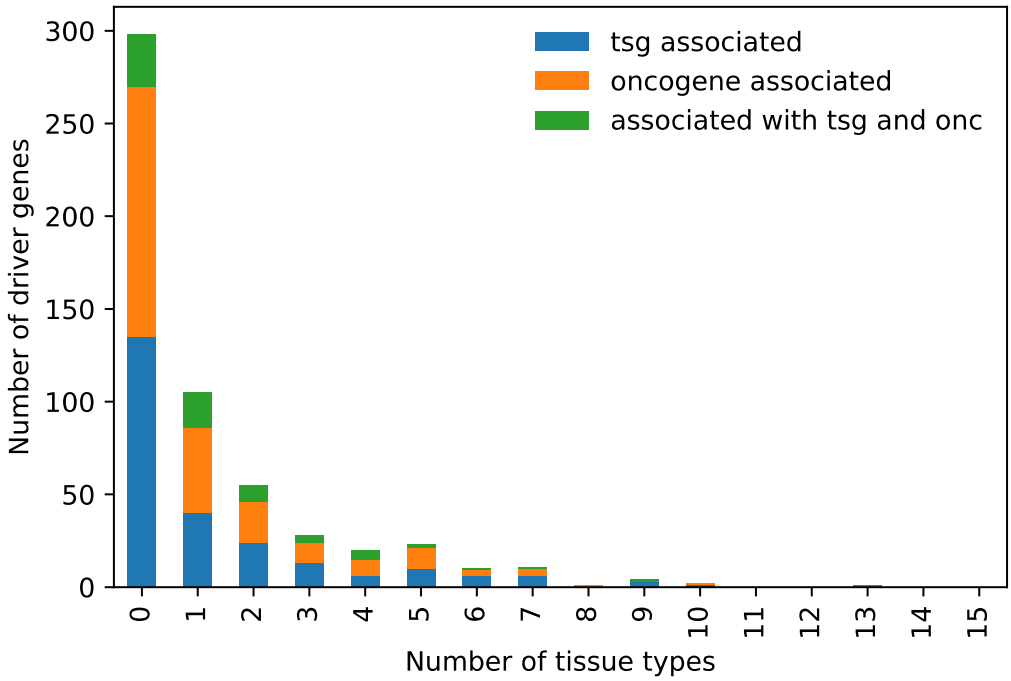


Figure 3.2: Histogram showing the number of driver genes that are frequently mutated in different numbers of tissue types. Most are never frequently mutated, whilst the remainder are generally specific to only one or two tissue types.

Next, I clustered cancers of different tissue types using the ward hierarchical clustering method using the frequency of recurrently mutated cancer-associated genes. See figure 3.3 for the resulting heatmap and cluster. The number of genes recurrently mutated is different in the different cancer types and the cancers broadly clustered into six cancer-associated gene clusters.

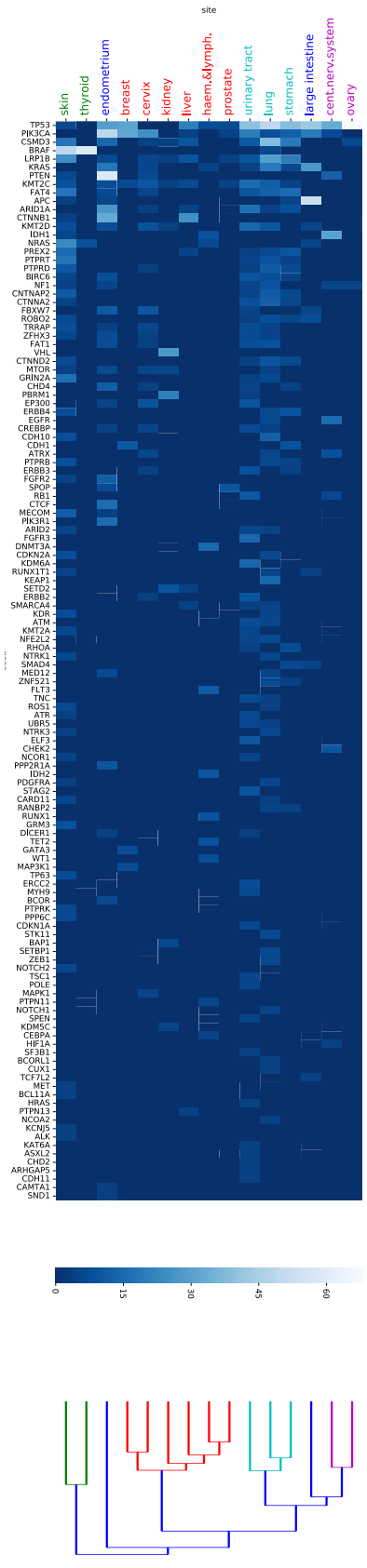


Figure 3.3: cancer sites clustered by proportion of samples with pathogenic mutations in cancer- associated genes.

Gene cluster 1 includes the skin and thyroid cancers. Although these cancer arise through different mutaional processes with very distinct mutational fingerprints both these cancers show markedly elevated levels of mutations in BRAF and elevated levels in NRAS. Overall there are many more mutations in skin cancers than thyroid cancers.

Gene cluster 2 just includes cancers of the endometrium. These are characterised by highly elevated levels of mutations in PIK3CA and PTEN.

Gene cluster 3 included cancers of the breast, cervix, kidney, liver, haematopoietic and lymphoid or prostate show no markedly elevated levels of gene mutation.

Gene cluster 4 included cancers of the urinary tract, lung or stomach and show elevated mutation levels in a wide range of genes: 16 genes mutated by at least 4% in all three tissue types: TP53, PIK3CA, CSMD3, LRP1B, KRAS, KMT2C, FAT4, ARID1A, PREX2, PTPRT, PTPRD, BIRC6, CNTNAP2, CTNNA2, ROBO2, CTNND2. Indeed, TP53 is mutated in at least 39% of samples in each of these cancers.

Gene cluster 5 included cancers of the large intestine and were characterised by elevated mutation levels in APC with 53% of the samples mutated.

Gene cluster 6 included cancers of the central nervous system or ovary show elevated mutation levels in TP53 and NF1. Indeed TP53 mutated by in more than 33% of samples of each cancer.

Clustering tissues by frequently mutated genes gives clusters are not statistically similar (Fowlkes-mallows score see figure 3.4) to those found when clustering them by mutational fingerprints. That is tissues which have similar mutational fingerprints may have very different highly recurrent genes, and tissues with very different mutational fingerprints can result in similar frequently mutated driver genes.

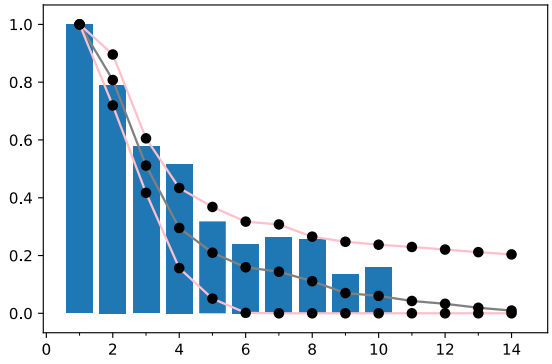


Figure 3.4: Bar chart of the Fowlkes-mallows scores between the tissue type dendrogram based on mean mutational fingerprints and that based on percentage mutations in cancer-associated genes. In order to compare cluster trees, I calculate the square root of the recall and precision for clusters formed with successively lower cut-offs, and compare with those obtained for a random cluster pattern. The grey plot shows the expected values for independent trees together with 2 standard deviation cut-offs (in pink) Statistically similar clusters will have bars higher than the 2sd marks.

3.4.1.2.1 The mutational status of major cancer-associated genes impacts on the signature breakdown for some tissue types

As tumorigenesis progresses the cancer genome becomes increasingly genetically unstable. In some cases, this may be due to failure of the DNA damage repair mechanisms in the cell caused by the failure or overactivation of the driver genes themselves. As mutational fingerprints reflect the mutational processes occurring in a tumour, I investigated the association between pathogenic mutations in the frequently mutated cancer-associated genes and changes to mutational fingerprints.

I analysed the impact of each frequently mutated cancer-associated genes individually. For each gene in each tissue type, I divided the samples into two sets; those with a pathogenic mutation and those with no mutation in the gene. I then decomposed each set into the COSMIC mutational signatures. [15], [20], [21]. The cosine similarity between paired sets was then calculated and the significance calculated using a permutation test. The results below are all significant at the 5% level following corrected for multiple testing using the Benjamini-Hochberg procedure. Statistically significant changes in patterns in mutational signatures were observed for several mutated cancer-associated genes in cancers of the breast, endometrium, stomach, lung, central nervous system, and cervix as shown in table 3.1 and figure 3.5 below.

Tissue Type	Gene	Function	Signatures elevated in mutated samples	Signatures lowered in mutated samples	P -value (permutation test)
Breast	CDH1	Regulation of cell-cell adhesions	2,13	1	e-4
Breast	PIK3CA	Activation of signaling cascades	2,13	1	e-4
Breast	TP53	Cell-cycle regulation and apoptosis	3,10,13	1,2	< e-4
Cervix	ARID1A	Regulation of transcription	6,7	1,2,10,13	2e-3
Endometrium	CAMTA1	Regulation of transcription	2,6,15	1,13	4e-4
Endometrium	KMT2D	Regulation of transcription	6,20	1,13	2e-4
Endometrium	KRAS	Regulation of cell proliferation	6	1,13	1.2e-3
Endometrium	TRRAP	Regulation of transcription and possibly cell-cycle progression	6,20	1,13	6e-4

Stomach	CSMD3	None known in stomach – potentially proliferation.[206][207]	7,13,15,17	1,6	5e-4
Stomach	ERBB4	Cell surface receptor for EGF family.	2,13,17,22	1,6	5e-4
Stomach	LRP1B	Potential extracellular signal transduction.	6,13,15,17	1	<e-4
Stomach	PREX2	Cell protection against oxidative stress.	2,6,7,13,15,17	1	3e-4
Stomach	PTPRD	Signalling protein involved in wide variety of cell processes.	2,13,15,17,22	1,6	7e-4
Central nervous system	IDH1	Metabolic enzyme.	15	1	<e-4
Lung	CSMD3	None known in lung – potentially proliferation.[207]	4,6,22,24	1,2,3,7,13	e-4
Lung	LRP1B	Potential extracellular signal transduction.	4,6,22	1,2,3,7,13,24	6e-4

Table 3.1: Signatures elevated or lowered in samples with mutations in cancer-associated genes.

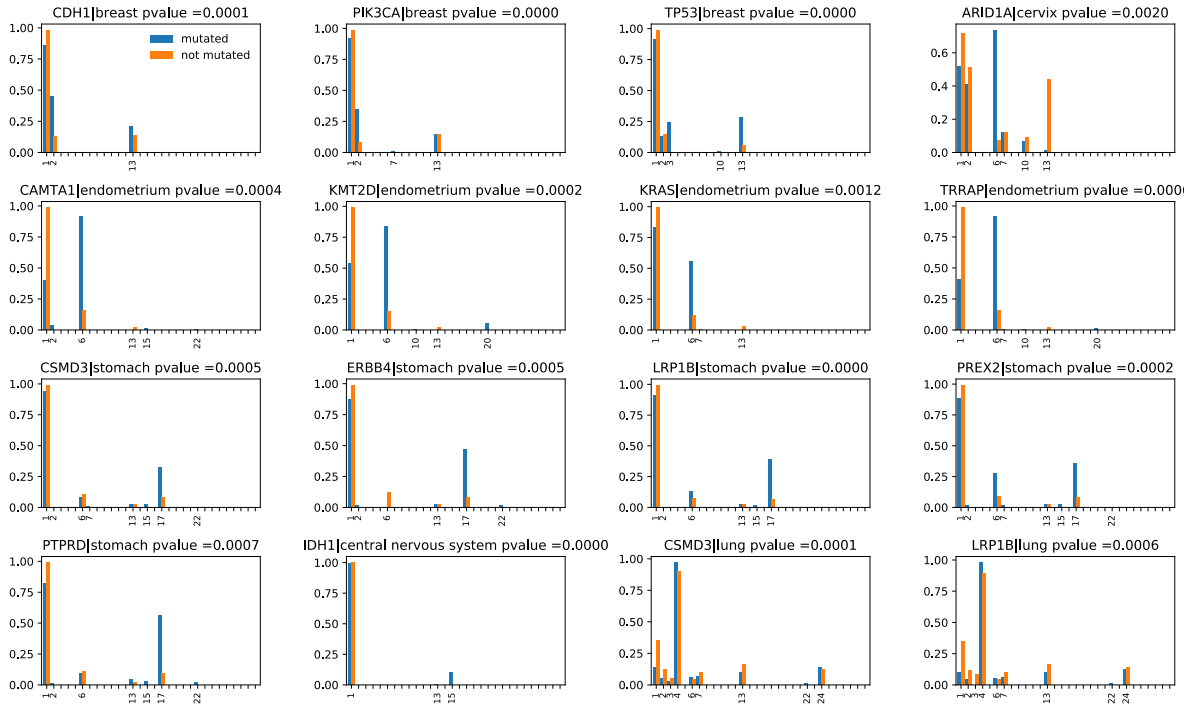


Figure 3.5: Changes in mutational signature seen in samples with and without mutations in cancer associated genes. The signatures are those set out in COSMIC version 2 and have the following associations: 1: spontaneous deamination of 5-methylcytosine – correlated with age, 2: activity of the AID/APOBEC family. 3: failure of homologous recombination. 4: smoking. 6: defective DNA mismatch repair. 7: ultraviolet light exposure. 13: activity of the AID/APOBEC family. 15: defective DNA mismatch repair. 17: aetiology unknown. 22: aristolochic acid. 24: aflatoxin.

In general, the samples that contained mutations in the genes of interest identified above show a loss of contribution of signature 1. Signature 1 is a clock-like mutation: spontaneous deamination of cytosine gradually giving rise to mutations[104]. This suggests that

mutations in the genes of interest are indicative of cancers in younger patients as a result of exogenous stresses.

In breast cancer, mutations in CDH1, PIK3CA and TP53 were associated with a statistically significant change of mutational fingerprints. CDH1 can act as a tumour-suppressor, PIK3CA as an oncogene, and TP53 can acquire mutations to enable it to act as both but is more commonly thought of as a major tumour suppressor [208]. Mutated CDH1 and PIK3CA are associated with an increase in signature 2 whereas mutations in TP53 are associated with a loss in signature 2 and an increase of COSMIC signature 13. COSMIC signature 2 is associated with the activity of the AID/APOBEC family. It is known that p53 is a transcriptional regulator of genes from the APOBEC family[209], and there is evidence that APOBEC activity can lead to mutations in PIK3CA gene in a number of cancers [210]. However, it was not possible to disambiguate any impact of CDH1 and PIK3CA mutation using this data set; all samples with pathogenic PIK3CA mutations also have a CDH1 mutation. Mutations in TP53 were also associated with an increase in COSMIC Signature 3 which is characteristic of patterns observed by the failure of double strand break repairs by homologous recombination. p53 has been shown to play a role in the regulation of homologous recombination [211] and this mutation may cause this effect.

For cancers of the cervix, mutations in ARID1A (which acts as a tumour suppressor in cervical cancer), were associated with the presence of signature 6, characteristic of a defect in mismatch repair.

In endometrial cancer, mutations in CAMTA1A, KMT2D, KRAS, and TRRAP were associated predominantly with an increase in signature 6 and a decrease in signature 13. CAMTA1A can act as a tumour-suppressor. KMT2D can act as both a tumour suppressor and an oncogene, but tends to be a tumour suppressor in endometrial cancers. KRAS and TRRAP are both oncogenes. Signature 6 is associated with defective mismatch repair and signature 13 attributed to activity of the AID/APOBEC family of cytidine deaminases. Mutations in these genes were not found independently of one another. Mutated KRAS was associated with mutated CAMTA1 (p-value < e-26, chi-square association test) with KMT2D (pvalue <e-6, chi-square test) and with TRRAP (pvalue<e-5, chi-square test). It is therefore possible that the signature changes are all caused by activations in KRAS.

In stomach cancer, mutations in CSM3, ERBB4, LRP1B, PREX2 and PTPRD were associated with change in mutational fingerprints. CSM3, LRP1B and PTPRD can act as tumour suppressors, PREX2 can act as an oncogene and ERBB4 can act as both although in stomach cancer it usually acts as an oncogene . Mutations in ERBB4 and PREX2 were both associated with increases in signature 6. Increases in signature 17 were seen in samples with mutations in LRP1B, PTPRD, PREX2, ERBB4. Signature 17 is characterised by NT>GT mutations where N in {C,G,T}. However, the aetiology of signature 17 is unknown. Mutations in the four genes were not independent: mutations of LRP1B were associated with mutations in PTPRD (pvalue =1.3e-6, chi-square test) , PREX2 (4.7e-4, chi-square test) and ERBB4 (pvalue 2.7e-5, chi-square test). This suggests that the changes in signature 17 may be primarily associated with mutations in LRP1B.

In cancers of the central nervous system, mutations in IDH1, an oncogene in glioblastoma, were associated with the presence of signature 15 which is characteristic of defective mismatch repair.

Finally, in lung cancers, mutations in LRP1B and CSMD3 was associated with change in mutational fingerprints. LRPB1 and CSMD3 can both act as oncogenes. The mutational rates of the two genes are associated (p-value 9.9e-6, chi-square test). Signature 4 (associated with smoking) is stronger in samples with mutations in either LRP1B or CSMD3. The picture is complicated because, there is evidence to suggest that LRP1B is more frequently mutated in lung cancer patients who have Chronic Obstructive Pulmonary Disorder. Cigarette smoking is known to be a principal cause of COPD, which in turn increases the risk of lung cancer. However, Xiao et al. found similar signatures in patients with and without COPD. This suggests that any link between LRP1B and smoking signature is more direct [212].

As can be seen from table 3.1 and the discussion above, each of the genes associated with signature change is specific to a single tissue type. The two exceptions to this are CSMD3 and LRP1B. These were both associated with signature changes in both stomach cancer and lung cancers. However, the genes were associated with differing changes in signature in the different tissue types. This specificity points to the varying significance of each of the genes in driving cancers in the different sites.

3.4.1.2.2 The mutational status of major cancer-associated mutations impact on the signature breakdown

The sites of driver mutations in many oncogenes are highly constrained to just a few positions within the protein product that are critical for altering the resulting protein function. I therefore repeated the analysis above looking at site-specific mutations where specific mutations are observed in at least 4% of samples in one or more cancers. There were 23 such mutations in oncogene-associated genes as shown in the heatmap (figure 3.6) below. I have included TP53 mutations within this list as although TP53 is generally thought of as a tumour suppressor there is evidence that some common TP53 mutations are oncogenic[213].

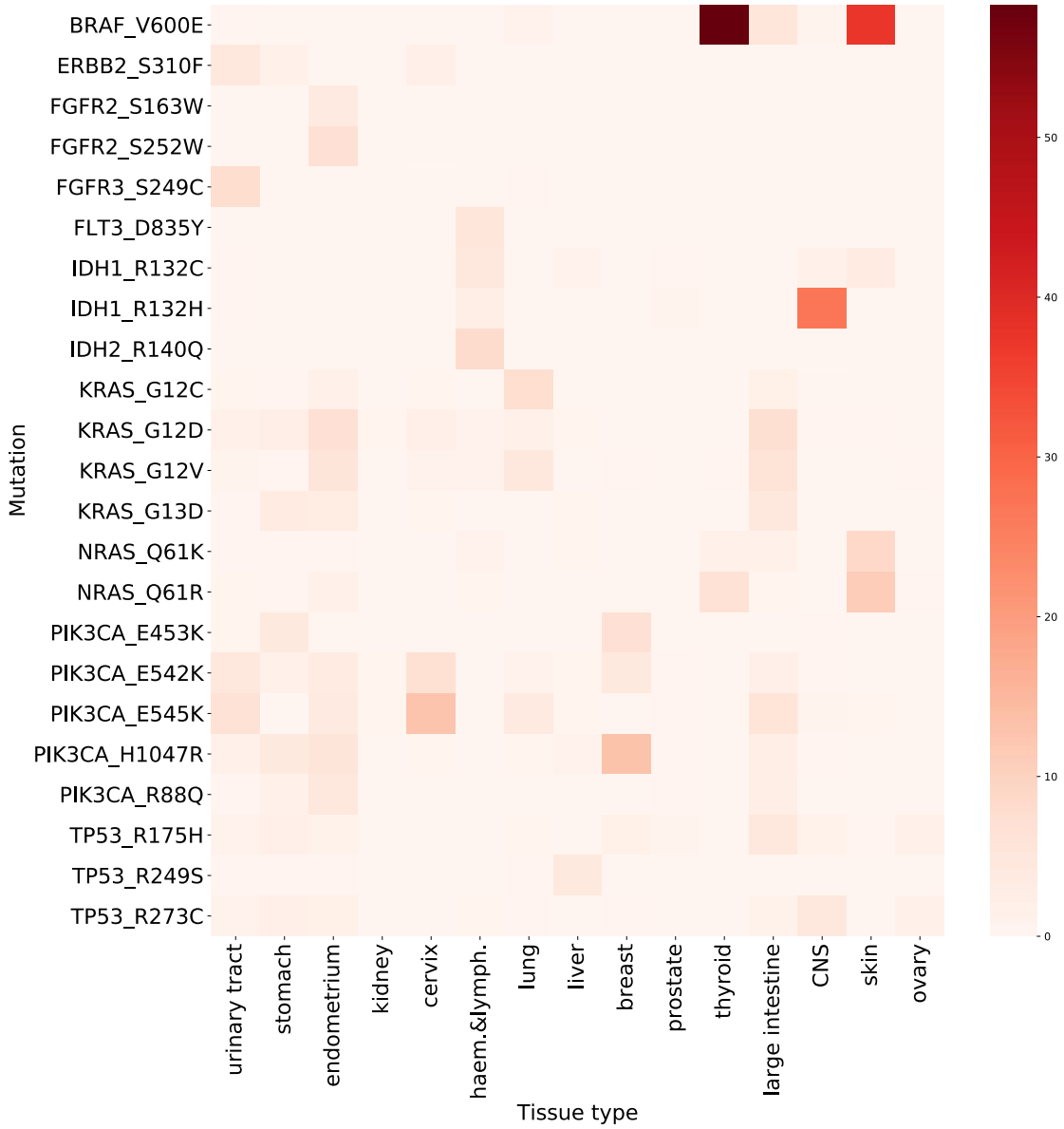


Figure 3.6: Heatmap showing percentage of samples with specific mutations for those mutations with at least 4% samples mutated in at least one tissue type.

Repeating the process of signature comparison using just those specific missense substitutions that are repeated in more than 4% samples in any one cancer, identified eight mutations that are associated with statistically significant differences in signatures. These are shown in figure 3.7 below and summarised in table 3.2.

Tissue Type	Gene	Signatures elevated in mutated samples	Signatures lowered in mutated samples	P-value permutation test
Breast	PIK3CA pE545K	2	1,13	0.0024
Central Nervous System	CHEK2 pY433C	2,21	1	0.0009
Central Nervous System	IDH1 pR132C	13,15	1	0.0018
Central Nervous System	IDH1 pR132H	15	1	<0.0001
Central Nervous System	TP53 pR175H	13,15	1	0.0012
Central Nervous System	TP53 pR273C	13,15	1	<0.0001
Stomach	TP53 pR273H	3,6,7,10,13,17,22	1,15	0.0021
Thyroid	BRAF pV600E	22	1	0.0025

Table 3.2: Signatures elevated or lowered in samples with mutations in oncogenic hotspots.

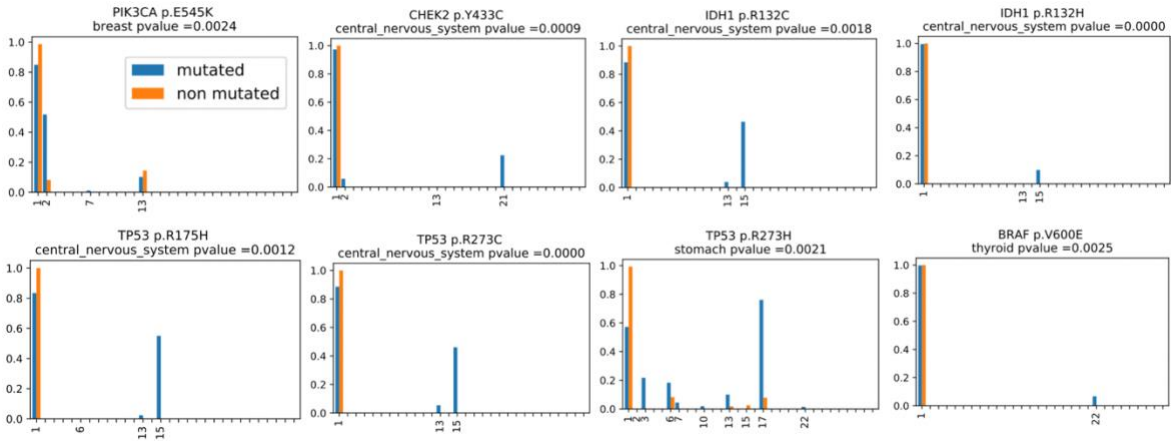


Figure 3.7: Statistically significant changes in mutational signature decomposition for samples with and without common substitutions in oncogenic-associated genes. The signatures are those set out in COSMIC version 2 and have the following associations:

1: spontaneous deamination of 5-methylcytosine – correlated with age

2: activity of the AID/APOBEC family.

3: failure of homologous recombination.

6: defective DNA mismatch repair.

7: ultraviolet light exposure.

10: altered activity of POLE.

13: activity of the AID/APOBEC family.

15: defective DNA mismatch repair.

17: aetiology unknown.

22: aristolochic acid.

In breast cancer mutations in PIK3CA E545K were associated with a reduction in signatures 1 and 13 and an increase in signature 2 which is associated with activity of the APOBEC family.

There is evidence that off-target activity by APOBEC proteins are a cause of PIK3CA mutations [214].

Five of the eight statistically significant results were found in cancers of the central nervous system. TP53 mutations p.R175H and p.273C were associated with similar changes to IDH1 p.R132H and IDH1 p.R132C. Mutations in these genes commonly co-occur (pvalue<e-19, fisher exact test). This means that it was not possible to identify whether any changes in signatures seen are as a result of the TP53 mutations or the IDH1 mutations. There may of course be an unseen factor influencing both mutations. It is worth noting that mutations in both genes change the extent of DNA damage repair in gliomas [215].

In stomach cancer, mutations in TP53 R273H were associated with reductions in signature 1 and an increase in signatures 3,6,7,10,13,17,22. Note that although the number of signatures showing an increase suggests that it is the reduction in signature 1 that is important here and
although the cohort with R273H mutations were younger, the age difference between cohorts is not statistically significant.

In thyroid cancers mutations in BRAF V600E were associated with a reduction in signature 1 and increase in signature 22, which has previously been associated with aristolochic acid.

3.4.1.2.3 The extent of driver mutations in a specific tissue type is associated with the nature of the stresses on the cells in question

It is clear that driver genes are very different in different tissue types but to what extent are they associated with the nature of the stresses on the cells in question? To examine this question further for each cancer-associated gene, the percentage of samples mutated at different cancer sites was calculated and compared the extent to which the mean mutational fingerprint at those sites aligns with an idealised fingerprint. The idealised finger print is one that would provide the maximum opportunity for pathogenic mutation. See methods for information on how they were calculated.

In total 339 cancer-associated genes were tested: 129 genes associated with oncogenesis, 160 associated with tumour suppression and 50 that were identified in the cancer gene census as belonging to both categories. Of these, 226 genes (66%) showed a statistically significant (pvalue<0.05 permutation test) positive correlation between the percentage of samples mutated and the alignment between the ideal fingerprint and the mean mutational fingerprint. This suggests that most mutations in cancer-associated genes are partially opportunistic, mutating in those tissues where the endogenous and exogenous pressures on the cells are of the right type to create mutations which are pathogenic. Similar percentages were found for the tumour suppressor-associated genes and for the oncogene-associated genes.

It is expected that passenger genes will show this type of opportunistic behaviour, so it is possible that the findings could be the result of passenger mutations being included in with genuine driver mutations. However, opportunistic genes include those such as TP53 where

the gene is experimentally verified to be a driver for a large number of mutations and in a large number of tissues[208] when I changed the cut-off off to look at just those genes mutated in at least 10% of samples in at least one tissue type, the number of genes showing a statistically significant positive correlation between strength and alignment increased to 78% leaving just 14 that do not. These were: BRAF, CTCF, CTNNB1, EGFR, ERBB2, FGFR2, FGFR3, KRAS, NRAS, PBRM1, PIK3CA, PPP2R1A, PTEN and VHL.

This suggests that, except in rare cases, including the fourteen genes listed above, even when genes are drivers, and thus subject to positive selection pressure, the proportion of samples exhibiting mutations often depends on the chance nature of the mutations.

Example plots of the fourteen genes without statistically significant positive correlations are shown in figure 3.8. For comparison plots of fourteen genes with high mutation rates demonstrating statistically significant positive correlations are shown in figure 3.9.

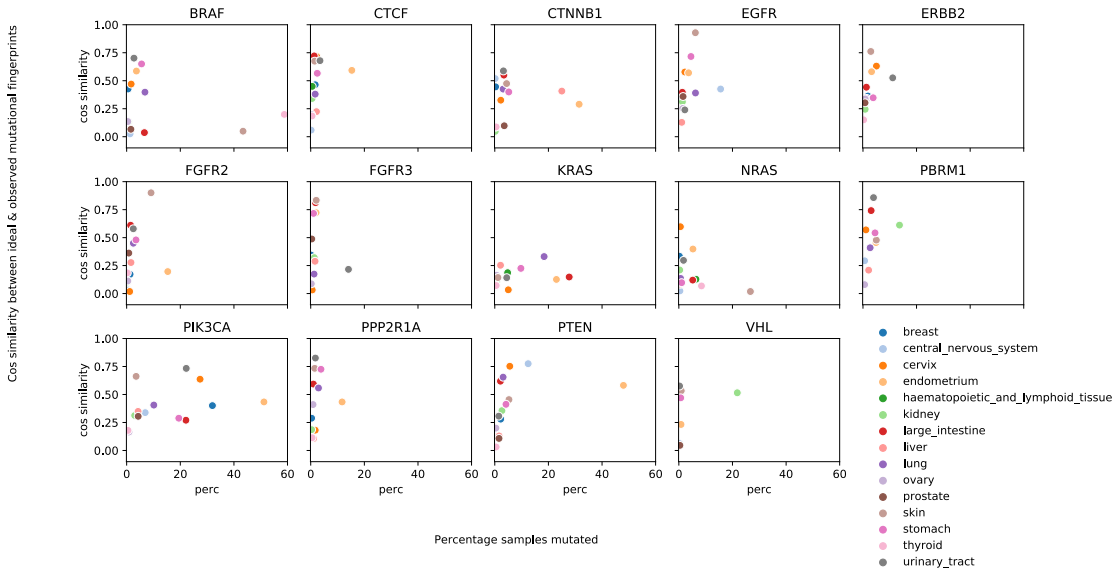


Figure 3.8: Scatterplots for the fourteen genes with a maximum mutation rate >10% that do not have a statistically significant correlation between the percentage of samples mutated

(x axis) and the cosine similarity between the ideal and observed mutational fingerprint (y axis).

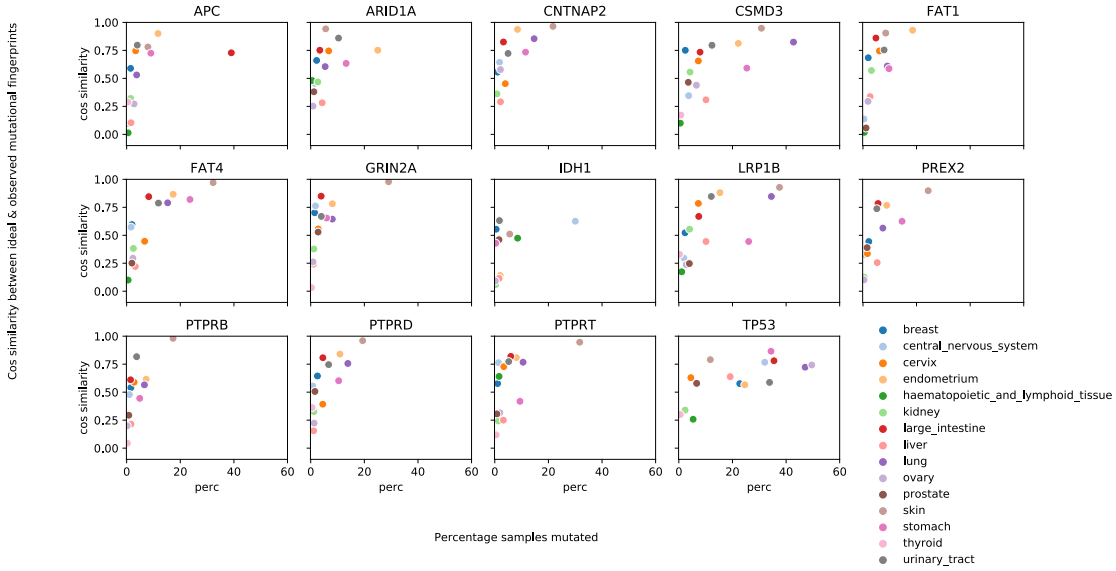


Figure 3.9 : Scatterplots for the fourteen genes with the maximum mutation rates that do have a statistically significant positive correlation between the percentage of samples mutated (x axis) and the cosine similarity between the ideal and observed mutational fingerprint (y axis).

3.4.1.3 Selective pressure on mutant residue choice in cancer hotspots

Experimental evidence evaluating the impact of uncommon substitutions at cancer missense hotspots is patchy. However, I posited that the mutational profiles of patients with a mutation in the gene in question can shed light on the question. For example, if an uncommon substitution would be likely to occur by chance given the mutational profiles it is less likely to be pathogenic than when the mutational profile allows for few mutations of a

type which would drive the substitution in question. I looked at extent to which mutational profiles are associated with the distribution of amino acid substitutions found at the most common cancer missense hotspots. BRAF, IDH1, KRAS, NRAS, PIK3CA and TP53 each have more than 100 TCGA samples mutated at a single position. These are shown in Figure 3.10.

Most of the missense hotspots have a distribution of mutant amino acids which is more closely aligned to that of the mutational fingerprints than would be expected by chance using a cosine similarity test. However, in some sites the mutational fingerprint accounts for up to 92% of the mutations seen whilst in others that drops to 50%. The clearest contrasts are provided by looking at the mutations at TP53 G245 and comparing these with those at BRAF pV600: Most of the possible mutations at TP53 G245 are represented at close to the rates I would expect from the mutational profile. This suggests that the different mutant amino acids are equally effective at changing the action of TP53 and their distribution occurs largely by chance. By contrast only BRAF pV600E is seen in my samples. The results summarised in table 3.3 are all significant at p-value <0.01. These p-values were generated using randomised trials where the expected distribution of mutations was used to generate random mutations, see methods for details.

Mutation	Sample size	Comments

BRAF V600	409	Distribution similar to mutational profile (pvalue =0.004). Selective pressure for V600E.
IDH1 R132	283	Distribution similar to mutational profile (pvalue <0.001). Selective pressure for R132H. R132C may be less effective.
KRAS G12	536	Selective pressure for G12A, G12D and G12V. G12S and G12R may be less effective.
KRAS G13	110	Distribution similar to mutational profile (pvalue <0.001). Selective pressure for G13D. G13C and G13V may be less effective.
NRAS Q61	154	Selective pressure for Q61K, Q61R. Q61H may be less effective. Absence of nonsense mutations confirms NRAS oncogenic status.
PIK3CA E542	244	Distribution similar to mutational profile (pvalue <0.001). E542G may be as effective as E542K.
PIK3CA E545	434	Distribution similar to mutational profile (pvalue <0.001). Selective pressure for E545K.
PIK3CA H1047	429	Distribution similar to mutational profile (pvalue =0.001). Selective pressure for H1047R. H1047Y may be less effective.
TP53 C176	115	Distribution similar to mutational profile (pvalue =0.007). Selective pressure for C176F. Complete destruction of protein through nonsense mutation may be less effective.
TP53 G245	155	Distribution similar to mutational profile (pvalue <0.001).
TP53 H179	126	Distribution similar to mutational profile (pvalue =0.002).No selective pressure seen.

TP53 H193	131	Distribution similar to mutational profile (pvalue <0.001). No selective pressure seen.
TP53 R158	114	Selection pressure for R158L and R158H. R158C and R158S may be less effective.
TP53 R175	327	Distribution similar to mutational profile (pvalue =0.027). Selection pressure for R175H. R175C and R175L may be less effective.
TP53 R196	114	Distribution similar to mutational profile (pvalue =0.002). Selection pressure for nonsense mutation.
TP53 R213	182	Distribution similar to mutational profile (pvalue =0.052). Selection pressure for nonsense mutation.
TP53 R248	455	Distribution similar to mutational profile (pvalue 0.008, cosine test). Some selection pressure for R248Q. R248L and R248W may be less effective.
TP53 R273	407	Distribution similar to mutational profile (pvalue <0.001, cosine test). Some selection pressure for R273L and R273C. R273H may be less effective.
TP53 R282	141	Distribution similar to mutational profile (pvalue =0.060). Selection pressure for R282 W. R282Q may be less effective.
TP53 Y220	151	Distribution similar to mutational profile (pvalue <0.001). There is an interesting lack of nonsense mutations suggesting incomplete destruction may be more useful.

Table 3.3 : Selective pressure at the site of most frequent missense mutations.

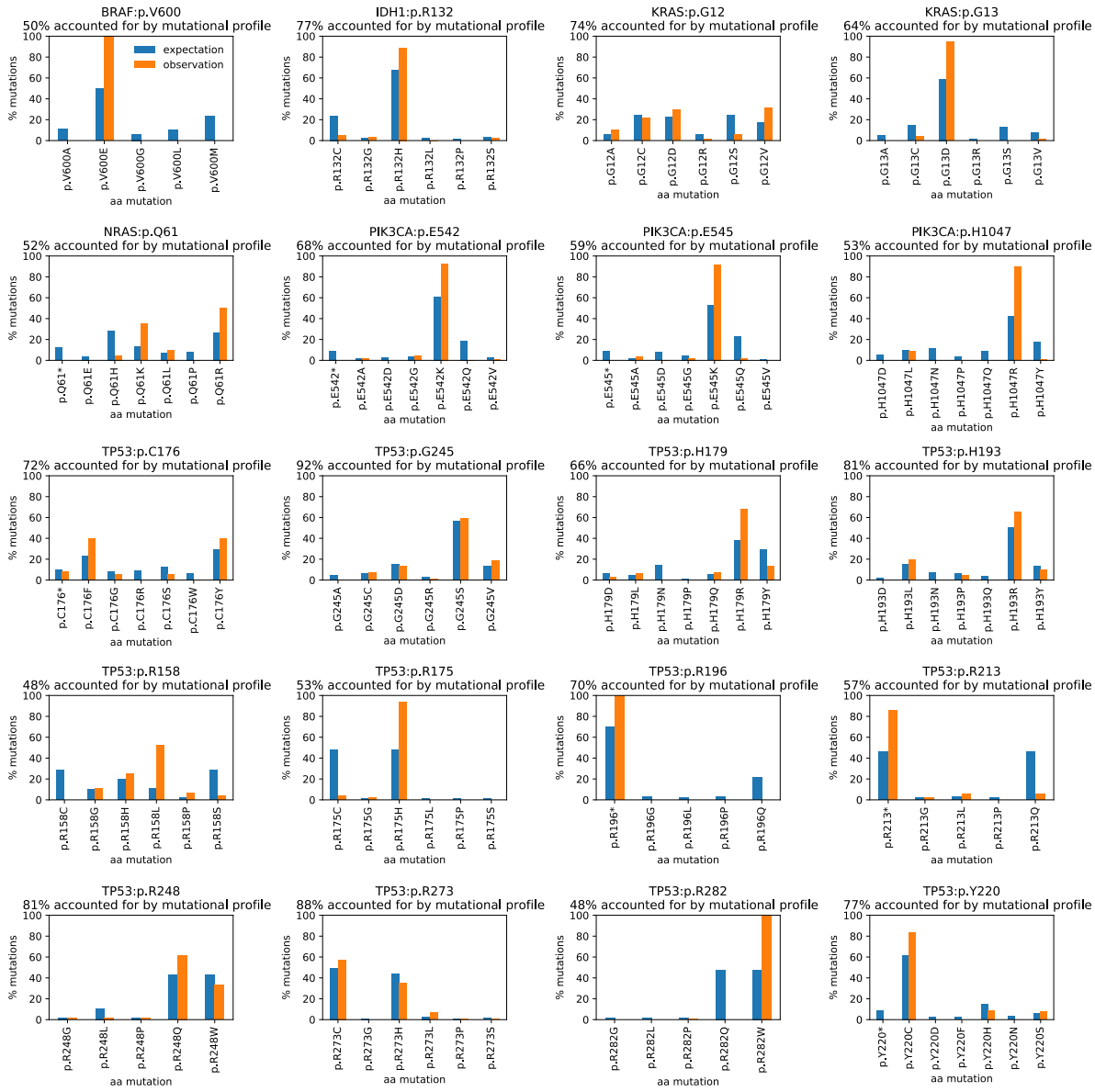


Figure 3.10: Comparison between the observed and expected frequency distribution of mutant amino acid in each of the 20 hotspots considered.

3.4.2 BRAF case study

BRAF was most frequently mutated gene in thyroid (59% of samples) and skin cancers (44%). In both cases V600E predominated but in skin cancers a handful of other mutations (G469A, K601E, P367S, K183E, N581H, K641E) also occur.

The V600E mutation is caused by a G (T>A) G transversion which is common mutation in thyroid cancer but not in skin or other cancers (see figure 3.1). Interestingly, there is no significant correlation between the percentage of samples mutated and the alignment between the observed mutational fingerprint and the idealised mutational fingerprint for the gene in either cancer type.

One possible explanation of the mismatch between that mutational fingerprint and the idealised mutational fingerprints may have been the selection of mutations predicted to pathogenic. FATHMM[216] is used to predict the pathogenicity of the mutations and could have been misclassified. I therefore reclassified the mutations using Polyphen2[217]. Polyphen2 confirmed 55 of the 76 BRAF missense mutations as either possibly or probably damaging, and did not give rise to a significant correlation.

An alternative explanation is provided by the experimental classification of BRAF mutations based on their mechanisms for activation of the MAPK pathway. These classifications separate the impact of the mutations into those which give rise to active BRAF monomers (the V600 mutants) from constitutively active dimers (K601E, L597Q, G469A, G469V, or G464V) and from those which result in little or no BRAF kinase activity (G466E and D287H) [194]. It is thus possible that no correlation can be seen because different interactions are selected for in different cancers.

Analysis at the V600 site identifies far more V600E substitutions than expected. If all available amino acid substitutions were equally effective as onco-proteins, 20% of the V600

substitutions would occur as V600M (see figure 3.10). However, V600M does not occur in these samples at all, and it is reported as very rare, around 0.3% in melanomas for example [218]. This suggests that there is strong selection for glutamate over the alternative amino acids.

3.4.3 Conclusions and discussion

In this study I have shown that there is an association between patients' mutational signature profile and their pathogenic mutations in fourteen genes, and with mutations at a specific sites in five genes with oncogenic hotspots. Most of the frequently mutated genes are highly specific to particular tissue types. That is, the mutations are very common in just one or two tissue types and are otherwise rare. I find that for most such genes the mutations most frequently occur in those tissues where the endogenous and exogenous pressures on the cells are of the right type to create mutations which are pathogenic. However, this is not universally true: in fourteen of the genes (BRAF, CTCF, CTNNB1, EGFR, ERBB2, FGFR2, FGFR3, KRAS, NRAS, PBRM1, PIK3CA, PPP2R1A, PTEN and VHL) there is no such connection.

Looking specifically at mutation hotspots, I find that the distribution of mutant amino acids is more closely aligned to that of the mutational fingerprints than would be expected by chance using a cosine similarity test. Nevertheless, there is variation. For mutations at TP53 G245, where mutations destabilize the protein [219], the sample fingerprints account for 92% of the distribution. On the other hand, for mutations at TP53 R248, where mutations lead to a gain of function [220], greater selectivity is seen and only 48% of the distribution is accounted for by the sample fingerprint.

After I completed this work Poulos et al. [221] and Temko et al. [222] published analyses identifying associations between mutated genes and mutational signatures. Poulos et al used logistic regression analysis to test associations between driver mutations and mutational signatures and found 39 significant associations. Temko et al. compared the levels of signatures in cancers harbouring the mutations to those in cancers that did not harbour the mutations. They tested whether observed frequencies of each driver mutation differed significantly from those expected based on mutational process activity alone, using a Mann-Whitney U test. Using this approach, they identified 43 significant correlations between signature activity and driver mutations. Temko et al. also looked for differences between the predicted and observed mutational likelihood of mutant amino acids for 9 genes, disaggregating the data into specific cancer types. They found selection pressure in 19% of the pairwise comparisons.

As well as the different statistical tests used, there are other differences between the approaches. For section 3.4.1.2.1 , this includes the methods used to identify mutational signatures within samples and the tissue types used to disaggregate samples. For section 3.4.1.3 Temko et al. use sample signatures rather than the sample fingerprint directly and aggregate different mutations when looking for selective pressure.

Despite this, the association between mutations in PIK3CA E545K and an increase in signature 2 in breast cancer was found in all three studies, and both Poulos and I found an association between IDH1 R132C and reduction in signature 1 in cancer of the central

nervous system. In addition, Temko et al. also found particularly strong selective pressure for PIK3CA H1047R, NRAS Q61R and BRAF V600E.

The large differences in the methodologies suggests that the association between PIK3CA and signature 2 in breast cancer, and between IDH1 and signature 1 in cancer of the central nervous system are particularly robust. Both PIK3CA and IDH1 play a role in DNA damage repair which suggests that the mutation may change the profile of further mutations, rather than the differences in signature giving rise to an increased frequency in mutations in IDH1/PIK3CA [223][224]. It also suggests that there may be substantive differences in the mutations found at PIK3CA H1047, NRAS Q61 and BRAF V600 which could potentially require different therapies.

In conclusion, there is clear evidence that the physical cause of the cancers, as represented by their mutational profiles, is associated with the particular genes that become mutated. This suggests that mutations amongst driver genes include an element of opportunism: a conclusion which is strengthened when I look at the association between mutational load and the probability of a mutation in a frequently mutated cancer-associated gene. Nevertheless, selectivity is still evident and some mutations, such as BRAF V600M are much less common that the mutational fingerprints would suggest. The results for TP53 are of particular interest because the hotspots include some mutations which cause loss of function and others which cause gain of function. It is possible that, with sufficient data, the method used here could provide a test to distinguish between loss of function and gain of function mutations in genes which may possess both.

4 Mutational signatures in bacteria

4.1 Abstract

Mutational signatures are characteristic combinations of nucleotide substitutions within the DNA that result from specific mutagenesis processes. Within the field of cancer, identifying mutational signatures has provided considerable insight into the biological mechanisms involved in both carcinogenesis and somatic mutagenesis in healthy cells.

In this paper using a similar technique, I investigate the underlying mutational processes involved bacterial evolution. First, I identify mutational fingerprints for 16 bacterial species, mostly human pathogens, and then decompose them into mutational signatures using non-negative factorisation. By comparing the signatures between species, those observed in human cancer samples and those observed in cell lines treated with environmental mutagens, I identify defective incorrectly-repaired alkylation, use of POLE and defective mismatch repair as a potential mechanisms in driving bacterial mutations.

4.2 Introduction

Work on human cancer samples has demonstrated that different sources of damage to DNA give rise to characteristic patterns of substitution mutations [152], [166], [167], [225]. These sources of damage may be exogenous or endogenous in nature. Defects in DNA damage repair pathways (DDR) are also detectable from sequencing data. In order to uncover these patterns, substitution mutations are commonly represented as mutational fingerprints and then decomposed to reveal common mutational signatures.

Mutational fingerprints are vector representations of all the substitution mutations that have occurred in a set of genes compared to the consensus genes. Typically, the substitution types are classified by considering both the substituted nucleotide and the flanking nucleotides and are written as a quadruplet. Thus CCG > CTG is written CCTG. Although there are twelve possible ways of substituting a nucleotide and thus 192 potential quadruplets, substitutions on the leading or lagging strand of DNA are considered equivalent. For example, CGT > CAT is considered equivalent to ACG > ATG and hence CGAT ~ ACTG. This is used to reduce the number of possible substitution types to 6, namely: C>A, C>G, C>T, T>A, T>C, T>G and the overall number of possible substitution types to 96. A mutational fingerprint is then formed by counting the number of substitutions in each of these 96 categories.

Once several mutational fingerprints have been identified it is possible to decompose them into mutational signatures and associate these signatures with different potential causal factors. Decomposition is frequently done using non-negative matrix factorisation as this allows a simple interpretation of the results as specific fingerprints can be identified as a linear sum of different signatures[226].

Within cancer samples most mutations are passenger mutations, in that they do not have a significant impact on the cell viability and are not under great selective pressure. Hence, the mutational fingerprint captures the mutational history of the cancer cell, and the decomposed mutational signatures can shed light on the action of specific exogenous or endogenous genotoxins to which the cell has been exposed. Within the field of cancer

genomics, considerable work has been undertaken to identify the cause of particular signatures[103], [205] and extensive collections of such signatures have been documented .

Signatures obtained from somatic cancer cells have revealed causal links with a number of generic mutational processes . These include the deamination of 5-methylcytosine at CpG not repaired prior to DNA replication [227], the C:G > T:A transitions occurring mainly at dipyrimidines caused by exposure to UV light [161], defects in DNA damage repair pathways and a number of signatures associated with alkylating agents [197].

In contrast to the chromosomal organisation of human DNA, the DNA in most bacteria is contained in a single circular molecule, called the bacterial chromosome together with several plasmids – small circular DNA molecules. Acquisition of mutations followed by a selective pressure results in most bacterial species having a variety of different sub-species comprising different strains, some of which have been fully sequenced[129]. Rapid bacterial evolution can take place through the acquisition of novel genetic elements, including new genes and fragments of genes, from the accessory gene pool, the transfer of circular DNA called plasmids through cell to cell contact known as conjugation, the inclusion of DNA from the environment known as transformation and finally transduction, the transfer of DNA by bacteria-specific viruses called bacteriophages [228].

However, in addition to these mechanisms, bacteria are also at risk from endogenous and exogenous mutational pressures, and have a suite of DNA damage response pathways. For example, the Nucleotide Excision Repair (NER) pathway is well conserved across many bacterial species [29]. However some of the genes present in Escherichia coli's pathways for

base excision repair (BER), mismatch repair (MMR), and direct repair (DR) do not have an orthologue in other bacterial species [229] suggesting that there may be significant differences in these pathways in other bacterial species.

In this chapter I investigate the underlying mutational processes involved bacterial evolution. First, I identify mutational fingerprints for 16 bacterial species, mostly human pathogens by identifying unique silent substitutions. Then I decompose these fingerprints into mutational signatures using non-negative factorisation. By comparing the signatures between species, those observed in human cancer samples and those observed in cell lines treated with environmental mutagens, I identify mutational processes that could underly bacterial mutations.

4.3 Methods

For each bacterial species, the genes from genomes of available strains were downloaded from ensemblgenomes.org by downloading the cDNA from the site as follows and 200 strains selected for analysis[129].

4.3.1 Clustering bacterial sub-species

The cDNA of all the bacterial genes were translated to the amino acid sequence and clustered using the MMseqs program to find orthologous genes [153]. 100 orthologous gene families that were shared by all 200 strains, had no paralogs and were of similar length between 900 and 2250 bps [230] were identified. The nucleotides corresponding to these orthologs were then clustered using ClustalOmega[154] and the sequence identity found for each gene in each bacterial strain. These sequence identities were used as a similarity

distance allowing me to hierarchically cluster the strains using Ward's method[231], using a cophrenetic distance cut-off of 0.7.

To identify the relationships between the sub-species and species I used the twenty-five primers from ten genes identified by Santos et al[232]. In the majority of strains these genes were labelled within the bacterial genomes, making identification straight forward. However, where this was not the case, I used MMseqs to identify the relevant, using a sample gene from *Escherichia coli* as a template. This method successfully retrieved all the annotated genes. The genes were then aligned against the primers and a vector built up showing for each nucleotide whether or not the primer was conserved. Finally for each of the subspecies the mean vector was found and the subspecies clustered again using Ward's method [231].

4.3.2 Generating mutational fingerprints

Mutational fingerprint analysis was carried out on the bacterial subspecies identified. For each cluster, genes were identified that were shared by all of the strains in the cluster and had no paralogues. For each of these genes, a consensus sequence was identified and single base substitutions (SBSs) called against the consensus sequence. In order to remove selection bias, I accepted only mutations that were silent and only unique silent substitutions were accepted, as I consider these more likely to reflect recent mutational pressures rather than the historical distribution of established alleles. The substitutions were classified according to the wild and mutant nucleotides and the value of flanking nucleotides. In common with protocol in cancer mutations I use the quadruplet TCTG to

indicate the point substitution from TCG to TTG. There are 192 such possibilities. However, it is assumed that mutations are equally likely along both DNA strands. To reduce the number of possibilities by two to 96, mutations of wild type A,G are read along the opposite strand so that all mutations are of the form $N^1X > N^2N^3$ where N^1N^2 and N^3 may be any nucleotide but X is constrained to be either C or T. So, to give a concrete example AGAC is short for AGC>AAC which may be read as GCT>GTT on the opposite strand i.e. GCTT. By doing this I built up a distribution of all the SNPs called, and also just those SNPs that are unique to a single strain at any one position.

So that I could compare these distributions between different bacterial species I also identified the distribution of all the quadruplets within the shared genes that could have arisen as a result of a silent substitution. So, for example threonine can be coded by ACA, ACC, ACG and ACT. As a result, the trinucleotide ACC followed by T could give rise to any of the three quadruplets CCAT, CCGT, CCTT each representing a silent substitution. By dividing the SNP distributions by the underlying distribution of possible silent SNPs I end up with a frequency distribution of all the silent SNPs. This provided an overall picture of the stable variation of the bacterial genome, and also for each strain. The mean mutational fingerprint found for each cluster was used as a similarity measure enabling me to cluster the different bacterial strain clusters.

4.3.3 Generating mutational signatures

For each cluster of bacterial strains, the matrix formed from the individual mutational fingerprints was decomposed using non-negative matrix factorisation to give five mutational signatures per bacterial strain cluster, and a weights matrix. Each signature was normalised

to sum to one, and the weights adjusted accordingly. For each signature the weights matrix was used to identify the percentage of strains where that signature accounts for more than 50% of the fingerprint. The mutational signatures were then clustered to identify similar signatures in different bacterial strain clusters. A cophrenetic distance of 2 was used as a cut-off giving 40 clusters. Of these 24 were shared by 3 or more strain clusters. For these 24 clusters the median signature of the clustered signatures was found.

4.3.4 Identification of DNA Damage Repair (DDR) genes

To identify the DDR genes for each of the downloaded bacterial strains I clustered the translated nucleotides with the 250 known human DDR genes as well as all the genes which have been identified in the reference genome for Escherichia coli as being involved in DNA repair pathways. Again this was done using the MMseqs package [73], [149]. I then counted the number of strains where an ortholog was identified. The number of strains with a specific DDR gene was used as a similarity measure enabling the different bacterial species to be clustered by the presence or absence of DDR orthologs.

4.3.5 COSMIC cancer signatures

To be able to compare the most common bacterial signatures with those found in the COSMIC cancer signatures single based substitutions (version 2) [185] normalisation of human based signatures were undertaken. The raw COSMIC signatures are derived from exomic frequencies and have not been corrected for the trinucleotide distribution in *Homo Sapiens*. The human exome was downloaded from Ensembl Biomart [16]. Using the trinucleotide distribution for each frame and the corresponding theoretical quadruplet distribution was calculated. Dividing the cancer signatures by this quadruplet distribution

allowed direct comparison with signatures derived from human to those derived from the bacteria.

To compare these signatures the cosine similarity between each of the signatures was calculated. To provide a random cosine similarity against which to measure statistical significant, ten thousand random signatures were generated and the cosine similarities calculated. These were compared with the cosine values between 24 bacteria and 30 cancer signatures and corrected for multiple testing using the Benjamini Hochberg method [202].

4.3.6 Carcinogen- derived human signatures

I also compared the bacterial signatures with the signatures derived experimentally by Kucab et al. These were derived by exposing human stem cells to environmental carcinogens [197]. The probabilities associated with these signatures were downloaded from Signal [196].

(https://signal.mutationalsignatures.com/explore/mutagens?group&hasSignature=true&name=)

These signatures are derived from human genomic mutation frequencies, rather than exomic frequencies, and had not been corrected for the trinucleotide distribution in *Homo Sapiens*. The human genome was downloaded from Ensembl Biomart [16] and the trinucleotide distribution in each frame – and hence the quadruplet distribution- was identified. The signatures were then converted to substitution frequencies, and the comparison with bacterial signatures carried out as before.

4.4 Results

4.4.1 GC content for different bacteria

In total I considered 16 bacterial species: *Acinetobacter baumannii*, *Bacillus cereus*, *Burkholderia pseudomallei*, *Clostridioides difficile*, *Enterococcus faecalis*, *Enterococcus faecium*, *Escherichia coli*, *Klebsiella pneumoniae*, *Listeria monocytogenes*, *Mycobacterium abscessus*, *Mycobacterium tuberculosis*, *Neisseria gonorrhoeae*, *Neisseria meningitidis*, *Pseudomonas aeruginosa*, *Salmonella enterica*, and *Streptococcus pneumoniae*.

For each of the species I considered 200 different strains. I first identified the GC content for each bacterial species as shown in figure 4.1 below. These range from a mean of 30% in *Clostridium difficile* to 68% in *Burkholderia pseudomallei*. For most of the bacterial species, the difference between the maximum and minimum level of GC content was less than 5%. However, for *Clostridioides difficile* there are outlying strains whose GC content varies by more than 24%.

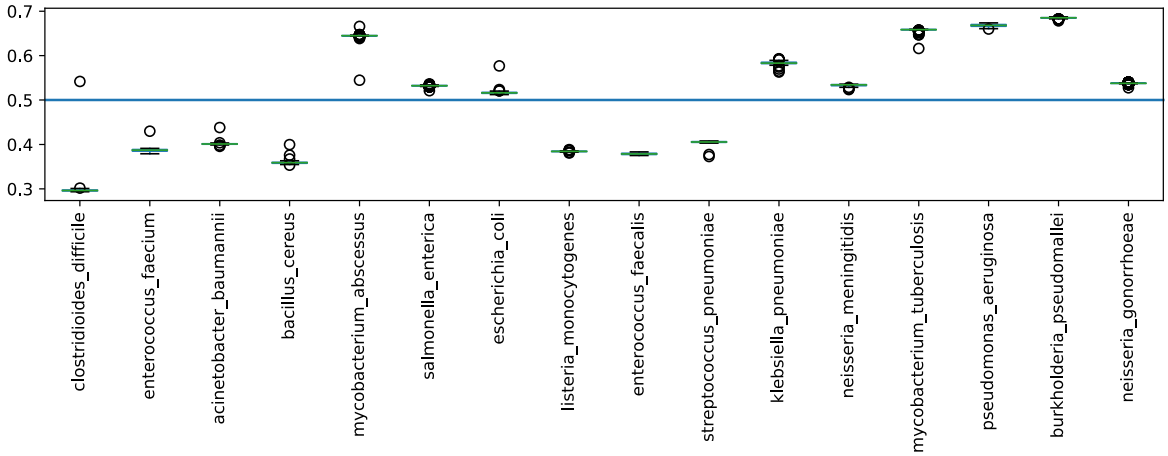


Figure 4.1 legend

Boxplot of the GC content for each of the bacterial species. These range from 30% to 68%.

4.4.2 Distribution of genes in strains within a species

Next, I analysed the conservation of genes within a species by identifying in how many strains each gene was observed, see figure 4.2 below. For each of the bacterial species considered, genes show a bimodal distribution, with each gene being either part of a very small gene family or shared by nearly all the strains. For most of the bacterial species around 60% of genes are nearly universal. *Bacillus cereus* is an exception to this: only 26% of genes belong to an orthologous gene cluster shared by nearly all of the strains. A small minority of genes belong to between 5 and 95% of the bacterial strains. These genes are likely to be adaptations giving strains individual characteristics (see for example [233]–[235]).

Between 7%-34% of all the genes are belong to a gene cluster shared by less than 5% of the strains. This is most pronounced in *Bacillus cereus* where 34% of the genes are only found in an orthologous gene cluster shared by less than 5% of the strains. In *Escherichia coli* some 15% of genes are thought to result from horizontal gene transfer, so it is possible that these anomalous genes may be candidates for horizontal gene transfer [236].

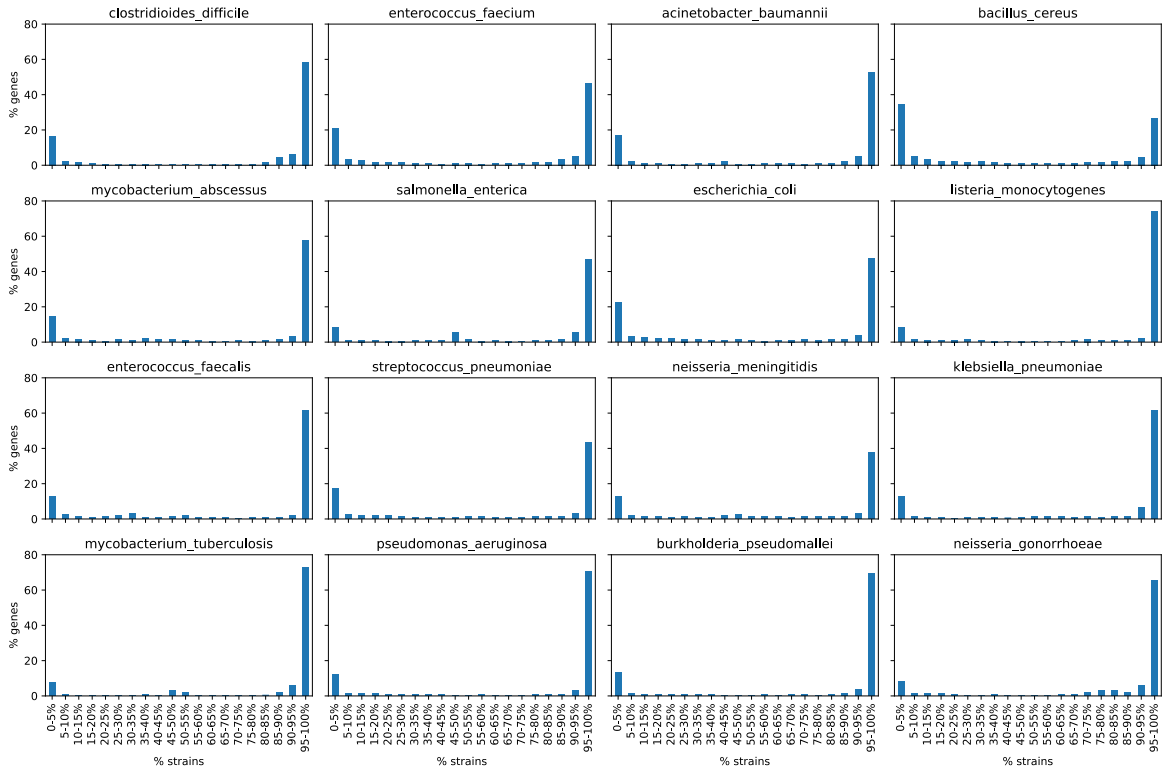


Figure 4.2: Histogram showing the percentage of total genes that belong to an orthologous gene family shared by a given percentage of different strains of the same bacterial species.

4.4.3 Deriving bacterial sub-clusters for each bacterial species

Previous work has shown that in genetic terms the membership of a bacterial species can be quite broad. Membership is often based on having a gene sequence similarity >97% for the ubiquitous 16S rRNA gene sequence [237]. However, this measure can be insufficient to distinguish between closely related strains [238]. Comparing groups of conserved proteins improves sensitivity and nucleotide identity has been shown to be a robust measure of evolutionary distance [232], [239].

To ensure that I am finding mutational fingerprints on sub-species that have close intraspecific genes I grouped the bacterial species into sub-species using sequence identity between 100 conserved gene families; each between 900 and 2250 bps long with no paralogs. Each gene family was aligned to form a consensus sequence, and the sequence identity found between the consensus sequence and each of the corresponding strain sequences. These sequence identities were used to cluster the strains into a phylogenetic tree. I used a cophrenetic distance 0.7 as a cut-off to group each of the species into sub-species for further analyses.

Each bacterial species showed distinctive evolutionary patterns structures with their phylogenetic tree branching into between one and four sub-clusters, as well as outlier strains that do not belong to any of the clusters. Figure 4.3 below shows the tree derived for *Bacillus cereus.* Each sub-species, shown as cyan, red or green, comprises at least 25 strains within a cophrenetic distance of less than 0.7 of one another. The maximum cophrenetic distance between bacterial strain clusters is generally between 0.8 and 3, but for *Enterococcus faecalis* the maximum cophrenetic distance is greater than 14 suggesting that there is considerably more genetic variation in *E. faecalis* than the other bacteria considered.

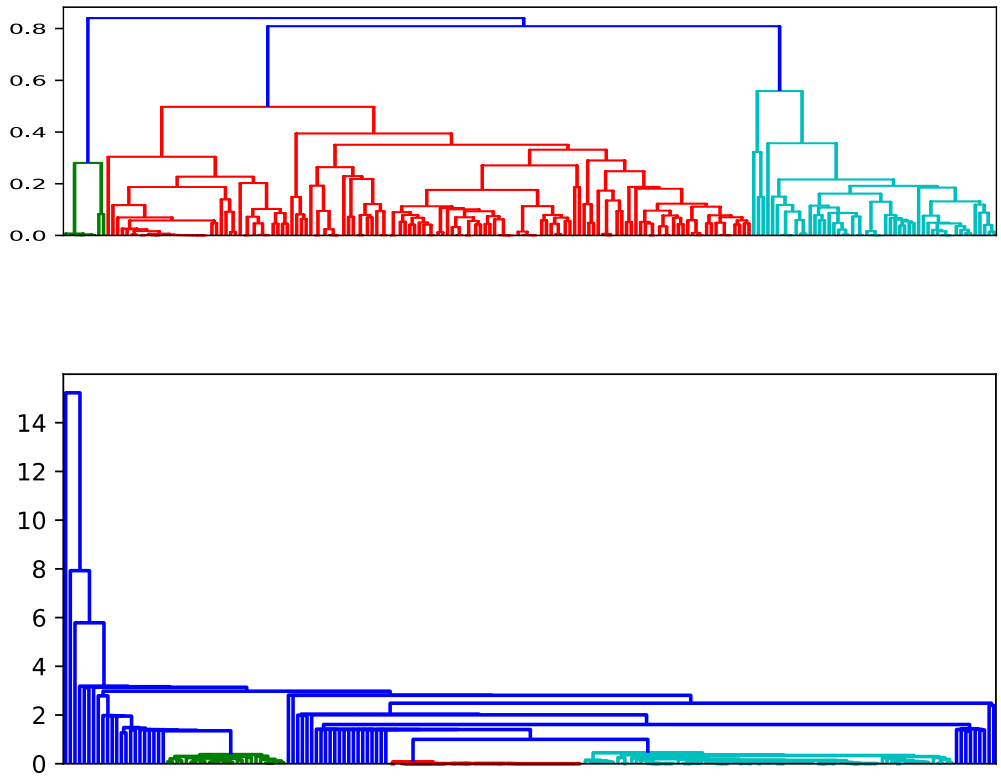


Figure 4.3: Bacterial strain clusters across conserved genes using sequence identity from the consensus gene as a measure of distance. This first example shows *Bacillus cereus* and is fairly typical in showing between 1 and 4 clusters using a cophrenetic distance cut-off of 0.7. The second example is *Enterococcus faecalis* which is not typical. Many of the strains do not cluster closely with any other strains and the maximum cophrenetic distance is greater than 14. Bacterial strain clusters for each of the bacterial species are shown in appendix 2 figures 4.1a-4.1o.

4.4.4 Mutational frequencies in different bacteria

All further analysis was done on the basis of the bacterial sub-species identified, which I denote red, green and cyan according to the colour of the cluster. For each gene in each bacterial sub-species, I constructed an ancestral sequence corresponding to the base

occurring most frequently at each position in the sequence. I identified all the observed mutations that occurred in each strain against this consensus sequence.

Silent mutations accounted for between 31% of all mutations (*Salmonella enterica*) and 74% of all mutations (*Enterococcus faecalis*) in the different bacterial sub-species. This is a far higher percentage of silent substitutions than might be expected if there was no selective pressure. If each substitution took place on each nucleotide with equal probability, I calculated that only 19%-25% of the mutations would be silent for the bacterial species examined. It is likely I am seeing signs that the silent mutations are much more tolerated than missense mutations, as they are unlikely to be under selective pressure as predicted by Kimura in 1968 [240].

For each gene I identified the unique silent mutations that occurred in each strain against the ancestral sequence. If a mutation occurred in more than one strain it was not included further in this analysis. These silent substitutions are unlikely to be under evolutionary pressure and so allow me to focus purely on mutational stresses acting on the cells, without considering the evolutionary consequences of particular mutations.

On average each of the bacterial strains had 5 unique silent substitutions (USSs). However, there was considerable variation: the *Clostridioides difficile* strains have an interquartile range of only 1-3 USSs whereas one of the sub-species of *Enterococcus faecalis* has an interquartile range of 43-170 USSs showing again the pronounced genetic divergence in *E. faecalis*. Full details are set out in table 4.1.

Bacterial species	**Cluster colour**	**Number of Strains**	**Number of Mutations (Median)**	**Number of Mutations (Interquartile range)**
Acinetobacter baumannii	red	63	2	1-10
Bacillus cereus	red	69	5	3-9
Bacillus cereus	cyan	64	27	8-51
Burkholderia pseudomallei	red	97	65	2-90
Burkholderia pseudomallei	cyan	93	1	0-2
Clostridioides difficile	red	33	1	1-3
Enterococcus faecalis	red	24	39	15-75
Enterococcus faecalis	cyan	51	12	4-56
Enterococcus faecalis	green	25	102	43-170
Enterococcus faecium	red	25	1	1
Escherichia coli	red	152	8	2-18

Escherichia coli	green	47	25	13-64
Klebsiella pneumoniae	red	118	1	0-9
Klebsiella pneumoniae	cyan	50	8	1-97
Listeria monocytogenes	green	34	2	1-3
Mycobacterium abscessus	green	54	6	2-8
Mycobacterium abscessus	red	47	2	1-9
Mycobacterium tuberculosis	red	83	12	5-21
Mycobacterium tuberculosis	green	103	7	2-15
Neisseria gonorrhoeae	red	37	5	2-10
Neisseria gonorrhoeae	green	62	2	1-11
Neisseria meningitidis	red	128	2	0-7
Pseudomonas aeruginosa	cyan	148	5	2-36

Salmonella enterica	red	63	3	1-8
Salmonella enterica	green	28	7	1-48
Salmonella enterica	cyan	46	8	1-25
Streptococcus Pneumoniae	cyan	44	19	7-79
Streptococcus Pneumoniae	red	87	4	2-11

Table 4.1: Range of unique silent substitutions found in samples of different bacterial sub-species.

4.4.5 Mutational fingerprints in bacteria

For each bacterial sub-species, I generated a raw mutational fingerprint: this was a count of all the unique silent substitutions (USSs). In common with the protocol used to describe cancer mutations, USSs were classified according to the wild and mutant nucleotides and the flanking nucleotides. For example, I used the quadruplet TCTG to describe the point substitution TCG > TTG. To compare mutational rates in different species, I then normalised the raw mutational fingerprint against the background silent substitution count within the bacterial sub-species.

4.4.5.1 Distribution of mutations in bacteria

For each bacterial subspecies I identified the mean of these mutational fingerprints (MMF). These are shown in the figure 4.4 below. More detailed figures for each of the six mutation types C>A, C>G, C>T, T>A, T>C, T>G, are included in the appendix 2, figures 4.2a-4.2f .

In general, nucleotide transitions were observed much more commonly than nucleotide transversions and of these T>C transitions were far more commonly observed than C>T transitions.

Transversions from cytosine, i.e. C>A and C>G, were comparatively rare, though elevated levels of CC>AC and ACC>AGC mutations were seen in some bacterial sub-species, including *E. coli* and *P. aeruginosa* respectively. Transitions to thymine were much more common, particularly CT>TT substitutions.

Transversions from thymine to adenine, T>A, were also comparatively rare for all bacterial species. However, substantially elevated levels of T>G in specific contexts were seen for seven bacterial subspecies including *B. pseudomallei* and *P. aeruginosa*.

Transitions of the form T>C were the most commonly observed transition, particularly GT>GC substitutions.

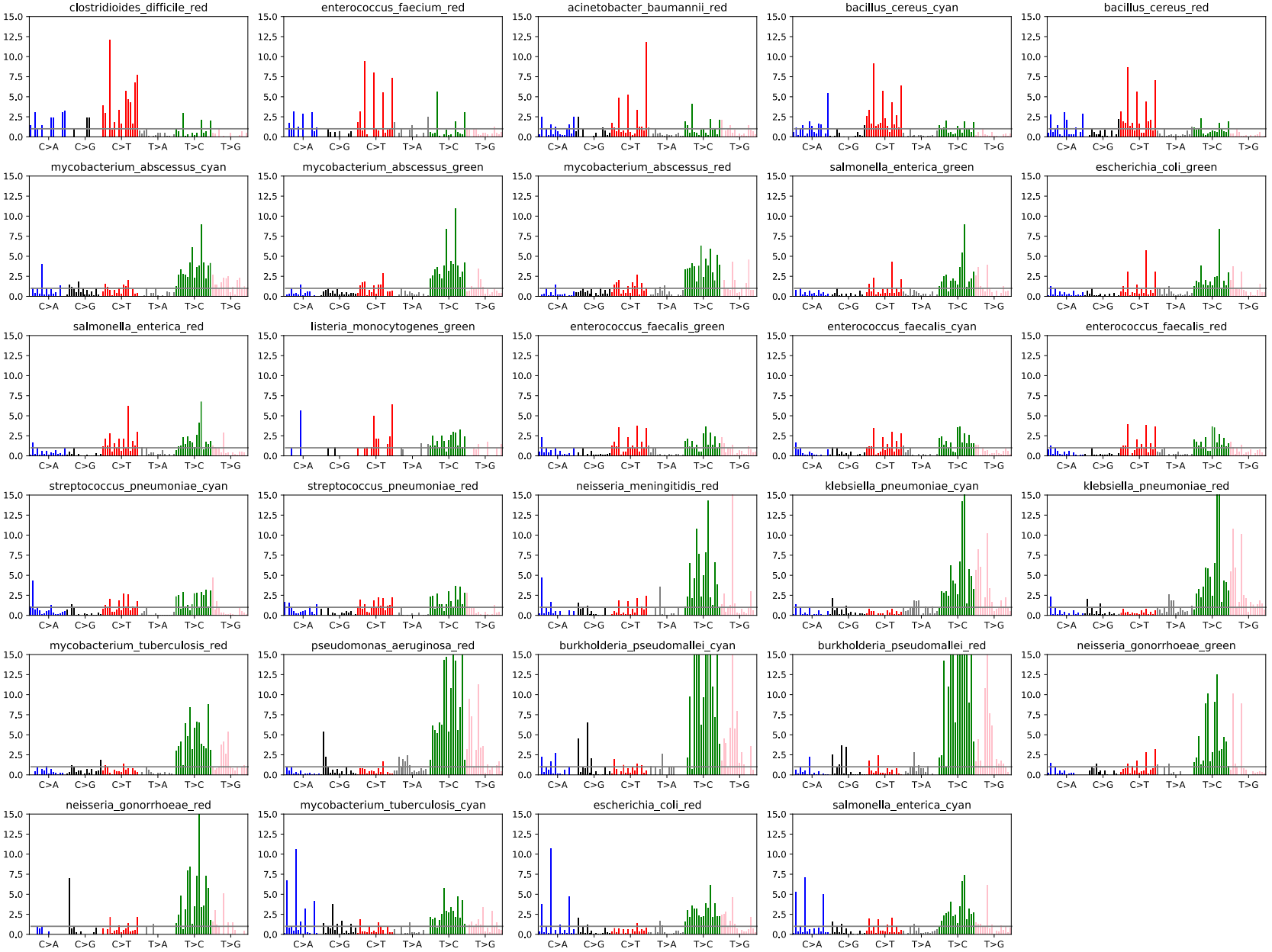

clostridioides_difficile_red
enterococcus_faecium_red
acinetobacter_baumannii_red
bacillus_cereus_cyan
bacillus_cereus_red
mycobacterium_abscessus_cyan
mycobacterium_abscessus_green
mycobacterium_abscessus_red
salmonella_enterica_green
escherichia_coli_green
salmonella_enterica_red
listeria_monocytogenes_green
enterococcus_faecalis_green
enterococcus_faecalis_cyan
enterococcus_faecalis_red
streptococcus_pneumoniae_cyan
streptococcus_pneumoniae_red
neisseria_meningitidis_red
klebsiella_pneumoniae_cyan
klebsiella_pneumoniae_red
mycobacterium_tuberculosis_red
pseudomonas_aeruginosa_red
burkholderia_pseudomallei_cyan
burkholderia_pseudomallei_red
neisseria_gonorrhoeae_green
neisseria_gonorrhoeae_red
mycobacterium_tuberculosis_cyan
escherichia_coli_red
salmonella_enterica_cyan
15.0
12.5
10.0
7.5
5.0
2.5
0.0
C>A
C>G
C>T
T>A
T>C
T>G

Figure 4.4: Mean mutational fingerprints for each of the bacterial sub-species. Each bar represents the relative frequency of unique silent substitutions seen in a particular nucleotide type e.g. ACC > ACG. The frequencies are normalised with respect to the number of possible silent substitutions across the consensus gene for each orthologous gene family. Mutations are generally dominated by T>C mutations.

4.4.6 Comparing mutational fingerprints

Visual inspection suggests that the fingerprints segregated into two major types: those where C>T mutations predominate and those where T>C mutations predominate (see figure 4). Subspecies from the same species frequently had similar fingerprints. For example, the signatures associated with subspecies of *Bacillus cereus* are very similar, as are those associated with *Mycobacterium abscessus* and *Klebsiella pneumoniae.* In some species there was more variation ,for example in both *Escherichia coli* and *Mycobacterium tuberculosis* one of the subspecies showed distinctive patterns of C>A mutations not present in the other subspecies in that bacterium.

The different mean mutational fingerprints shown in figure 4.4 above for each bacterial sub-species were clustered by comparing the Euclidean distance between them. The main clusters naturally split into 5 groups, see figure 4.5. Even though the fingerprints have been normalised for all the possible silent mutations in the consensus sequences, the sub-species within each cluster all share a similar GC content, with more T>C mutations in those bacterial sub-species that have balanced CG/AT nucleotides.

In fingerprint group 1 there are elevated numbers of CT>TT mutations. The group members are: *Clostridiodes difficile, Enterococcus faecium, Acinetobacter baumannii, Bacillus cereus.* All of these are CG poor, with the median number of CG pairs accounting for 30-40% of the total nucleotide pairs.

Bacterial sub-species in fingerprint group 2 show elevated levels of T>C mutations. The members included *Mycobacterium abscessus* and some, but not all, strains of *Escherichia coli,* and *Salmonella enterica.* These bacteria are relatively CG ric with a median CG content 52-64%.

Subspecies in fingerprint group 3 have slightly elevated levels of both C>T and T>C mutations. The group includes *Listeria monocytogenes, Enterococcus faecalis,* and *Streptococcus pneumoniae.* These bacteria are slightly CG poor with median CG content 38-41%.

The largest fingerprint group is group 4 which includes *Burkholderia pseudomallei*, *Pseudomonas aeruginosa*, *Klebsiella pneumoniae*, *Neisseria gonorrhoeae*, *Neisseria meningitidis* and one strain of Mycobacterium tuberculosis. These bacteria show highly elevated levels of T>C mutations, as well as some T>G mutations particularly CTC>CGC. These bacteria are relatively CG rich: median CG content 53-69%.

Finally the smallest group, fingerprint group 5, shows slightly elevated levels of T>C mutations. Members include one strain each of *Mycobacterium tuberculosis*, *Salmonella enterica* and *Escherichia coli.* The fingerprints are distinguished from those of fingerprint

group 2 by the elevated levels of NCC>NAC mutations where N is either A,C or T. These bacteria are slightly CG rich: mean CG content 52-66%.

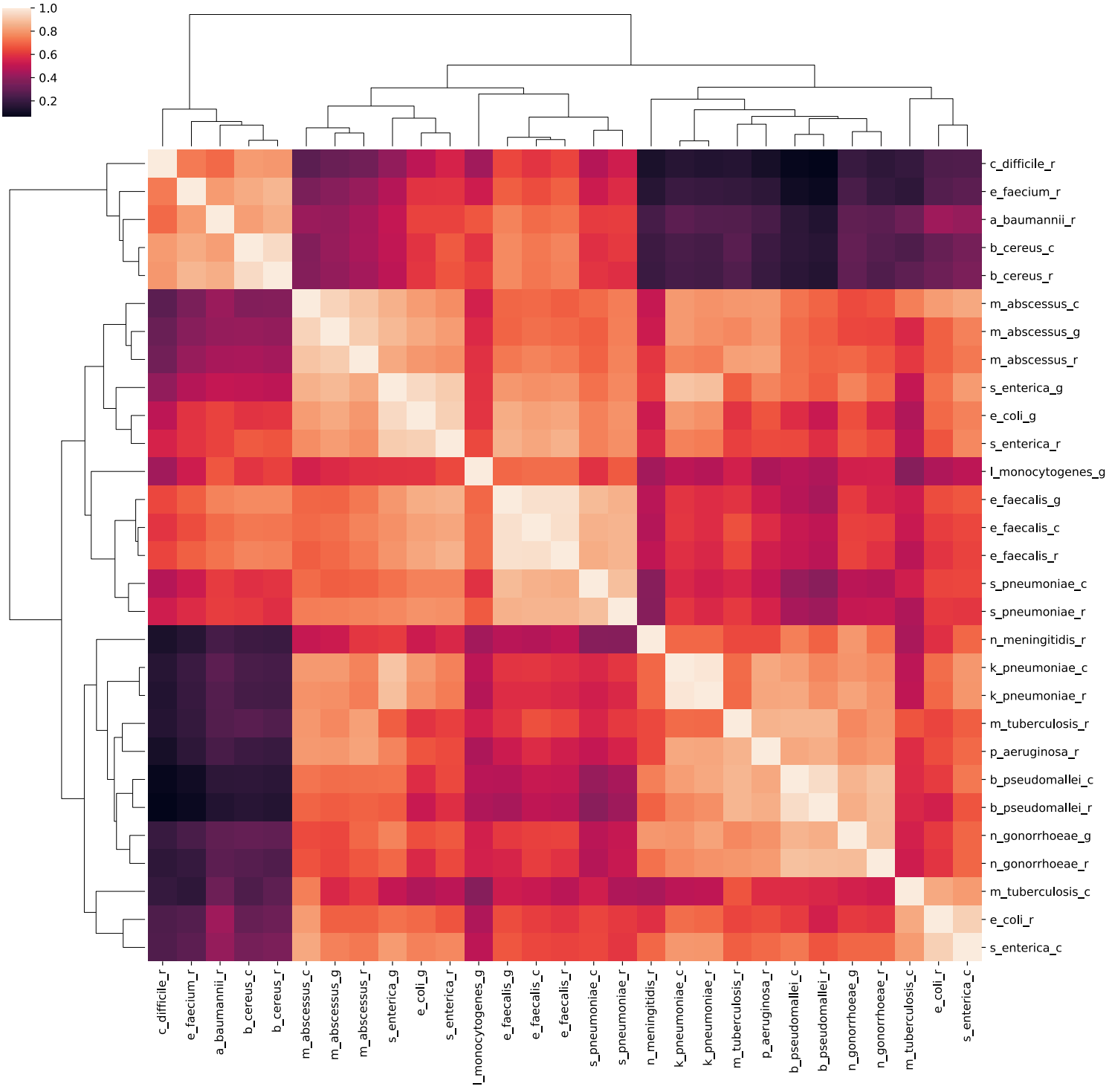


Figure 4.5: Cluster map showing how bacterial subspecies cluster by their mean mutational fingerprint.

4.4.7 Comparison between mutational fingerprints and position in phylogenetic tree

In section 4.4.6 I show that bacterial subspecies with similar GC content have similar mutational fingerprints. I therefore hypothesised that this could mean that the bacterial subspecies with similar mutational fingerprints are closest in the evolutionary tree. In 2004 Santos et al. identified 25 conserved motifs from 10 conserved proteins : fusA, gyrB, ileS, lepA, leuS, pyrG, recA, recG, rplB, rpoB which could be used to identify bacterial phylogeny [232]. I used these motifs to determine the phylogeny between the different sub-species identified. The phylogeny is shown in figure 4.6 below.

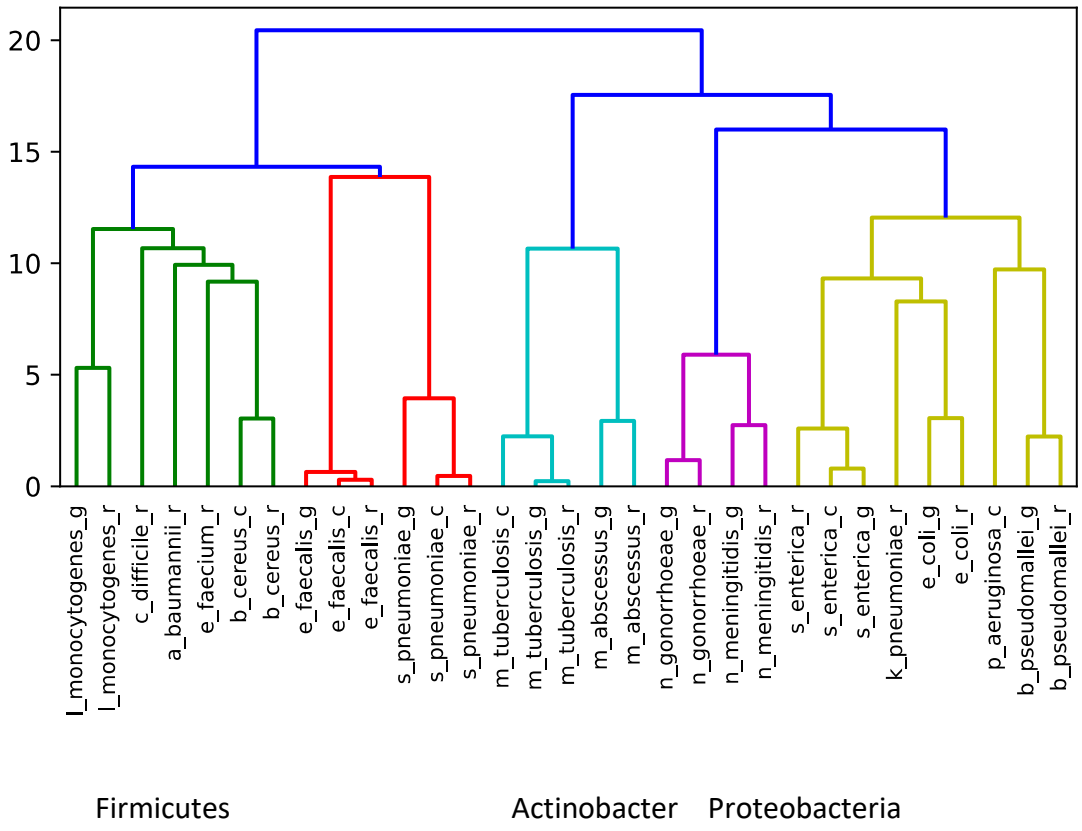


Fig 4.6: The bacterial phylogeny derived using 25 conserved motifs identified by Santos et al [232].

The mutational fingerprints of the firmicutes fall into groups 1 and 3. That is they tend to have fewer T>C mutations. Subspecies from the phylla actinobacter and proteobacteria have mutational fingerprints which fall into groups 2, 4, and 5, and have comparatively more T>C mutations. However, beyond these categories the link between phylogeny and mutational fingerprints breaks down. There is some association between mutational fingerprints and GC content, but no further clear association with genome.

4.4.8 Decomposition of mutational fingerprints into mutational signatures

The mean mutational fingerprints provide an aggregated picture of the mutational stresses on a particular bacterial sub-species. To understand further the causes that could be contributing to this picture, I decomposed the mutational fingerprints for each bacterial sub-species into five mutational signatures (giving 145 signatures in total and the corresponding weights for each of the bacterial strains). I then clustered these signatures together to identify signatures that were seen in more than one bacterial subspecies and identified the median signature for each of 40 groups.

These mutational signatures are shown below in figure 4.7, and the heatmap showing in which bacterial subspecies they occur is given in figure 4.8. The clustering diagram is shown in appendix 2 as supplementary figure 4.4. The first 24 median signatures shown formed the dominant signature in samples from at least 3 bacterial subspecies. More details about these more common signatures is given in supplementary figures 4.5.

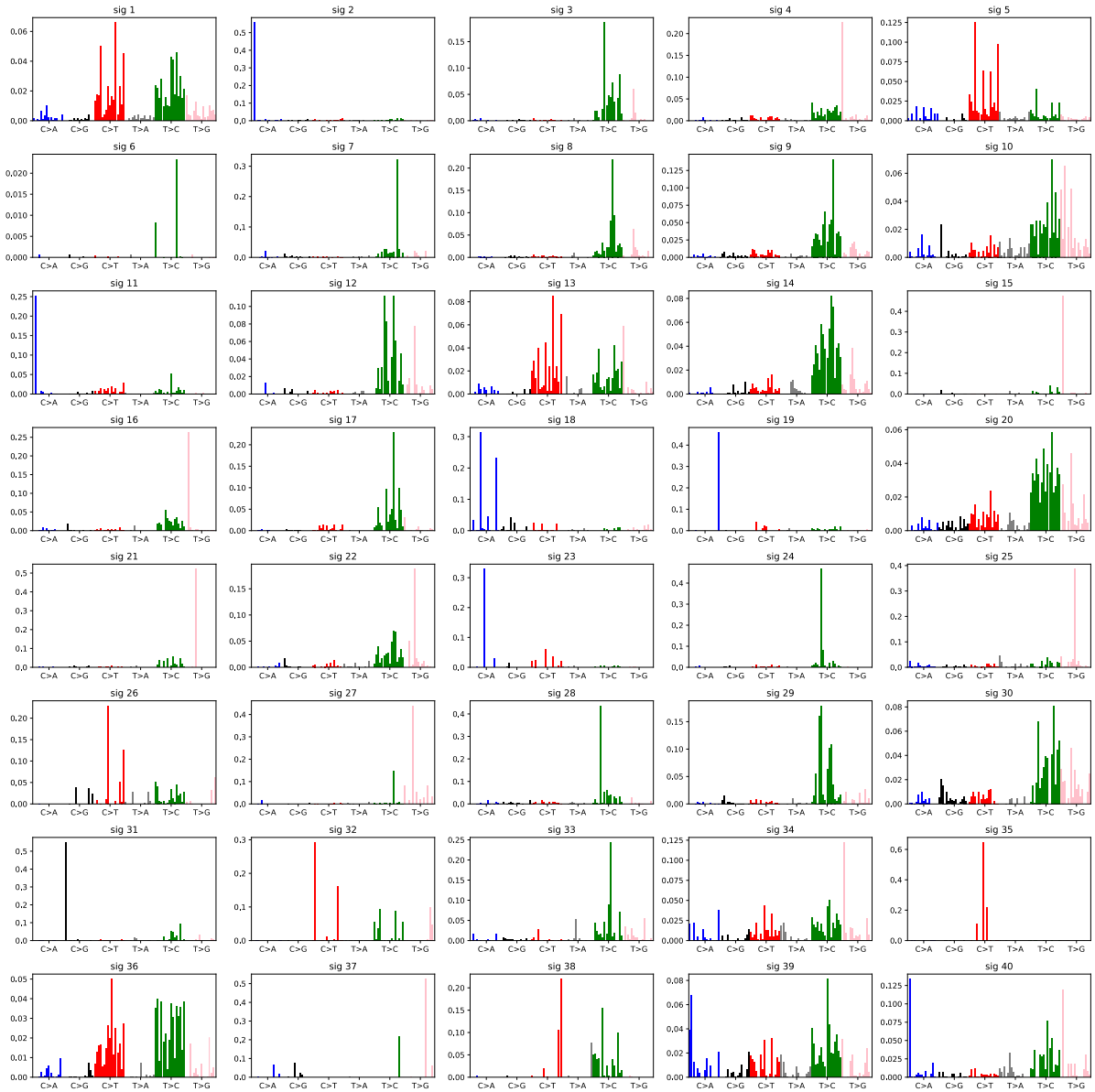


Figure 4.7: Mutational fingerprints from the 29 bacterial strain clusters were decomposed into five mutational signatures each (giving 145 signatures). The signatures were then grouped into 40 similar signatures. The median of these clustered signatures is shown above.

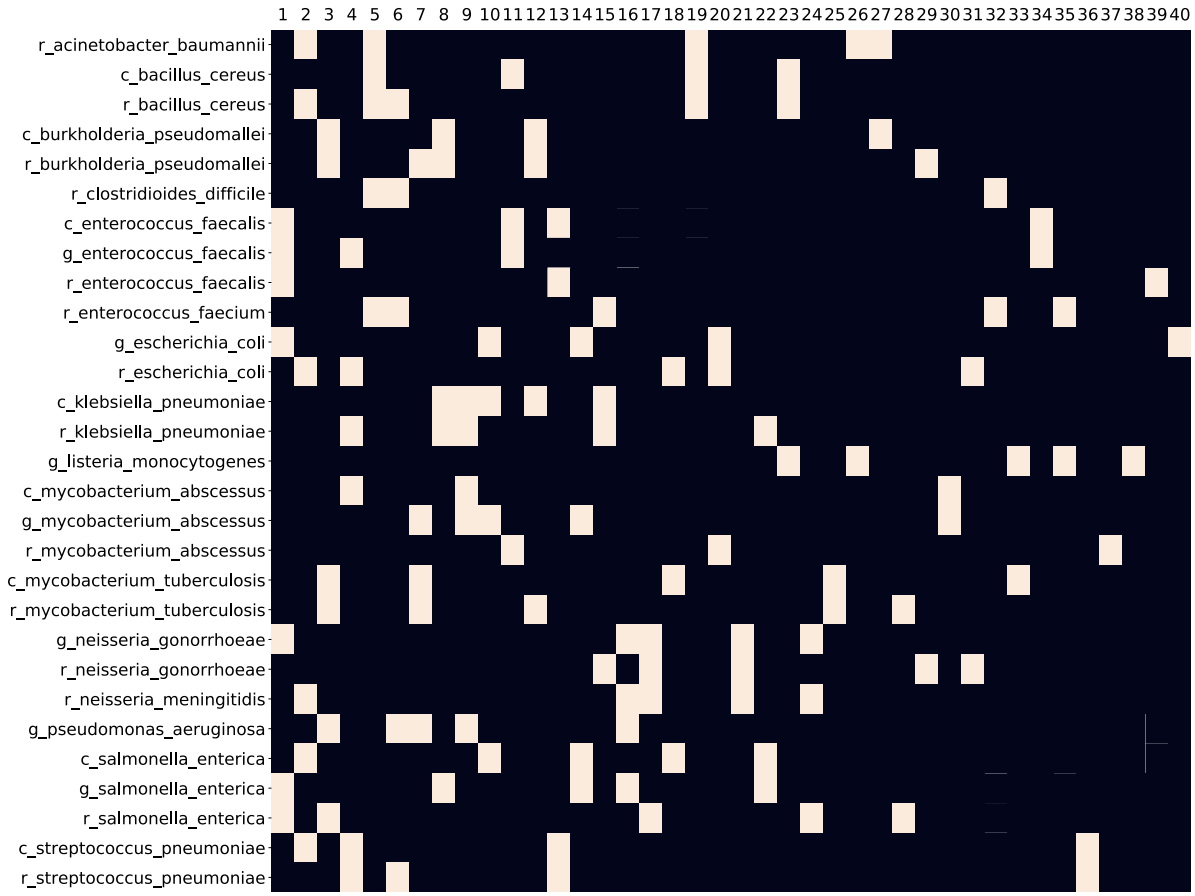


Figure 4.8: The 145 signatures cluster to provide 40 signatures. Here I show the breakdown of these signatures into the subspecies.

4.4.9 Potential aetiology of bacterial signatures

4.4.9.1 Comparison with cancer signatures

At present little is known about the aetiology of different mutational signatures in bacteria. However, considerable work has been done to improve understanding of mutational signatures in human cancer tissues. The COSMIC signatures version 2 [101], [102], [152], [167], [173] shown in figure 4.9 below are based on all somatic base mutations in the exome. The signatures published by COSMIC are all for a single species; *Homo sapiens*, and

include both silent substitutions and missense mutations, because the majority of cancer mutations are passenger mutations having little impact on the ability of the cell to survive and reproduce.

To enable an appropriate comparison, I normalise these signatures by counting the human trinucleotide distribution in the exome and converting this into a vector representing the distribution of mutations if each nucleotide substitution had the same probability. I then divide the raw signatures by this background count.

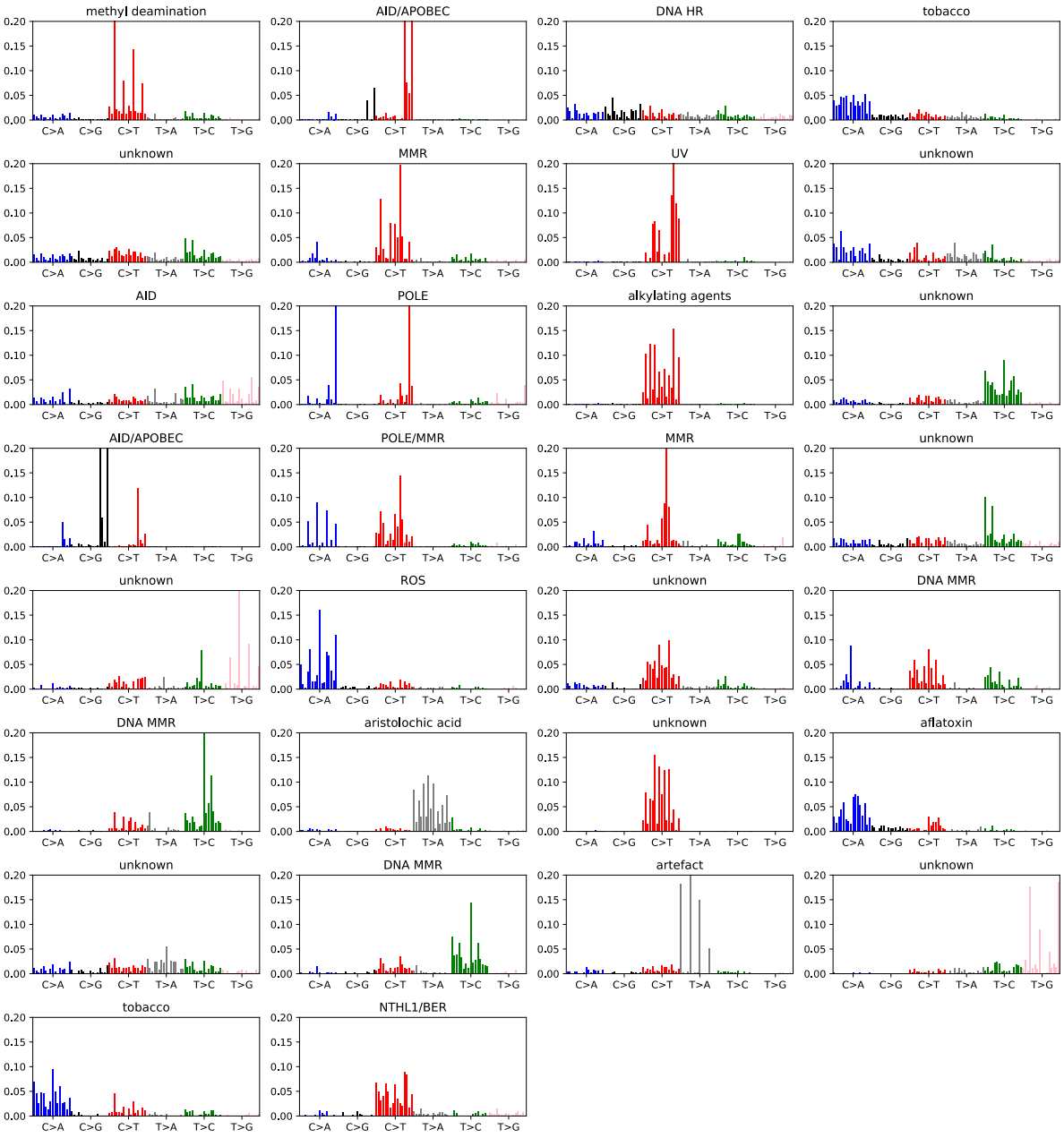


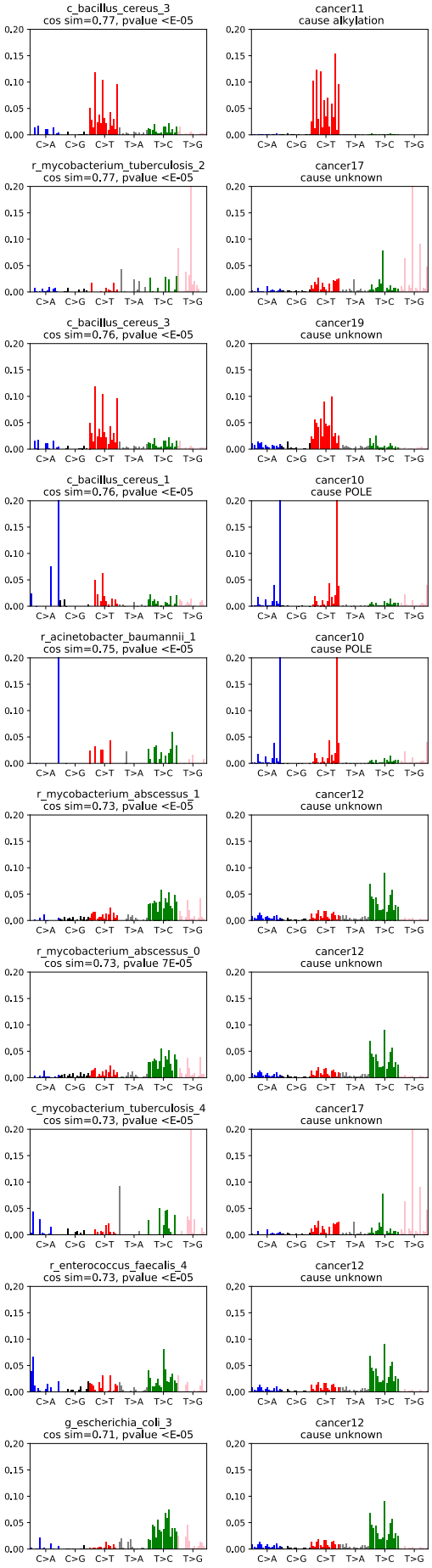
Figure 4.9 legend: Shows the COSMIC signatures version 2 normalised for the composition of the human exome. Version 2 of the COSMIC signatures is used because they are based on somatic mutation counts in the exome.

To compare the bacteria and cancer signatures I then calculated the cosine similarity between them. To calculate the significance of these similarities I generated ten thousand random mutational signatures and found the cosine similarity of each when compared to each of the 30 bacterial mutational signatures. The highest cosine similarity score seen ranged between 0.29 and 0.65 depending on the signature.

Fifty-three pairs of cancer/bacteria signature pairs have a higher cosine similarity score than expected with a p-value <0.05 after correcting for multiple testing. The highest cosine similarity score between a cancer signature and a bacterial signature was 0.79, with a p-value < 1e-5.

The ten pairs shown in figure 4.10 have the highest cosine similarity between bacterial signatures and those found in cancer. These include signatures relating to failure of DNA MMR, action by alkylating agents and the error prone polymerase POLE. All pairs showing significant cosine similarity are shown in the appendix 2 in figures 4.6a-f.

Figure 4.10: The ten pairs of bacterial and cancer signatures showing a high cosine similarity.



4.4.9.2 Comparison with signatures of known mutagens

Mutational signatures have also previously been experimentally ascertained for environmental agents acting on human pluripotent stem cells [197]. These fingerprints are based on the entire human genome so to enable an appropriate comparison, I counted the human trinucleotide distribution across the whole genome and convert this into a vector representing the distribution of mutations if each nucleotide substitution had the same probability. I then divided the raw signatures by this background count. The resulting signatures are shown in figure 4.11 below. A description of each environmental mutagen is given in table 4.2.

Interestingly, the signatures of alkylating agents MNU and ENU are quite distinct from COSMIC signature 11 which has also been attributed to the impact of derived alkylating agents. Diverse cellular repair pathways give rise to different patterns of damage from alkylation in different cells which may go some way to explaining this phenomena [241].

As before, I calculated the cosine similarity between the environmental agent signatures and the bacteria signatures and found the statistical significance using a permutation test. The closest match is between the alkylating agent DMH and a common signature found in enterococcus faecalis with a cosine similarity score of 0.86. A further 40 out of the 145 bacterial signatures were more similar to one or more mutagenic signature than would be expected by chance, after correcting for multiple testing. In particular, eleven of the sixteen bacteria have a signature similar to at least one of the alkylating agents that preferentially cause T>C transitions (Temozolomide, MNU). Figure 4.12 shows the most similar bacterial

signature to each of the seventeen environmental agents, where these are statistically significant.

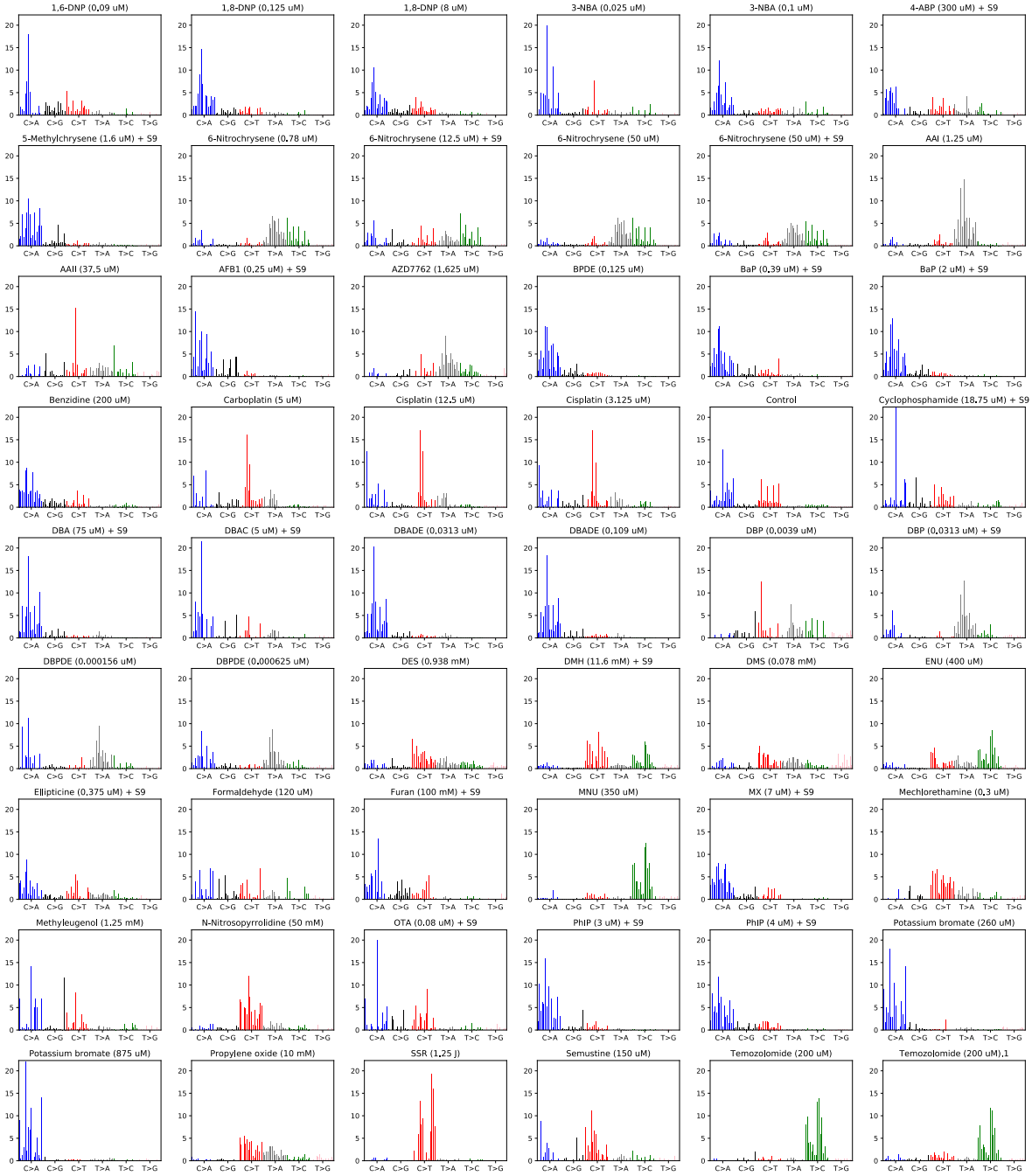
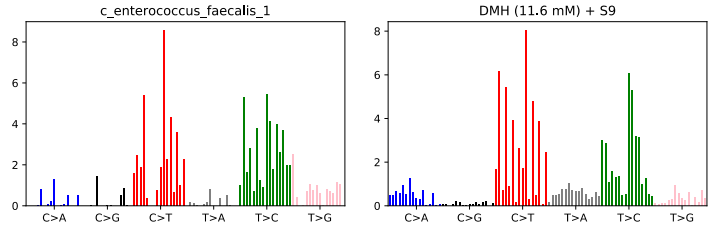


Figure 11: Normalised signatures for the environmental mutagens tested by Kucab et al.[197]. The signatures were normalised by the genomic distribution of trinucleotides in homo sapiens to enable direct comparison with bacterial signatures.

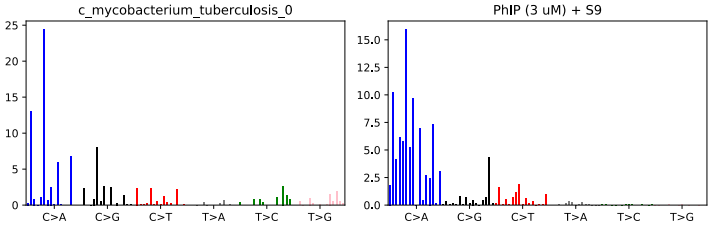
Mutagenic agent	Description
DMH	Powerful DNA alkylating agent and carcinogen
PhIP	Food mutagen - heterocyclic amine found in cooked meat.
ENU	Alkylating agent and carcinogen.
Propylene Oxide	Alkylating agent and expected human carcinogen used primarily as a chemical intermediate in the production of polyethers and propylene glycol.
MNU	Alkylating agent and carcinogen. No known commercial use.
Temozolomide	Cytotoxic alkylating agent, with chemotherapeutic uses.
DES	Alkylating agent expected to be human carcinogen. Used as commercial ethylating agent in organic synthesis.
Mechlorethamine	Metabolised to reactive ethylene immonium derivative, which alkylates DNA and inhibits DNA replication, with chemotherapeutic uses.
AAI	Carcinogen used in some folk medicines.
DMS	Alkylating agent and an immunosuppressive agent, expected to be human carcinogen. Used as methylating agent in organic synthesis.
MX	Strong bacterial mutagen. Disinfectant by-product[242].
N-Nitrosopyrrolidine	Food and tobacco mutagen. Expected human carcinogen.
BPDE	Carcinogen/mutagen. Component of smoke
Ellipticine	Mutagen, antineoplastic agent and a plant metabolite[243].

Benzidine	Highly toxic carcinogen, intermediate in chemical synthesis.

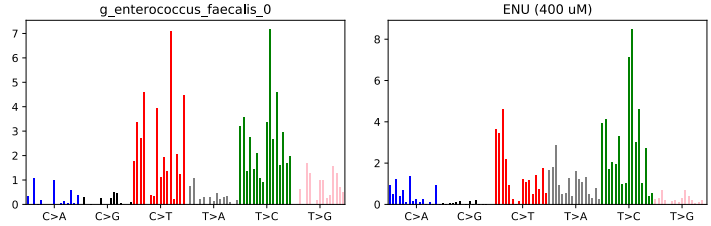
Table 4.2 : Mutagens used by Kucab et al.[197] to derive mutagenic signatures [244].



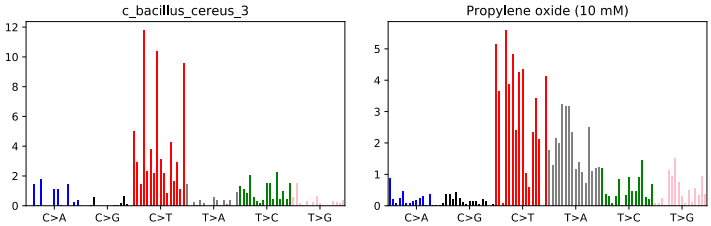
Similar signatures to DMH (11.6mH)+S9 occur in *E. faecalis*, *N. gonorrhoeae, B. cereus, S. pneumoniae, and S.enterica.* The closest match is c_enterococcus_faecalis_1 with cosine value = 0.86 adjusted pvalue <1.00E-04



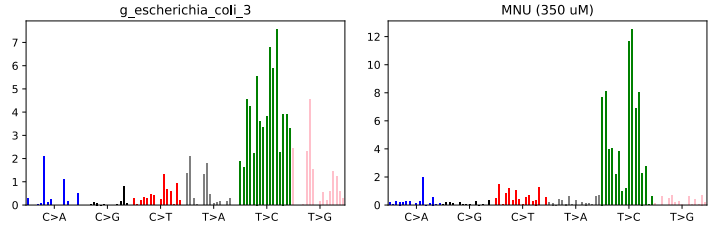
Similar signatures to PhIP (3uM)+S9 occur in *S. enterica, E. coli, M. tuberculosis*. The closest match is c_mycobacterium_tuberculosis_0 with cosine value = 0.79 adjusted pvalue <1.00E-04



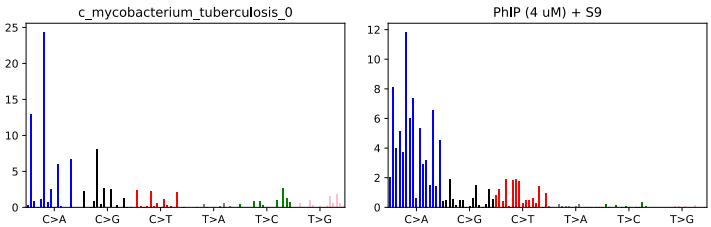
Similar signatures to ENU (400uM)+S9 occur in *L. monocytogenes, E. faecalis, M. abscessus, E.coli and S.enterica.* The closest match is g_enterococcus_faecalis_0 with cosine value = 0.77 adjusted pvalue <1.00E-04



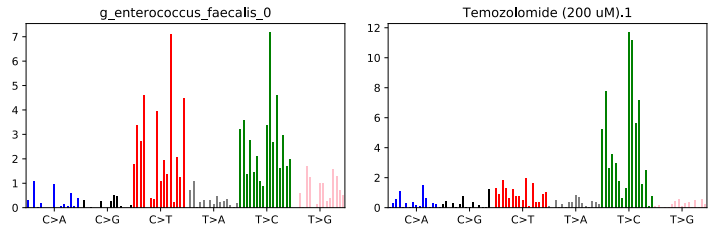
Similar signatures to Propylene Oxide (10mM) occur in *C.difficile, B.cereus, A. baumannii, and E. faecium.* The closest match is c_bacillus_cereus_3 with cosine value = 0.77 adjusted pvalue <1.00E-04



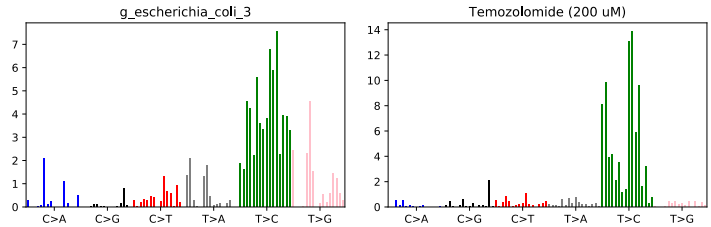
Similar signatures to MNU (350 uM) occur in *M.tuberculosis, L.monocytogenes, N.gonorrhoeae, K. pneumoniae, E.coli, P.aeruginosa, S.enterica and B.pseudomallei.* The closest match is g_escherichia_coli_3 with cosine value = 0.76 adjusted pvalue<1.00E-04



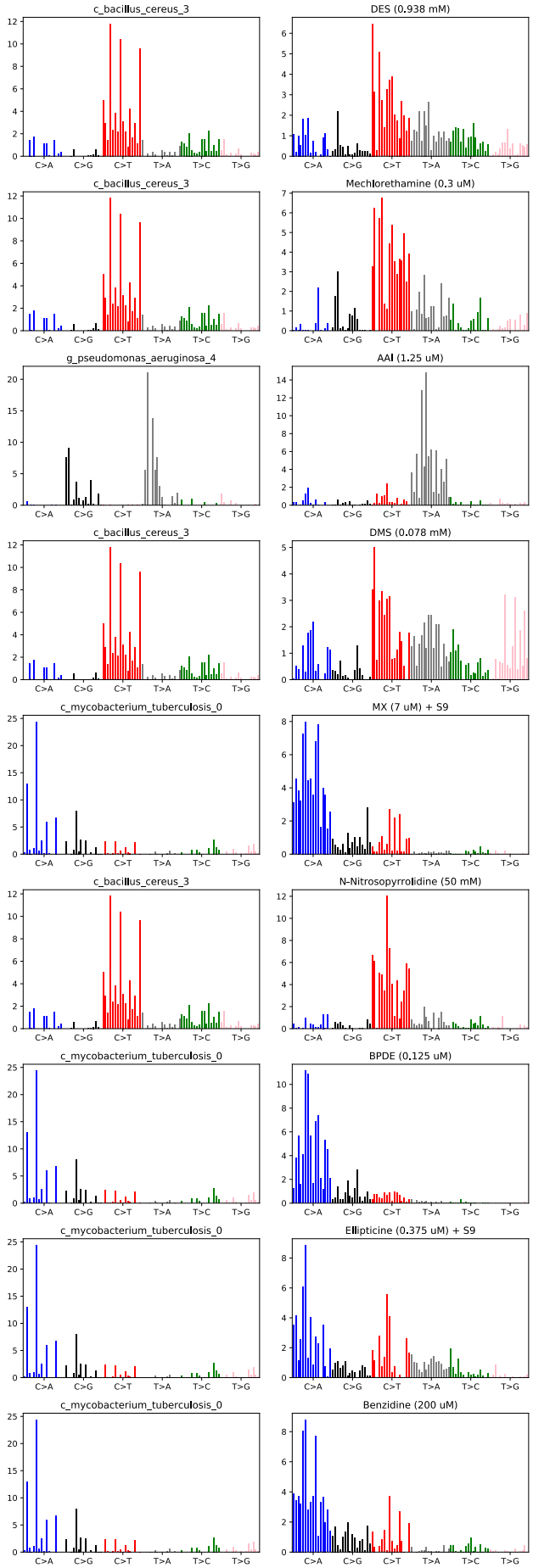
Similar signatures to PhIP (4uM) +S9 occur in *M.tuberculosis, E.coli, and S.enterica.* The closest match is c_mycobacterium_tuberculosis_0 with cosine value = 0.75 adjusted pvalue<1.00E-04



Similar signatures to Temozolomide (200uM).1 occur in *M.tuberculosis, L.monocytogenes, E.faecalis, M.abscessus, N.gonorrhoeae, K.pneumoniae, P aeruginosa and S.enterica.* The closest match is g_enterococcus_faecalis_0 with cosine value = 0.74 adjusted pvalue<1.00E-04



Similar signatures to Temozolomide (200uM) occur in *M.tuberculosis, L.monocytogenes, E.faecalis, M.abscessus, K.pneumoniae, E.coli, S.pneumoniae, P aeruginosa and S.enterica.* The closest match is g_escherichia_coli_3 with cosine value = 0.74 adjusted pvalue<1.00E-04



Similar signatures to DES (0.938mM) occur in *B.cereus.* The closest match is g_bacillus_cereus_3 with cosine value = 0.73 adjusted pvalue<1.00E-04

Similar signatures to Mechlorethamine (0.3uM) occur in *B. cereus, C.difficile, E.faecalis, and E.faecium.* The closest match is c_bacillus_cereus_3 with cosine value = 0.73 adjusted pvalue<1.00E-04

Similar signatures to AAI (1.25 uM) occur in *P. aeruginosa.* The closest match is g_pseudomonas_aeruginosa_4 with cosine value = 0.64 adjusted pvalue<1.00E-04

Similar signatures to DMS (0.078mM) occur in *B.cereus.* The closest match is c_bacillus_cereus_3 with cosine value = 0.60 adjusted pvalue<1.00E-04

Similar signatures to MX (7uM) +S9 occur in *M.tuberculosis.* The closest match is c_mycobacterium_tuberculosis_0 with cosine value = 0.57 adjusted pvalue<1.00E-04

Similar signatures to N-Nitrosopyrrolidine (50mM) occur in *B.cereus.* The closest match is c_bacillus_cereus_3 with cosine value = 0.71 adjusted pvalue< 1.01E-02

Similar signatures to BPDE (0.125uM) occur in *M.tuberculosis.* The closest match is c_mycobacterium_tuberculosis_0 with cosine value = 0.63 adjusted pvalue<2.20E-02

Similar signatures to Ellipticine (0.375uM) occur in *M.tuberculosis.* The closest match is c_mycobacterium_tuberculosis_0 with cosine value = 0.62 adjusted pvalue<3.54E-02

Similar signatures to Benzidine (200uM) occur in *M.tuberculosis.* The closest match is c_mycobacterium_tuberculosis_0 with cosine value = 0.61 adjusted pvalue<3.60E-02

Figure 4.12: Plots show comparisons between seventeen bacterial signatures against the normalised signatures caused by damage from environmental agents on human pluripotent stem cells. In each case there is a statistically significant cosine similarity between the two signatures.

4.4.9.3 Conclusion from comparisons

Finally, I now return to the clustered bacterial signatures identified in figures 4.7 and 4.8 and ask whether there is sufficient evidence to propose an aetiology for these signatures. The analysis above sheds light on the potential aetiology of 11 of the 40 grouped signatures. These split into four categories.

Five bacterial signatures (1, 14, 20, 36 and 39) were similar to the cancer signature 26 which is associated with deficient MMR. The relatively high cosine similarity between cancer signature 26 and the alkylation signatures of MNU (0.81), Temozolomide (0.80/0.82), ENU (0.76), DMH(0.63) means that many of these signatures were statistically similar to both cancer signature 26 and the signatures of the alkylation agents MNU and Temozolomide and often ENU and DMH as well. These signatures are found in *Enterococcus faecalis*, *Escherichia coli*, *Mycobacterium abscessus*, *Neisseria gonorrhoeae*, *Salmonella enterica*, and *Streptococcus pneumoniae*. Other bacterial signatures were not similar to the MMR cancer signature but were similar to some of the alkylation agents with a similar signature. These were: bacterial signature 9 (MNU and Temozolomide) which is found in *Klebsiella pneumoniae*, *Mycobacterium abscessus*, and *Pseudomonas aeruginosa*; bacterial signature 13 (DMH) which is found in *Enterococcus faecalis* and *Streptococcus pneumoniae*, and bacterial signature 33 (ENU,MNU, Temozolomide) which is found in *Listeria*

monocytogenes and *Mycobacterium tuberculosis*. One bacterial signature (signature 5) which is found in *Acinetobacter baumannii*, *Bacillus cereus*, *Clostridioides difficile* and *Enterococcus faecium* was similar to the cancer signature for alkylation (cancer signature 11) and also similar to the signature for alkylating agents (DES, mechlorethamine, propylene oxide). These are very different from the MMR signature having a strong C>T component. Bacterial signature 18 which is found in *Escherichia coli*, *Mycobacterium tuberculosis* and *Salmonella enterica* was similar to the cancer signature associated with use of PhIP, a heterocyclic amine. Finally, bacterial signature 19 which is found in *Acinetobacter baumannii* and *Bacillus cereus* was similar to the cancer signature associated with use of the error prone polymerase POLE. These results are shown in table 4.3 below:

Bacterial Signature	**Similar cancer signatures with known aetiology**	**Cos similarity**	**pvalue**	**Similar mutagenic signatures**	**Cos similarity**	**pvalue**
Signature 1	Signature 26 (MMR)	0.70	<e-4	DMH ENU MNU Temozolomide,	0.80 0.71 0.66 0.65/0.71	<e-4 <e-4 <e-4 <e-4
Signature 5	Signature 11(alkylation)	0.71	<e-4	DES, Mechlorethamine, Propylene oxide	0.65 0.66 0.67	1e-4 <e-4 <e-4

Signature 9				MNU Temozolomide	0.67 0.67/0.68	<e-4 3e-4/<e-4
Signature 13				DMH	0.65	1e-4
Signature 14	Signature 26 (MMR)	0.61	3e-4			
Signature 18				PhIP	0.67 0.7	2e-4 <e-4
Signature 19	Signature 10 (POLE)	0.75	4.2e-3			
Signature 20	Signature 26 (MMR)	0.66	<e-4	MNU, Temozolomide	0.69 0.68/0.70	<e-4 <e-4
Signature 33				ENU, MNU, Temozolomide	0.65 0.69 0.70/0.70	<e-4 <e-4 <e-4
Signature 36	Signature 26 (MMR)	0.73	<e-4	ENU,	0.70 0.69	<e-4 <e-4

				MNU, Temozolomide, DMH	0.73/0.69 0.76	<e-4 <e-4
Signature 39	Signature 26 (MMR)	0.69	<e-4	ENU, MNU, Temozolomide	0.70 0.67 0.70/0.66	<e-4 <e-4 <e-4

Table 4.3: Comparison between bacterial signatures, cancer signatures of known aetiology and signatures arising from environmental mutagens, showing cosine similarity and p-values where these are statistically significant. All p-values derive from a permutation test on 10,000 shuffled signatures. Where none of these random signatures provided as high a cosine similarity the pvalue is given as <e-4.

4.5 Conclusion and Discussion

The rate of bacterial substitutions is sufficient to leave a mutational fingerprint. By looking at just those substitutions that are unique and silent it is possible to build up a history of mutations in each strain which can hint at the underlying causes of DNA damage. On average the patterns of mutations reflect the CG content of the bacteria: bacterial species with low CG content having higher rates of C>T than those with higher GC content.

Often there is little variation between the mean mutational fingerprint of subspecies within the same species (for example *Bacillus cereus*). However, a richer picture emerges, with

some bacterial subspecies (for example in *Salmonella enterica*) having strikingly diverging patterns of mutations.

When human signatures are compared the criteria for similarity is often very stringent. For example, the cosine similarity between the COSMIC cancer mutational signature 1 in versions 2 and 3 is 0.95. However, here I am comparing signatures between different species (bacteria, human) with different DDR mechanisms, and significant methodological differences in the way in which the signatures are derived (exome, genome, unique silent substitutions). By 'similar' I mean that the cosine similarity is higher than for the signatures randomly permuted. Using this method, a cosine similarity between two signatures of 0.7 would normally be highly statistically significant.

Using this approach, the majority of the bacterial signatures do not have any similar human mutational signature. However, eleven of the signatures are statistically significant to those found either in cancer cases or as a result of the action of environmental mutagens on pluripotent human stem cells. Most of these similarities are to the signatures from alkylating agents. These examples that are C>T rich (DES and cancer signature 11) , T>C rich (MNU) and with both C>T and T>C peaks (DMH). However, the comparatively high cosine similarities between some of the alkylating signatures and cancer signature 26 (which is considered to be the result of deficient MMR), mean that deficient MMR could also play a role in the creation of bacterial signatures. In addition, I also found signatures that are similar to cancer signature 10 which arises from use of the error prone polymerase POLE, and from the environmental mutagen PhIP which is a heterocyclic amine.

5 Using mutual exclusivity to identify therapeutically actionable synthetically lethal gene pairs

5.1 Introduction

Recent estimates suggest that it takes approximately 13 years and a 'capitalized' cost of approximately US$1.8 billion to bring a new drug to the market [245]. Reducing costs and amount of time required for each of the different steps in the drug discovery pipeline is the key to deliver better drugs to patients in a timely manner. One approach that has been utilised to increase the efficiency of the drug discovery process involves drug repositioning, whereby treatments for one disease are exploited in treating another. This approach has been successfully employed in the development of personalised medicines for a variety of different types of cancers[246].

Personalised cancer therapies offer an opportunity for more effective and less harsh cancer treatments[247]. Many of the personalised therapies licensed so far involved the development of oncoprotein inhibitors, for example the use of Vemurafenib to treat melanomas carrying the BRAF V600E mutation, use of Gefitinib or Erlotinib to treat breast cancers with over-expression of the EGFR family, and the use of Imatinib to treat chronic myeloid leukaemia [248][249][44]. However, not all cancers are treatable in this way and acquired resistance remains a problem [250]. An alternative strategy, exemplified by the use of PARP inhibitors in the treatment of BRCA deficient tumours [251] is to selectively kill those cells which have lost the function of specific tumour suppressor genes.

PARP inhibitors exploit the existence of genetic interactions within a cell. A genetic interaction is when a genetic inactivation or activation in one gene in a cell can then be accentuated or attenuated by inactivation or activation in a second gene [252][106].

Synthetic lethality is an example of a genetic interaction. This is when pairs of genes can be found such that a cell that can survive the loss of protein product from either one of the genes, but not the loss of protein product from both. These interactions can be used as the basis for selectively killing cells that have inactivated tumour suppressor genes [253].
To illustrate this in more detail, let me assume that I have identified a synthetically lethal pair of genes where the first gene, is the tumour suppressor gene that I wish to target, and the second gene is a gene whose protein products are druggable, that is they can be inhibited using a known drug. The tumour suppressor gene is inactivated predominantly within the body of the tumour. When the protein products of the synthetically lethal partner are inhibited, non-tumorous cells will generally suffer the loss of protein products from just the druggable gene and survive. By contrast, cells within the tumour will not have the protein products of either of the synthetic lethal gene pair and should therefore be selectively killed [254]. Within each cancer cell numerous genetic interactions exist, and some of these may be exploited therapeutically.

A number of experimental approaches have been developed to identify genetic interactions. The majority of experiments have been undertaken in model organisms such as *Saccharomyces cerevisiae* or *Drosophila melanogaster* [255][256][257][258]. However, genetic interactions tend not to be highly conserved. For instance, *S. cerevisiae* and *S.pombe*

share only around 30% of their genetic interactions, suggesting that these model organisms may not be the most effective way of predicting synthetic lethality in humans[259].

Experimentally verified synthetic lethal gene pairs based around specific target genes are now beginning to emerge for humans based on CRISPR-CAS9 knockouts [260][261][262][263]. Even these do not provide a gold standard. Cell lines differ from in vivo cells both genetically and epigenetically, but also because they lack an appropriate tumour microenvironment [264].
An alternative approach is to predict synthetic lethality computationally.

Synthetically lethal gene pairs have been successfully predicted using conserved patterns in protein interaction networks [265][266][267]. Although genetic interactions are not reliably conserved between species, a number of research teams have managed to use orthological and evolutionary data to infer synthetically lethal interactions in humans from those from model organisms [268][269][270][271], whilst other teams have integrated information from GO terms to compare the functionality of genes involved in genetic interactions [272][273]. An alternative approach is to analyse patterns of alterations in cancer cell lines and samples to identify mutually exclusive interactions .

Different definitions of mutual exclusivity exist. A “hard” definition would be to say that two events are mutually exclusive if they never co-occur. However, in practice synthetic lethality is never this clear cut, so a correspondingly softer definition of mutual exclusivity can be used. Given the number of samples with an alteration in either gene A or gene B, the number with alterations in both gene A and gene B is lower than expected i.e.

$Prob\ (A\ given\ B) < Prob\ (A)$ or equivalently $Prob\ (B\ given\ A) < Prob\ (B)$.

Similarly, co-occurring events are taken to be ones where given the number of samples with an alteration in either gene A or gene B, the number with alterations in both gene A and gene B is higher than expected. i.e.

$Prob\ (A\ given\ B) > Prob\ (A)$ or equivalently $(B\ given\ A) > Prob\ (B)$.

Where a pair of genes have a synthetically lethal interaction, the alterations that have the effect of preventing effective protein production would occur in a way which is mutually exclusive. A therapeutically actionable gene pair is where gene A is a known tumour suppressor, and gene B gives rise to protein products that can be inhibited with a small molecule. If the small molecule is a licensed drug or a compound in late stage clinical trial this may point the way to potentially drug repositioning.

By comparison, where a gene pair co-occurs it suggests that they many both belong to a genetic pathway with built-in redundancy. In some such cases more than one hit may be needed to provide optimal selectivity for the tumour, or co-operative resistance to chemotherapies [274][275].

A number of teams have looked for mutually exclusive interactions using different statistical tests and different data sets, of which a subset is shown in table 6.1.

Team	Statistical test
Srihari et al. [276]	The team used the hypergeometric test to identify mutually exclusive gene pairs in order to predict synthetically lethal pairs, confirming some experimentally. They used The Cancer Genome Atlas (TCGA) copy-number and gene-expression datasets to identify gene copy-number amplifications and deletions, and gene up- and downregulation in four cancers. An assumption was made that the changes in gene expression would act as a proxy for mutations and for epigenetic changes.
Canisius et al. [277]	The Discover algorithm was developed to compare the hypergeometric test and Poisson binomial test on simulated data sets. The tests were used on both simulated data and TCGA copy number and somatic mutation data for 118 genes.
Leiserson et al. [278]	Leiserson et al. developed the CoMEt algorithm to look for multiple combinations of mutually exclusive alterations, using a Markov chain Monte-Carlo algorithm combined with the hypergeometric test. They used simulated data as well as somatic mutation data including substitutions, indels, gene-fusions, rearrangements and aberrant gene splicing as well as copy number variance where this accorded with gene expression data from TCGA.

Babur et al. [279]	Babur et al. used detailed prior pathway information together with an extension of the hypergeometric test to look for groups of mutually exclusive genes that have a common downstream target in the network. They used TCGA mutation and copy number variance data where this accords with gene expression data.
Ciriello et al. [280]	The team developed the MEMo algorithm in order to look for networks of oncogenic modules, looking for groups of genes that are frequently altered, that are in the same biological pathway; and that have mutually exclusive alteration events, using the hypergeometric test. The algorithm was tested on TCGA mutation and copy number data, restricting the genes to those where gene expression accorded with copy number.
Bradley et al. [281]	The team looked at whether observed levels of overlap were consistent with complete mutual exclusivity based on known rates of misclassification using a binomial test.

Table 6.1: Approaches used to identify mutual exclusivity

Several approaches that have previously identified mutually exclusive relationships have done so in order to identify genes, whose protein products may be in the same biological pathway [279] [280]. The theory is that inactivating two genes in the same pathway or contributing to the same functional process will generally not confer a significant selective advantage compared to the single inactivation in that pathway and so alterations will tend to be mutually exclusive. The goal here is different. I specifically want to find mutual exclusive behaviours that cannot be explained because the associated proteins are in the same pathway, as these are more likely to reflect synthetically lethal behaviours.

This raises three interesting questions. The first question is, what base case should be used for mutual exclusivity? That is, how many inactivations should be expected in a pair of genes if there is no co-occurrence or mutual exclusion between them? The second question is what type of experimental evidence is reliable measure for loss of function of the protein product? And finally, is the evidence sufficient to predict mutual exclusive pairs?

In this chapter, I have identified mutually exclusive and co-occurring patterns of alterations in pairs of genes in cancer samples in seven different cancer types. Research teams have previously used mutual exclusivity to predict synthetic lethality. My approach is novel in that: In order to identify therapeutically exploitable interactions which may lead to repositioning of known drugs for new cancers I focused on identifying mutually exclusive patterns of inactivation in gene pairs where one of the pairs is a tumour suppressor, and the other a gene where known drugs inactivate the corresponding protein product. I also extend the approaches previously used by integrating data on the somatic mutations, copy

number variations and methylation status of the gene. Calculations were run using these data both independently and then in combination.

Initially, I used the hypergeometric test to identify significantly mutually exclusive and co-occurring gene pairs. This is an industry standard test and has been implemented by several groups on a variety of other different datasets [276]–[280]. However, next, I modified the approach pioneered by Canisius et al. [277] to implement the Poisson binomial test to test for significant mutually exclusive and co-occurring gene pairs in a novel way that improves the reliability of the results and sensitivity to stratification of the samples.

Finally, I used the STRING network clusters were used to identify which gene pairs were in the same or closely connected protein pathways. In order to display the results, I designed and implemented a website MexDrugs

https://users.sussex.ac.uk/~skw24/mexdrugs1/index.html

where users can browse and search for potential therapeutic gene pairs.

5.2 Methods

5.2.1 Druggable genes.

In this analysis I needed a set of proteins with licensed inhibitors. To obtain this, the database of druggable genes was downloaded from the Druggable genome databank DGIdb on 2nd July 2020 and filtered for those genes associated with proteins that can be inactivated [148].

5.2.2 Cancer data.

Seven cancer types were analysed; breast, kidney, large intestine, liver, lung, ovary and prostate. A list of tumour suppressor genes was identified from the Cancer Gene Census [88]. I have named these as Tumour Suppressor Associated genes (TSa), because although they have been identified as a Tumour Suppressor in at least one sort of cancer, they may not be identified as such in the cancer being analysed.

I used The Cancer Genome Atlas (TCGA) data https://www.cancer.gov/tcga [282] to identify genomic alteration data because it is data rich, providing data on mutations, RNA seq expression, copy number variants (CNV) and methylation data. Somatic mutation and CNV data were downloaded from COSMIC, and methylation data were downloaded from the Genomic Data Commons [150][147]. The list of methylation probes where methylation is negatively correlated with gene activation was downloaded from the Broad Institute TCGA Genome Data Analysis Center [146].

None of these 'omic types provide a straightforward proxy for permanently altered protein expression and processing of the data was required.

5.2.3 Identifying inactivated genes

Genes may not be fully functional for a number of different reasons and separate data types were processed individually.

5.2.3.1 Missense mutations

The most frequent mutations observed in cancer are missense mutations, capable of preventing the formation of active protein, activating a protein but more often entirely benign. I use the precomputed FATHMM value included within the COSMIC database to predict whether or not a specific mutation would be pathogenic [201].

The other main problem in using any type of cancer mutation data is that it often provides information about only one of the sister chromatids. Tumour suppressor genes may provide sufficient active protein from one functional chromatid. In practice mutations in TS genes within cancer cells are often mirrored on the other chromatid due to loss of heterozygosity. The working assumption throughout this is that one mutation is generally sufficient, but that missense mutations dubbed as pathogenic in the TCGA database should be assumed to lead to loss of function only if either the gene in question is a known tumour suppressor, or has been predicted as a tumour suppressing kinase.

5.2.3.2 Frameshift mutations

Frameshift indels not only lead to the alteration of all amino acids past the point of mutation followed by premature truncation but they also give rise to nonsense mediated decay (NMD) so that in many cases no protein is produced. However, only 4% or so of all mutations are indels. For both tumour suppressor genes and druggable genes I treated all frameshifting mutations and whole gene deletions as inactivating.

5.2.4 CNV data

TCGA Data is also available for copy number variants (CNV). It is well known that copy number variants are not an effective proxy for protein abundance, because of possibility of gene dosage mitigating aneuploidy. Nevertheless, total absence of the gene will give rise to absence of protein, so the data is can be informative for loss of function alterations.

5.2.5 RNA seq data

It would be possible to use RNA seq expression data. However, gene expression data is a snap shot of the cells in time. Although I can expect the expression levels of essential genes to remain roughly constant, and thus interpret any reduction in RNA seq expression as detrimental, essential genes are of little interest to me precisely because they are essential and thus are likely to cause severe side-effects when inhibited. The second caveat is that gene expression is not well correlated to protein expression for all genes. As a result, I did not use RNA seq data within my analysis.

5.2.6 Methylation data

I considered that the use of methylation data is more informative than RNA-seq data, as cancer- associated genes include a number of epigenetic regulators, and once hypermethylated, the changes are stable, being inherited across cellular generations [283]. TCGA Firehose data exists identifying those methylation probes that are negatively correlated with RNA seq expression data, effectively identifying those probes which turn off gene expression in this more permanent fashion.

For methylation data I identified the pairs of methylation probes/genes that were negatively correlated with a pvalue <0.05. GDC data provided beta distribution values for sample/methylation probes. Methylation data is provided as β-values, that is the proportion of CpG dinucleotides that are methylated. The values of β have been shown to be bi-modal with peaks between 0-0.2 for hypomethylated sites and between 0.8-1 for hypermethylated sites[284]. I therefore use a cut-off of 0.8 to identify hypermethylation at the site of probes leading to gene inactivation .

5.2.7 Combined data types

I identified genes that were differentially in-activated within samples of the relevant cancer type in 4% of samples. That is, genes which are inactivated in less than 4% of samples or more than 96% of samples are removed from the set, as these would be unlikely to provide statistically significant findings.

I carried out analyses using somatic mutations, CNV data and methylation data independently, as well as using all the data together. In each case, I constructed an alteration matrix with 1s for all the sample/gene pairs where the gene is inactivated and 0s otherwise. Separately I identified for each type of data, how many genes were inactivated in each sample.

5.2.8 Statistical Tests Used

In order to distinguish mutually exclusive and co-occurring inactivations from independent inactivations it is necessary first to establish the most appropriate statistical distribution to use as the base case. In this study two different statistical frameworks were used, the hypergeometric and the Poisson binomial.

5.2.8.1 Hypergeometric distribution

If inactivations in *gene A* and *gene B* form independent and identically distributed random variables then I would expect the number of inactivations to follow a hypergeometric distribution. This distribution is most easily described as drawing coloured shapes from a bag without replacement. If the bag has S shapes of which n_1 are red and n_2 are square

then if the two features red and square are independent, then if I draw x red squares, the probability that $x = k$ is given by the hypergeometric distribution:

$$P(x = k) = \frac{\binom{n_1}{k}\binom{S - n_1}{n_2 - k}}{\binom{S}{n_2}}$$

If the shapes are analogous to samples, then this looks like a good candidate for the base case for assessing mutual exclusivity, and it has been used for that purpose by Leiserson et al.

and Babur et al. [278] [279]. One particularly useful feature of the hypergeometric distribution is that it is computationally tractable. In this analysis, I used hyp from python's scipy.stats library to calculate the hypergeometric probability for each gene pair.

5.2.8.2 Poisson binomial distribution

In cancer it is known that even within the same cancer the distributions of genetic alterations differ widely in different samples [78][285] In particular, Canisius et al. [277] showed whilst most genes are more likely to be mutated in samples that have many mutations, this is not true for many tumour suppressor genes. They proposed that the probability of a gene being mutated was a function of the overall number of genes mutated, and that the base case should be a Poisson binomial distribution rather than the hypergeometric distribution. The Poisson binomial distribution describes the probability of k successful outcomes (in this case a sample with *gene A* and *gene B* inactivated) from S

independent trials (the samples) where the probability of each trial being successful varies. This probability is given by:

$$P(x = k) = \sum_{A \in F_k} \prod_{i \in A} p_i \prod_{j \in A^C} (1 - p_j)$$

Here F_k is the set of all subsets of k integers that can be selected from $\{1,2, \ldots . s\}$ and A^c is the complement of A.

The disadvantages of using the Poisson binomial are that the individual probabilities p_i are not so straight forward to calculate and the distribution is not computationally tractable in its raw form. I took the following approach.

Let $g_1, g_2 \ldots g_k$ be k genes, $s_1, s_2 \ldots s_n$ be n samples. Using the Poisson binomial distribution, there are two steps to calculating the probability of mutually exclusive or co-occurring pairs. Firstly, I estimate $\overrightarrow{P_i}$; the probability of inactivation of gene g_i in all samples. I then calculate the Poisson binomial mass function, and use this to find the probability that I would get as extreme a number of joint inactivations as I observe by chance, before correcting for multiple testing.

Step 1: estimate the probability of each gene being inactivated in each sample.

Canisius et al. [277] do this by solving a constrained optimization problem to estimate the probability of each gene being inactivated in each sample. However, I found that many hundreds of the tumour suppressor associated (TSa) genes and druggable genes have variable methylation status and are therefore worthy of consideration. As a result, this

approach is too slow. I therefore take the simpler approach of estimating the probabilities as piecewise linear functions over approximately ten bins.

Let $t_1, t_2 \ldots t_N$ be the total number of inactivated genes of each sample. Further, let t_{max} be the maximum number of inactivations. For each gene g_i I assume that the probability of that the gene is inactivated in s_j is purely a function of t_j, i.e. $P_{ij} = f_i(t_j)$. I assume that these functions are continuous, that is that samples with similar total numbers of inactivated genes should have similar probabilities of each gene being inactivated. I then estimate these functions, f_i , as a piecewise, linear average of the number of inactivations of g_i in each of roughly 10 bins. I firstly separate the samples into 10 clusters on the basis of similar numbers of total gene inactivations. However, the distribution of the total inactivations t_i typically has 1-3 peaks (see Figure 6.1a)), and I consider it likely that the samples in these groups are biologically distinct, so I ensure that the clusters do not mix samples from different groups. For each cluster of samples, I then find the mean probability of inactivation for each gene, and the mean total inactivations. For each group, there will be two samples with the lowest and highest mean total inactivations. The probability for each gene inactivations for these samples is assumed to be the same as for the closest cluster. For each gene the piecewise linear functions f_i is then constructed from these fixed points. I assume that $\overrightarrow{P_i}$ are independent for all i , so that the probability of gene g_{i_1} and g_{i_2} being inactivated in the same sample is given by $\overrightarrow{P_{i_1 i_2}} = \overrightarrow{P_{i_1}}\, x\, \overrightarrow{P_{i_2}}$.

Thus $P_{i_1 i_2}(j) = f_{i_1}(t_j).f_{i_2}(t_j)$. I noticed that some of the $f(t)$ looked like bowls. To quantify the number of genes with these characteristic functions, I use the following cut-off procedure. I normalise t so that $t_{max} = 1$ and fit a quadratic curve. my cut-off is then that the quadratic portion should be <-1 (upturned bowl) or >1. To ensure that I only pick up the

functions with a full bowl shape I further require that the edges should be <2/3 the height of the maximum (for the upturned bowl). I consider that the function is strongly correlated if Pearson's correlation is >0.7 (strongly anti-correlated if Pearson's correlation was <-0.7).

Step 2: I use Hong's algorithm for calculating the probability mass function, using fast Fourier transforms[286]. I have the made the distribution available as the python package Poisson_binomial. It can be installed using *pip install Poisson_binomial*.

5.2.9 Simulating Data

In order to determine how effective, the Poisson binomial test is at identifying co-occurring and mutually exclusive gene pairs, I ran a simulation with 500 samples and 100 genes. I included within the samples a mutation distribution that roughly mimicked those seen in the data. That is, half of the samples had a total of roughly *t* = 400 mutations (i.e. the variable *t* was normally distributed, mean = 400, standard deviation = 50), and the other half had a total of roughly *t* =3000 mutations (the variable *t* was normally distributed, mean = 3000, standard deviation = 50). I designated 50 of the 100 genes as TSa genes and the other half I designated as druggable genes. The probability of a mutation in each gene was allowed to vary as a function $f(t)$ of the total number of mutations t where:

$$f(t) = \frac{Ae^{t'} + Be^{-t'}}{e^{t'} + e^{-t'}} \text{ where } t' = k(t - a)$$

These curves have a characteristic logit-shape tending to A for large t, B for small t (i.e. highly negative t'), and equalling $\frac{A+B}{2}$ when t = a. The curves resemble many of the $f(t)$

seen in our data. The parameter k was set at 0.001, whilst a was allowed to vary randomly between 300 and 700, A between 0.5 and 0.9 and B between 0 and 0.1.

I used these probability functions to generate independent alteration matrices, and then made 30 gene pairs $g_1 g_2$ that co-occurred. To do this I set the probability of joint inactivation midway between the independent probability of joint inactivation and the minimum probability of inactivation of g_1 or g_2. That is:

$$P(g_1 \,\&\, g_2) = \frac{p_1 p_2 + \min(p_1, p_2)}{2}$$

where $p_i = P(g_i inactivated)$. The probabilities $P(g_1 \,\&\, \sim g_2)$, $P(\sim g_1 \,\&\, g_2)$, and $P(\sim g_1 \,\&\, \sim g_2)$ were adjusted accordingly to hold $P(g_1)$, and $P(g_2)$ constant.
Similarly, I made a second alteration matrix where 30 gene pairs were mutually exclusive with the probability of joint inactivation 0.75 that of the independent probability of joint inactivation.

I then tested both the hypergeometric algorithm and Poisson binomial algorithm on the cooccurring and mutually exclusive gene pairs. In order to test the Poisson binomial model, I did not allow recourse to the known inactivation probabilities but instead used the inferred inactivation probabilities as I had for the genuine data.

5.2.10 Cluster Analysis

I used the protein-protein interaction clusters identified in STRING database [151] as proxy for genes whose protein products are in the same pathways. Most of these clusters are very small, but some of the STRING clusters include many thousands of genes, that are not in the same pathway. The networks were therefore filtered so that only clusters with less than 1,000 genes included in the analysis. This cut-off was modified to allow clusters with 2,000 3,000 and 4,000 to test the robustness of the cut-off point. These are very generous cut-off points and so should include the largest protein pathways.

5.3 Results

5.3.1 Results using Simulated Data

I ran a simulation with 200 alteration matrices for 500 samples with 100 genes. In 100 of these matrices I included 30 co-occurring gene pairs whilst in the other 100 I included 30 mutually exclusive gene pairs (see methods for details). I then used both the hypergeometric test and the Poisson-binomial test with inferred inactivation probabilities on all of the simulated alteration matrices to identify the genetic interactions.

For the mutually exclusive pairs g_1, g_2 I set the probability of joint inactivation to be 75% that of the probability g_1 and g_2 both being inactivated if there was no genetic inactivation. The Poisson binomial test found on average 29 of the 30 pairs (17 after correcting for multiple testing) whilst the hypergeometric test found on average 15 (10 after correcting for multiple testing).

For co-occurring pairs g_1, g_2 I set the probability of joint inactivation to be midway between that of the probability g_1 and g_2 both being inactivated if there was no genetic inactivation and the minimum inactivation probability of g_1 and g_2. The Poisson binomial test found on average 29 of these pairs (24 after correcting for multiple testing) whilst the hypergeometric test found all the pairs but incorrectly identified an additional 2184 pairs (1088 after correcting for multiple testing).

Thus, I found that with no correction for multiple testing, the Poisson binomial test found almost all mutual exclusive and co-occurring pairs, with few false positives whilst the hypergeometric test found roughly half the mutually exclusive pairs, and correctly identified only half the independent pairs in the co-occurring test. Following correction for multiple testing the sensitivity of the Poisson binomial test fell to roughly 60% for mutually exclusive pairs compared to 35% for the hypergeometric test. For co-occurring tests, inclusion of correction for multiple testing made the sensitivity of the Poisson binomial fall to 80%, whilst the specificity of the hypergeometric test improved to 70%. I concluded that the Poisson binomial test was more accurate and finds significantly more of the mutually exclusive pairs and finds significantly fewer false positive co-occurring pairs.

The sensitivity, specificity, and accuracy of the two tests is shown in table 6.2 below.

		Poisson Binomial			Hypergeometric		
		Sens.	Spec.	Accuracy	Sens.	Spec.	Accuracy
Mutual exclusivity	Benjamini-Hochberg Correction	58.1%	100%	99.5%	34.7%	100%	99.2%
	No correction for multiple testing	96.1%	99.9%	99.9%	52.5%	99.9%	99.4%
Co-occurrence	Benjamini-Hochberg Correction	80%	100%	99.8%	99.9%	69.3%	69.6%
	No correction for multiple testing	97.8%	99.9%	99.9%	100%	53.1%	53.4%

Table 6.2: Sensitivity and specificity of the Poisson binomial and hypergeometric tests in finding mutually exclusive gene pairs in simulations.

5.3.2 Data

On the basis of information from the Druggable genome databank, 1806 genes were identified as potential targets for drug inhibition [148]. I refer to these as druggable genes. I

also identified 263 tumour suppressor-associated genes (TSa genes) that gave me 2069 potential genes of interest for my analysis.

I analysed cancer associated alterations to these genes in seven different cancer types where data were available for somatic mutations, methylation and CNVs. The cancers chosen were cancers of the breast, kidney, large intestine, liver, lung, ovary and prostate, and analysed them independently. Table 6.3 shows the number of differentially altered genes identified in each cancer type.

	#TSa genes	#Druggable genes
Breast	611	121
Kidney	1419	239
Large intestine	754	145
Liver	471	96
Lung	1323	248
Ovary	150	24
Prostate	525	101

Table 6.3: the number of genes of interest (either tumour suppressor associated genes TSas or druggable genes) identified in each cancer type. These genes were inactivated in between 4 and 96% of samples.

5.3.3 Results for individual tissue types

5.3.3.1 Breast cancer

In total there were 843 breast cancer samples included in the analysis with mutation, CNV and methylation data. I initially calculated mutually exclusive or co-occurring gene pairs for the data types individually and then reran the calculations combining the data.

5.3.3.1.1 Calculations using individual data types

5.3.3.1.1.1 Mutational Data

Five TSa genes, CDH1, GATA3, KMT2C, MAP3K1 and TP53 had differential somatic mutation status none of the druggable genes did. Consequently, mutational data could not be analysed independently.

5.3.3.1.1.2 CNV Data

Nineteen genes had differential CNV status, of which 3 were TSa genes and 16 were druggable. This gave 48 gene pairs with CNV data to analyse. Using both the hypergeometric test and the Poisson binomial test on the CNV data none of the pairs of genes were found to be mutually exclusive. Using either test, the same 28 gene pairs 58% were identified as co-occurring meaning that they were more likely to be altered together. All co-occurring genes were on chromosome eight.

5.3.3.1.1.3 Methylation Data

Compared to mutation and CNV data, methylation data is a much richer source for identifying differentially inactivated genes. Methylation data alone enabled the identification of 107 TSa genes and 583 druggable genes with differential methylation status

giving rise to 62,381 gene pairs which could be potentially have mutually exclusive or co-occurring relationships.

5.3.3.1.1.3.1 Analysis of the methylation samples

The hypergeometric test assumes that the probability that a given gene is mutated is the same for each sample, whilst the Poisson binomial test allows me to take variations in these probabilities into account. I therefore analysed the methylation samples to see whether or not there is much variation in the probability of inactivation across the different samples.

Analysis of the 1,104 breast cancer methylation samples suggested that the samples split into two distinct clusters. The first comprises 531 samples each having in total less than 500 inactivations and the other cluster contains 573 samples each having more than 1400 inactivations (see figure 6.1a)).

There were no obvious distinctions between the clusters, for example in terms of days to death. However, it is possible that the distinct clusters reflect important differences in disease and that each cluster comes with its own different set of gene inactivations. In that case a gene pair may appear to have mutually exclusive inactivations because the two genes are associated with different clusters, rather than any genetic interaction. The Poisson binomial test is more sensitive to such gene pairs of this form.

I am therefore interested both in whether the two tests give broadly similar numbers of mutually exclusive pairs, and also whether or not there is much overlap between the set of gene pairs found using each algorithm.

The hypergeometric test can be used without explicitly estimating the probability of a gene being inactivated. However, in order to use the Poisson binomial test, I estimated these probabilities as a piecewise linear function of the total number of inactivations t in the sample, to give me an inactivation function, $f(t)$, for each gene (see methods for more details). The inactivation functions are normally roughly increasing i.e., the probability of a specific gene being inactivated is well correlated with the total number of inactivations in almost all genes. However, three distinct inactivation functions emerge, as shown in subplots 6.1b), 6.1c) and 6.1d). 52% of genes show a switching behaviour (figure 6.1b))– the gene is predominantly turned on in those samples with few inactivations and off in those with many. In a further 30% of genes the probability of the gene being inactivated grows steadily as the total number of inactivations grows (figure 6.1c)). Finally, in 6.1d)) typically some increase in probability is seen as the total number of inactivations grows but the relationship is less distinctive. The probability of gene inactivation is anti-correlated with the number of total inactivations in only five genes.

5.3.3.1.1.3.2 Calculations using hypermethylation data

Using methylation data alone, 352 gene pairs (0.56%) were identified as mutually exclusive using the hypergeometric test and this rose to 752 pairs (1.2%) using the Poisson binomial test. If a hypergeometric test is used, co-occurrence of genes inactivated by methylation is the default state with 50,265 of the possible 62,381 gene pairs (80.6% of possible gene pairs) being identified as co-occurring even after correcting for multiple testing. Using the Poisson binomial test, which accounts for the baseline distributions the number of co-occurring gene pairs fell dramatically to just 1,037 (1.7% of possible gene pairs)

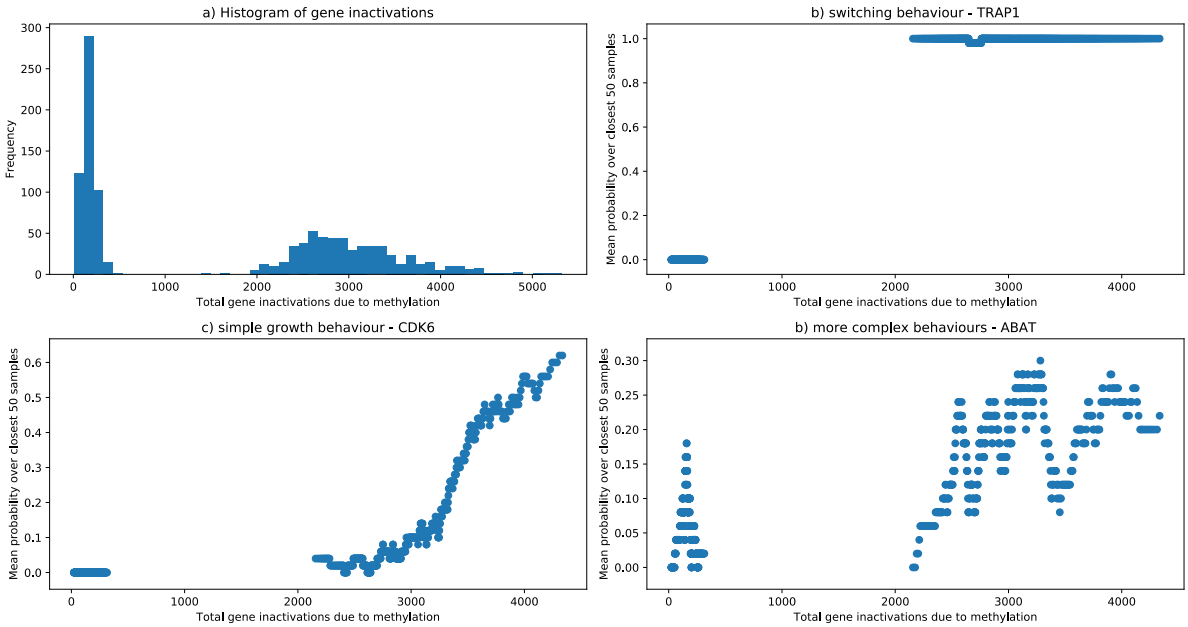


Figure 6.1: In breast cancer the number of genes inactivated by methylation is highly variable. In subplot a) I plotted a histogram of the number of methylated genes in each sample. There appear to be two distinct groups of samples, with either a few hundred inactivations or between 2000-5000. In subplot b-d I plotted the mean probability of a gene being inactivated against the total number of gene inactivations due to methylation. The mean probabilities were calculated over the closest 50 samples within the same cluster. Three types of behaviours were seen. For many genes, the gene is not inactivated in the first group, but is inactivated in the second (shown here by TRAP1 in subplot b), but for around one third the probability of inactivation grows steadily with the overall number of inactivations (shown by CDK6 in subplot c). Others show more complex behaviours (shown by ABAT in subplot d).

5.3.3.1.2 Calculations combining data types

When I included all somatic mutation, CNV and methylation data 462 gene pairs (0.63%) were identified as mutually exclusive using the hypergeometric test and this rose slightly to 492 gene pairs (0.66%) using the Poisson binomial test, shown in Figure 6.2 below. I distinguish here between pairs that share one or more clusters in the string database (shown in blue) and those that do not (shown in pink).

The hypergeometric test identified 47,762 gene pairs (64.6%) co-occurring of a possible 73,931. This fell to 854 (1.16%) using the Poisson binomial test. These data suggest that the Poisson binomial is more stringent.

I define 'jointly captured' as the percentage of pairs identified by the Poisson binomial that are also identified by the hypergeometric test and 'jointly rejected' as the percentage of pairs not identified by the Poisson binomial test that were also not identified by the hypergeometric test. Note that if the Poisson binomial test were a gold standard these would equate to the sensitivity and specificity respectively of the hypergeometric test. Then 24.7% of mutually exclusive pairs are jointly captured and 99.6% jointly rejected. For co-occurring pairs, 99.6% are jointly captured and 35.8% jointly rejected.

Using the Poisson binomial test, 452 gene pairs were identified as mutually exclusive using both methylation data on its own and all data together. Mutually exclusive pairs were found for 36 of the 121 TSa genes, using 114 of the 611 druggable genes. These include 9 hits for GATA3 and 20 hits for TP53. All the mutually exclusive pairs are shown in figure 6.2. The number of mutually exclusive or co-occurring inactivated gene pairs are also displayed in

tables 6.4 and 6.5 in the section on Pan-Cancer Analysis. The gene pairs for each inhibiting drug on our website *MexDrugs*.

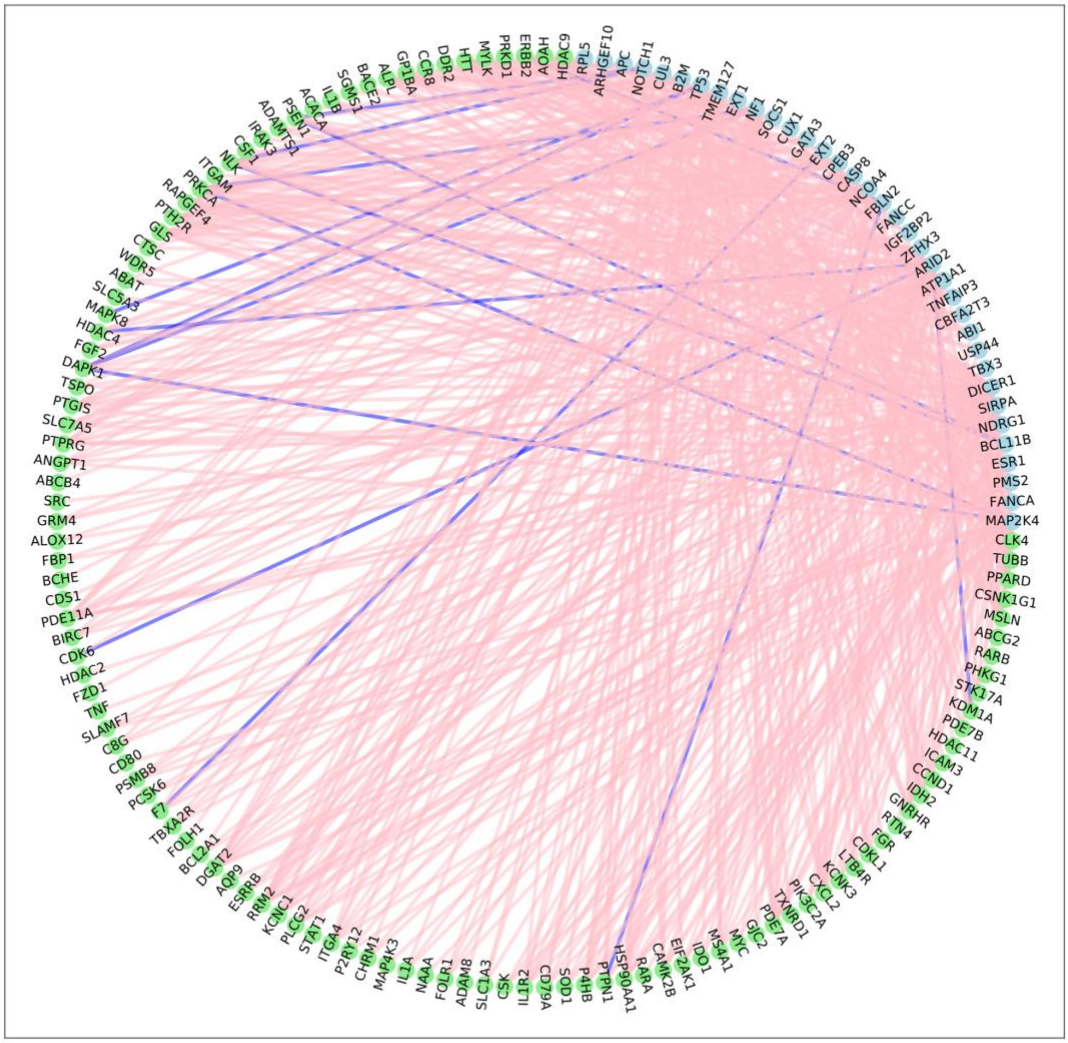


Figure 6.2: Using multi-omic data 492 gene pairs are identified as having mutually exclusive gene inactivations in samples with breast cancer. The gene pairs are shown above in blue for tumour suppressors (36) and green for druggable genes (114). In 16 pairs the genes

share one or more clusters in the string database (shown in blue) and in 472 of the pairs, TSa and druggable genes belong to different clusters (shown in pink).

5.3.3.1.3 Therapeutic Opportunities

Demonstrating mutual exclusivity does not demonstrate synthetic lethality. However, it is worth noting that only 16 (3%) of mutually exclusive gene pairs were found to belong to the same protein-protein interaction network clusters as defined in the STRING database, suggesting that most of the mutually exclusive gene pairs are found in different pathways. In addition, a number of the druggable targets which are identified as mutually exclusive partners to TSas have previously been reported to play a role in breast cancer.

For example, Trastuzumab (Herceptin) is a standard treatment for HER2 positive breast cancer either early-stage or advanced-stage/metastatic which works by inhibiting ERBB2 (HER2). There are ten mutually exclusive partners of ERBB2 (FBLN2, SIRPA, EXT2, BCL11B, ATP1A1, SOCS1, B2M, NDRG1, IGF2BP2 and NOTCH1). These genes are inactivated in 528 of 843 samples (63%).

5.3.3.2 Kidney Cancer

5.3.3.2.1 Calculations using individual data types

5.3.3.2.1.1 Mutation and CNV data

Seven TSa genes; BAP1, CSMD3, KDM5C, KMT2C, PBRM1, SETD2 and VHL, had differential somatic mutation status, but none of the druggable genes did. By way of contrast, no TSa

genes had differential CNV status, but thirteen druggable genes did. Consequently, neither mutational data and CNV could not be analysed independently.

5.3.3.2.1.2 Methylation data

Analysis of the 793 kidney samples with methylation data (66 with Kidney Chromophobe - KICH, 480 with Kidney renal clear cell carcinoma – KIRC and 247 with Kidney renal papillary cell carcinoma- KIRP) suggested that the samples split into three clusters. Around five hundred had under 400 inactivations. Around 250 had roughly one to five thousand and around forty were hyper-inactivated with roughly 13-15 thousand inactivations. These three clusters did not correspond to KICH, KIRC and KIRP classifications of kidney cancer. Whilst KICH samples all have low numbers of methylation in its samples, both KIRC and KIRP samples are split between the groups see figure 6.3.

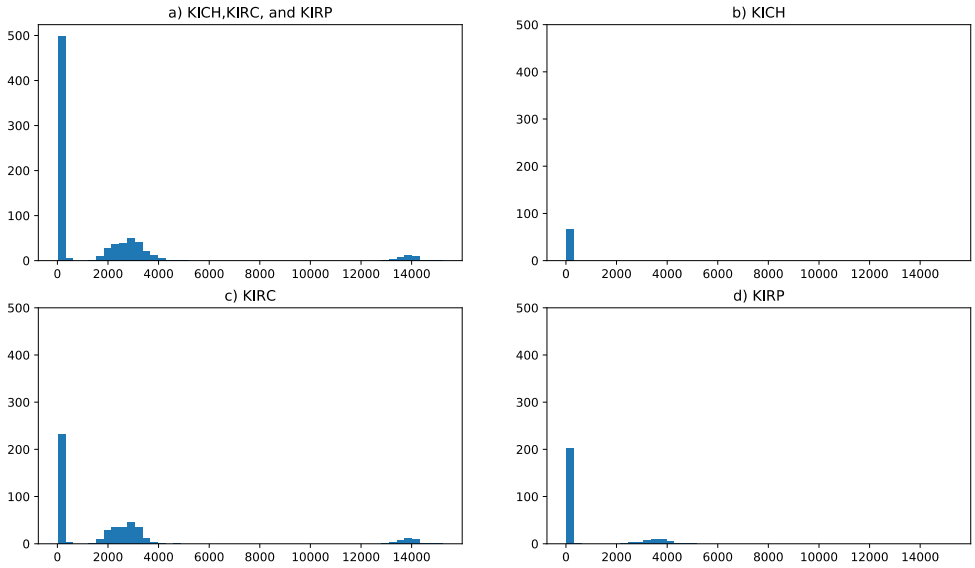


Figure 6.3: Histograms of the number of methylated genes in each sample. Subplot a) shows all kidney cancer types aggregated, whilst subplots b), c) and d) show samples from KICH, KIRC and KIRP respectively. Methylation samples from patients with kidney cancers show widely differing numbers of inactivated genes. In total 504 of 793 patients had few gene inactivations (between 11 and 371 inactivations), whilst 251 had between 1337 and 4893 and 38 were hyper-inactivated with between 12875 and 15238 inactivations. B) All KICH patients fell into the category with few inactivations, no KIRP patients were hyper-inactivated and KIRC patients fell into all three categories.

To implement the Poisson binomial test, I estimated these probabilities as a piecewise linear function of the total number of inactivations t in the sample, to give me an inactivation function, $f(t)$, for each gene. For 35% of genes, $f(t)$ was strongly correlated to t. However, other common patterns also emerged. In particular: for 11% of the genes $f(t)$ was strongly anti-correlated to t; for 9% of genes, $f(t)$ was shaped like an upturned bowl for the central group of samples; and in 3% genes $f(t)$ was shaped like a bowl for the central group of samples. See figure 6.4 for representative examples.

TSa genes are slightly over-represented in the group of genes where $f(t)$ was strongly anti-correlated to t (p = 0.01 Fisher exact test). Genes falling into this category include: ARHGEF10L, BCL10, BRCA1, LARP4B, MEN1, PIK3R1, POT1, PTK6, SDHD, SETD2, STK11, TCF3, USP44, WIF1 and ZBTB16.

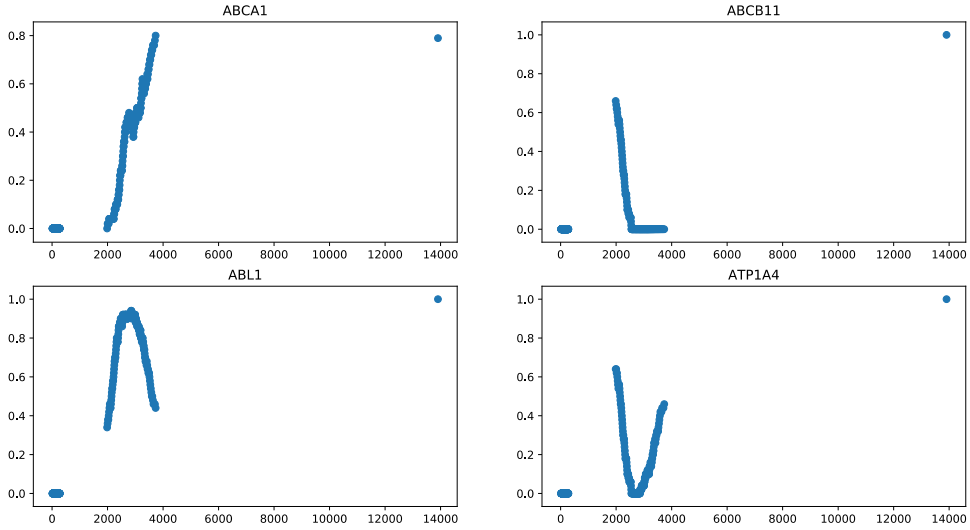


Figure 6.4: I plotted the average probability of a gene being methylated against the total number of genes methylated in the sample. Average probabilities were calculated across the 50 closest samples respecting the three clusters. The average probability of a gene being inactivated varies both as a function of the gene and of the total number of inactivations. Four common patterns emerge, and examples of these patterns are shown here. The most common pattern, illustrated by ABCA1 is that there is a zero probability of inactivation in patients with few inactivations, the probability then rises steeply for the group of samples with between 1337 and 4893 inactivations and is close to 1 for those samples that are hyper-inactivated. For some 9% of genes including ABCB11 shown here, the mid-group of samples shows a clear anti-correlation between the number of total inactivations and the probability of inactivation. I also found two bowl-shaped patterns. The most common was the inverted bowl shown here by ABL1, though V shapes also turned up, shown here by ATP1A4.

5.3.3.2.1.2.1 Calculations using hypermethylation data

228 TSa genes and 1373 druggable gene had differential methylation status giving a possible 313,044 gene pairs. 691 gene pairs (0.2% of possible gene pairs) were identified as mutually exclusive using the hypergeometric test and this rose to 2,453 gene pairs (0.8% of possible gene pairs) using the Poisson binomial test. If a hypergeometric test is used, co-occurrence of genes inactivated by methylation is the default state with 279,727 of the possible 313,044 gene pairs (89.4% of possible gene pairs) being identified as co-occurring even after correcting for multiple testing. Using the Poisson binomial test, which accounts for the baseline distributions the number of co-occurring gene pairs fell to 180,681 (57.7% of possible gene pairs).

5.3.3.2.2 Calculations combining data types

When I included all somatic mutation, CNV and methylation data 532 gene pairs (0.15%) were identified as mutually exclusive using the hypergeometric test. Using the Poisson binomial test this rose to 3,014 gene pairs (0.89%). Mutually exclusive pairs were found for 61 of the 250 TSa genes, using 239 of the 1419 druggable genes. These include one hit for VHL (PSMD9) and 94 hits for BAP1. Results are shown in Figure 6.5 below. I distinguish here between pairs that share the one of more PPI network cluster in the STRING database (shown in blue) and those that do not (shown in pink).

The hypergeometric test identified 270,134 gene pairs (79.7%) co-occurring of a possible 339,141. This fell to 5,230 (1.5%) using the Poisson binomial test.

In total, 10.4.7% of mutually exclusive pairs are jointly captured and 99.9% jointly rejected. For co-occurring pairs, 99.9% are jointly captured and 20.7% jointly rejected.

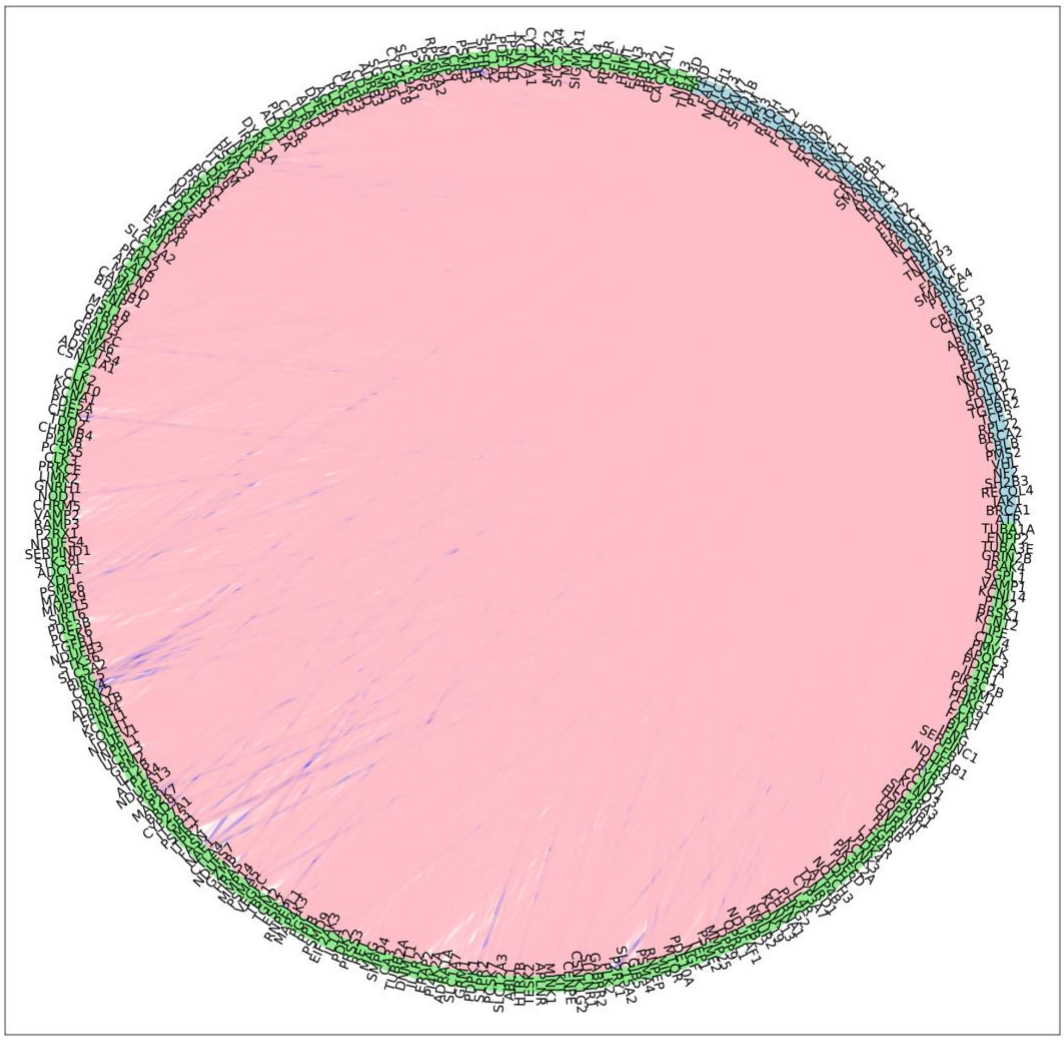


Figure 6.5: Using multi-omic data 3014 gene pairs are identified as having mutually exclusive gene inactivations in samples with kidney cancer. The gene pairs are shown above in blue for tumour suppressors (61) and green for druggable genes (250). In 2895 of the pairs, TSa

and druggable genes belong to different PPI network clusters (shown in pink). In 119 pairs the genes share one or more string cluster (shown in blue).

5.3.3.2.3 Therapeutic Opportunities

Renal cell carcinomas are often chemotherapy-resistant tumours, so whilst mutually exclusivity can occur for a number of reasons as well as synthetic lethality the large number of mutually exclusive gene pairs is encouraging. Of the pairs identified here, it is worth noting that a number of the druggable genes are inhibited by drugs which are known to have links with cancer. A recent study [287] indicates that Genistein induces cell apoptosis and inhibits cell proliferation of kidney cancer cells. Genistein is an inhibitor for ABL1, BIRC5, ESRRA, IL1R1, RET and TGFB1, which together form part of 86 mutually exclusive pairs.

The isoflavonoid me-344, mitochondrial inhibitor is in clinical trials OXPHOS complex 1 PMC4706149[288]. Me-344 has been shown to have anti-cancer properties, activating cell death pathways. It is an inhibitor for NDUFA13, NDUFAB1, NDUFAF1, NDUFB3, NDUFS4, NDUFS7 and NDUFV1, which together form part of 140 mutually exclusive pairs [289]. The list of partnered TSa genes for this and other drugs can be found in MexDrugs.

A wide range of potential uses for the senolytic, Dasatinib have been put forward, including anti-cancer treatments. Dasatinib is an inhibitor for ABL1, BLK, FGR, and KIT, which together form part of 97 mutually exclusive pairs [290].

The drug Cabazitaxel is being investigated for its role in renal cell carcinoma chemotherapy. It inhibits TUBA1A, TUBA3D, TUBA3E,and TUBB2A which together form part of 74 mutually

exclusive pairs in kidney cancer[291]. In addition, analysis of clinical trial suggests that Atezolizumab plus Bevacizumab can increase longevity by a mean of 3 months for patients with untreated metastatic renal cell carcinoma and sarcomatoid features. Bevacizumab is an inhibitor for HRAS, IDH1, KIT and VEGFC, which together form part of 56 mutually exclusive pairs [292].

5.3.3.3 Cancer of the large intestine

In total there were 488 cancer of the large intestine samples with mutation, CNV and methylation data.

5.3.3.3.1 Calculations using individual data types

5.3.3.3.1.1 Mutation and CNV data

Seven TSa genes; APC, FBXW7, KMT2D, ROBO2, RUNX1T1, SMAD4, and TP53 had differential somatic mutation status, but none of the druggable genes did so the mutational data could not be analysed independently. Four genes had differential CNV status, of which MMP26, RHD were TSa genes and FHIT, SMAD4 were druggable. This gave 4 gene pairs with CNV data to analyse. Using both the hypergeometric test and the Poisson binomial test on the CNV data none of the pairs of genes were found to be mutually exclusive. Using either test, SMAD4 and RHD were identified as co-occurring meaning that they were more likely to be altered together. Surprisingly, SMAD4 and RHD are found on different chromosomes (SMAD4 chromosome is found on 18 and RHD is found on chromosome 1).

5.3.3.3.1.2 Methylation data

Analysis of the 601 samples with methylation data suggested that they split into three distinct clusters. The first comprises just over 400 samples with less than about 400 inactivations, the second cluster roughly is 150 samples with between roughly 1000 and 4500 inactivations and roughly 50 samples with between 11,000 and 14,500 inactivations (see figure 6.6a).

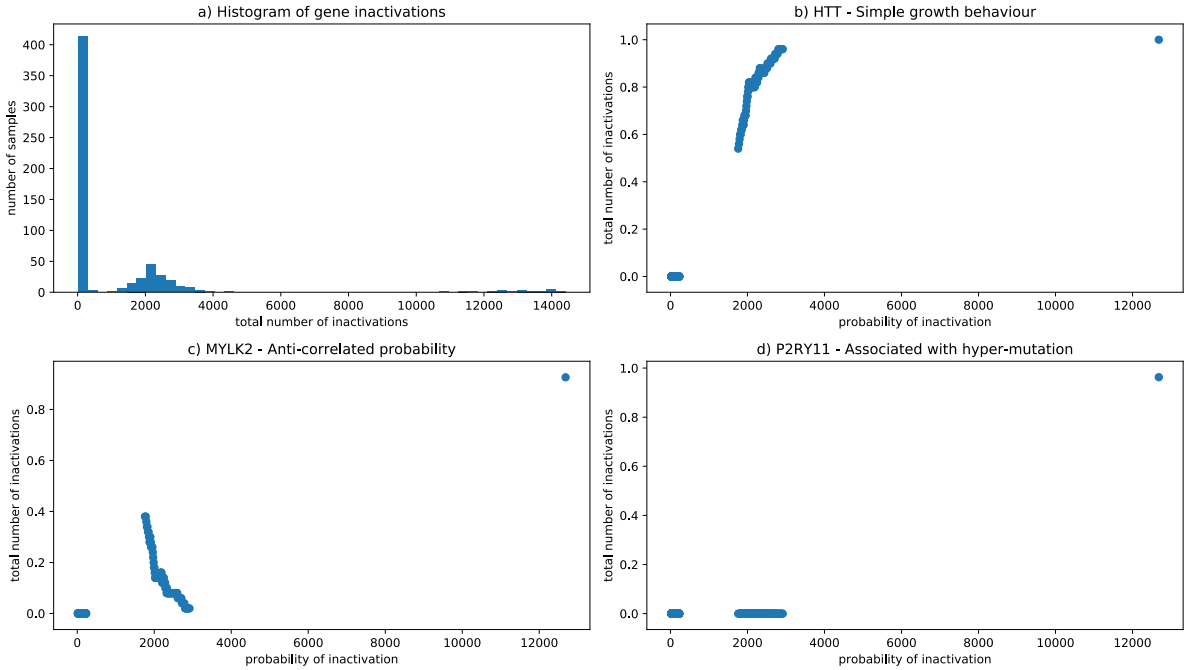


Figure 6.6: In cancer of the large intestine the number of genes inactivated by methylation is highly variable. In subplot a) I plotted a histogram of the number of methylated genes in each sample. There appear to be three distinct groups of samples, with 416 samples having less than 365, 158 samples having between 1,016 and 3,866 inactivations and 25 samples having between 11,340 and 14,411 inactivations. In subplot b-d I plotted the mean probability of a gene being inactivated against the total number of gene inactivations due to methylation. Three types of behaviours were seen. In many genes exemplified here by HTT in subplot b, the gene is not inactivated in the first cluster, probability grows in the second

cluster and the gene is normally inactivated in the third cluster. In a minority exemplified here by in MYLK2 in subplot c) for the middle cluster the probability of inactivation decreases with the overall number of inactivations . A switching behaviour is also seen in around 30% of genes shown here by P2RY11 in subplot d).

For 39% of genes the inactivation function, $f(t)$, was strongly correlated to t (see HTT, figure 6.7b)). However, other common patterns also emerged. In particular: for 15% of the genes $f(t)$ was strongly anti-correlated to t (see MYLK2, fig 6.6c)) ; and for 30% of genes, $f(t)$ was 0 except for the hyper-activated samples (see P2RY11 figure 6.6d)). The pronounced bowl shapes seen in kidney cancer were absent.

5.3.3.3.1.2.1 Calculations using hypermethylation data

There were 191 TSa genes and 1150 druggable genes with differential methylation status. 90 gene pairs of the possible 219650 gene pairs (0.04%) were identified as mutually exclusive using the hypergeometric test and this rose to 3,007 pairs (1%) using the Poisson binomial test. If a hypergeometric test is used, co-occurrence of genes inactivated by methylation is the default state with 201,371 gene pairs (92% of possible gene pairs) being identified as co-occurring even after correcting for multiple testing. Using the Poisson binomial test, which accounts for the baseline distributions the number of co-occurring gene pairs fell dramatically to just 2,369 (1% of possible gene pairs).

5.3.3.3.2 Calculations combining data types.

When I included all somatic mutation, CNV and methylation data only 147 gene pairs of a possible 109,330 pairs (0.134%) were identified as mutually exclusive using the

hypergeometric test. Using the Poisson binomial test this rose tenfold to 1847 gene pairs (1.69%). Mutually exclusive pairs were found for 35 of the 145 TSa genes, using 196 of the 754 druggable genes. Results are shown in Figure 6.7 below. I distinguish here between pairs that share the one of more PPI network clusters in the string database (shown in blue) and those that do not (shown in pink). The hypergeometric test identified 73,754 gene pairs (67.5%) co-occurring. This fell to 2116 (1.9%) using the Poisson binomial test.

For mutually exclusive pairs, 4.7 % of pairs found using the Poisson binomial test were jointly captured by the hypergeometric test as well, whilst 99.9% of pairs were jointly rejected. For co-occurring pairs, 99.9% of pairs found by the Poisson binomial test were jointly captured by the hypergeometric test and 33.1% were jointly rejected.

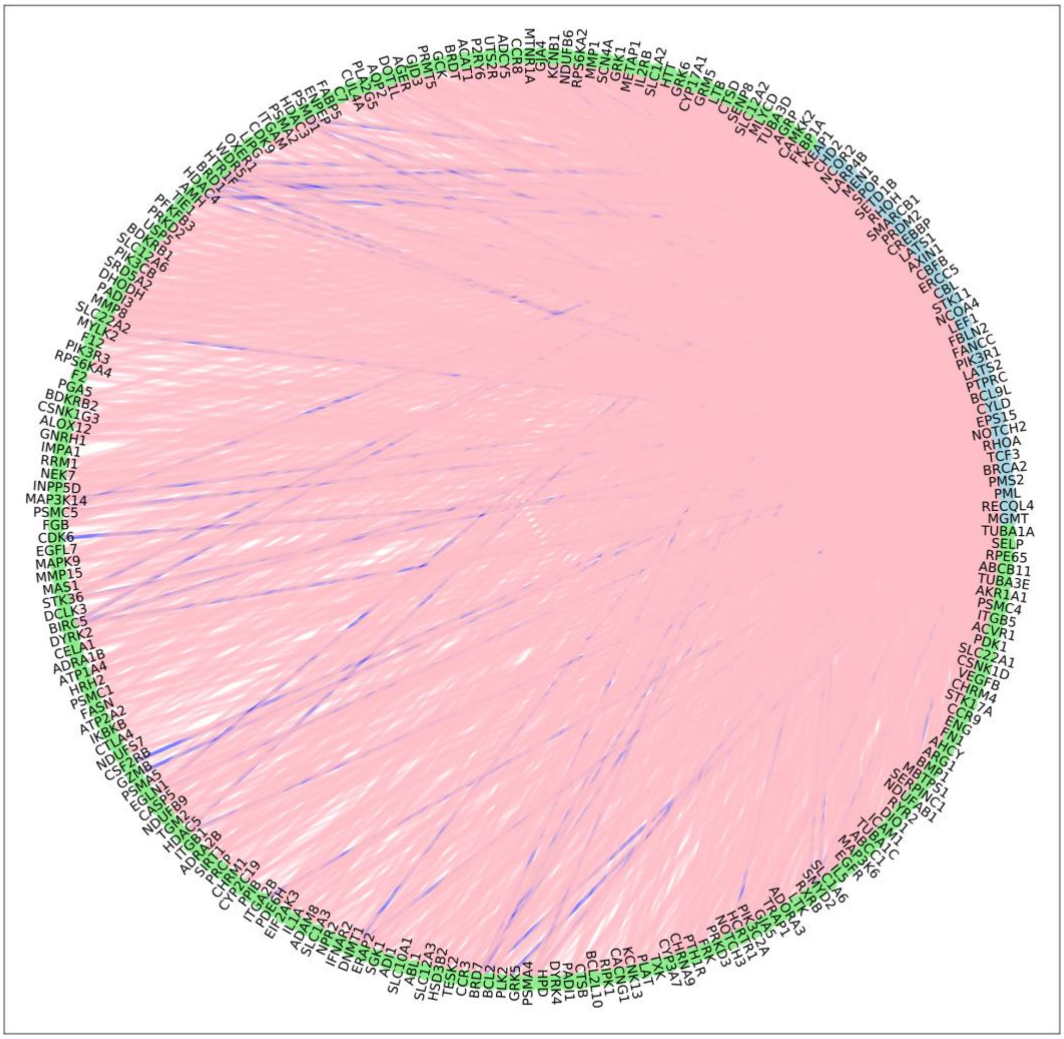


Figure 6.7: Using multi-omic data 1,847 gene pairs are identified as having mutually exclusive gene inactivations in samples with cancer of the large intestine. The gene pairs are shown above in blue for tumour suppressors (35) and green for druggable genes (196). In 1787 of the pairs, TSa and druggable genes belong to different PPI network clusters (shown in pink). In 60 pairs the genes share one or more pathway (shown in blue).

5.3.3.3.3 Therapeutic opportunities

70% of the samples included had inactivated genes which showed mutual exclusivity with one or more genes that can be inhibited by Indibulin. Indibulin is a synthetic small molecule which destabilizes tubulin polymerization thus inducing tumour cell cycle arrest and apoptosis. Indibulin has been found to have a potent activity in a number of cancer cell lines including colon cancer cell lines[293][294].

5.3.3.4 Lung Cancer

5.3.3.4.1.1 Mutation and CNV data

Differential somatic mutation status was found for 34 TSa genes and one druggable gene (NTRK3). However, no mutually exclusive or co-occurring pairs were found. Differential CNV status was found for four TSa genes (CDKN2A, FHIT, LRP1B, PTPRD) as well as 49 druggable genes, giving rise to 196 gene pairs that could be either mutually exclusive or co-occurring. One pair, CDKN2A, TRDV1 were mutually exclusive using the Poisson binomial test.

5.3.3.4.1.2 Methylation data

As before, I analysed the methylation samples to see whether or not there is much variation in the probability of inactivation across the different samples, necessitating the use of the Poisson binomial test.

Analysis of the 729 lung cancer methylation samples suggested that the samples split into two distinct clusters, but that there is less inactivation via methylation than in most cancers. Roughly 650 samples had less than 250 genes inactivated. However, around 30 samples

have between roughly 650 and 2900 inactivations and the remainder had more than 9,000 inactivations as shown in figure 6.8a) below.

I estimated these probabilities as a piecewise linear function of the total number of inactivations t in the sample, to give me an inactivation function, $f(t)$, for each gene (see methods for more details). The inactivation functions for each gene take two main shapes. For almost all genes the inactivation function is low for all of the first cluster, then climbing or is flat for the second cluster see figure 6.8b) However, for 36 genes it climbs rapidly for the first cluster of samples (to at least 0.1), and shows distinctly different behaviour for the second cluster of samples, in three cases then falling see figure 6.8c).

5.3.3.4.1.2.1 Calculations using hypermethylation data

In total, there were 197 TSa genes and 1,157 druggable genes with differential methylation status giving rise to 227,929 gene pairs to analyse. 26 gene pairs (0.01%) were identified as mutually exclusive using the hypergeometric test and this rose marginally to 35 pairs (0.02%) using the Poisson binomial test. If a hypergeometric test is used, co-occurrence of genes inactivated by methylation is the default state with 223,814 gene pairs (98.2%) being identified as co-occurring. Using the Poisson binomial test, which accounts for the baseline distributions the number of co-occurring gene pairs fell dramatically to just 1,389 (0.61% of possible gene pairs).

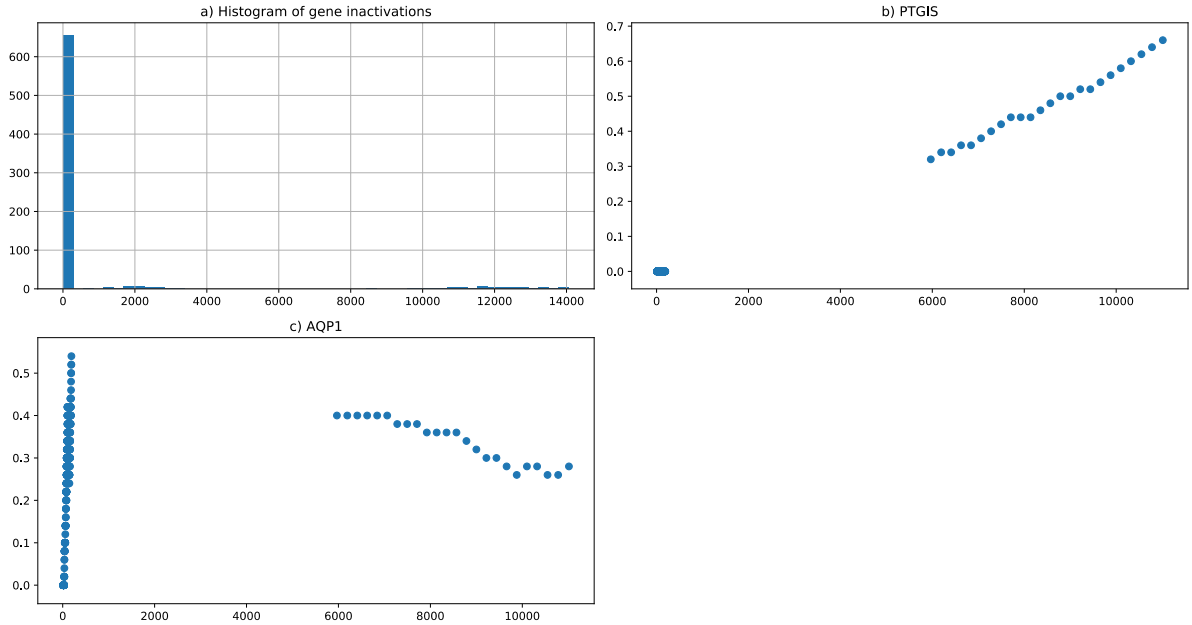


Figure 6.8: In lung cancer the number of genes inactivated by methylation is less variable than in many other cancers. In subplot a) I plotted a histogram of the number of methylated genes in each sample. There appear to be three distinct groups of samples, Generally, patients with lung cancers have less than 250 genes inactivated. However, roughly 30 have a few thousands and roughly 50 have more than 9,000. The mean probabilities were calculated over the closest 50 samples within the same cluster. Two behaviours were seen. For most genes the probability of gene inactivation is zero for most samples (illustrated here in figure 6.8b) for PTGIS), and rises consistently for those samples with more gene inactivations. However , 36 genes stand out and the probability of inactivation in these genes (illustrated by AQP1 in figure 6.8c) grows strongly as a function of total inactivations. Samples with many inactivations show very different behaviours.

5.3.3.4.2 Calculations combining data types

When I included all somatic mutation, CNV and methylation data, 20 gene pairs of a possible 328,104 pairs (0.006%) were identified as mutually exclusive using the hypergeometric test. This rose to 96 gene pairs (0.29%) using the Poisson binomial test. Mutually exclusive pairs were found for 91 of the 248 TSa genes, using just 5 of the 1323 druggable genes with differential inactivation. See figure 6.9. The hypergeometric test identified 279,943 gene pairs (85.3%) as co-occurring. This fell to 814 (0.25%) using the Poisson binomial test.
For mutually exclusive pairs, 3.1 % were jointly accepted and 100% jointly rejected. For co-occurring pairs 100% were jointly accepted and 14.7% jointly rejected.

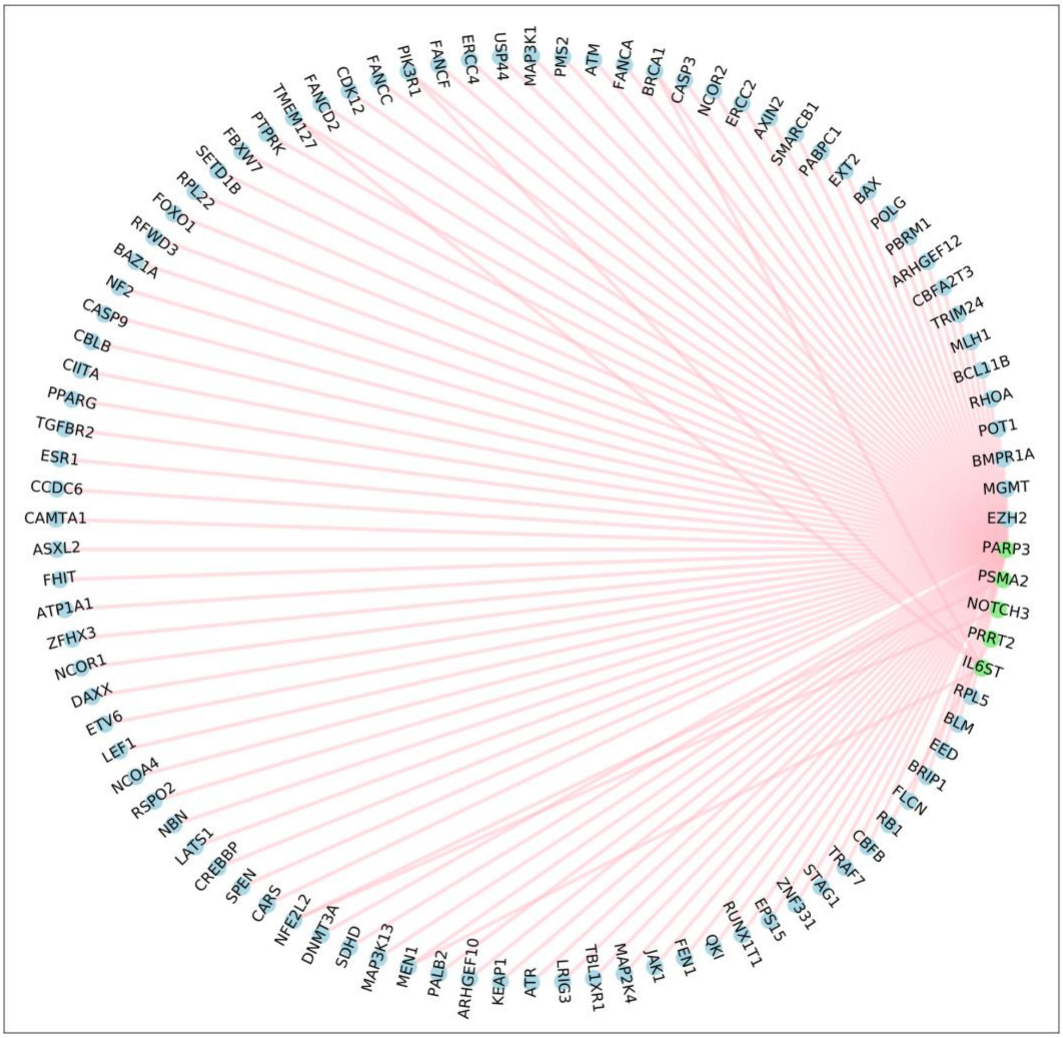


Figure 6.9: Using multi-omic data 96 gene pairs are identified as having mutually exclusive gene inactivations in samples with lung cancer. The gene pairs are shown above in blue for tumour suppressors (91) and green for druggable genes (5). In each of the pairs, TSa and druggable genes belong to different string clusters.

5.3.3.4.2.1 Therapeutic Opportunities

The druggable gene PARP3 forms part of 89 of the 96 mutually exclusive pairs suggesting that a PARP3 inhibitor maybe of utility in the treatment of lung cancer. PARP3 inhibitors include rucaparib, olaparib and niraparib [295]. A number of trials have been undertaken or are underway to look at the impact of parp inhibitors in lung cancer, with some, albeit limited, signs of promise. Overall response rate for participants in a part II lung cancer trial were higher for those who took veliparib alongside temozolomide compared with those taking temozolomide and a placebo.

Olaparib is currently being used in combination with cediranib in stage II trials for patients with advanced or metastatic solid lung tumours (NCT02498613). Rucaparib is currently in stage II trials for patients with Recurrent Non-small Cell Lung Cancer (NCT03845296).

5.3.3.5 Liver Cancer

5.3.3.5.1 Calculations using individual data types

5.3.3.5.1.1 Mutation and CNV data

Eight TSa genes; ARID1A, CSMD3, KMT2C, LRP1B, PTPN13, SETD2, SMARCA4, and TP53 had differential somatic mutation status, but none of the druggable genes did. By way of contrast, no TSa genes had differential CNV status, but three druggable genes have differential CNV loss, namely; DDC, MMP26, RHD. Consequently, mutational data and CNV could not be analysed independently.

5.3.3.5.1.2 Methylation data

5.3.3.5.1.2.1 Analysis of the methylation samples

I found that methylation samples appear to cluster into two groups with around 200 samples having less than 300 inactivations, whilst the remaining 100 had between roughly 900 and 4300 inactivations. See figure 6.10a). For 83% of genes the inactivation function, $f(t)$, was strongly correlated to t see figure 6.10b). For 10% of the genes $f(t)$ was a switch the gene was not inactivated for samples with fewer inactivations, but then inactivated amongst those with more inactivations see figure 6.10c). Just 3 genes showed a strong anti-correlation.

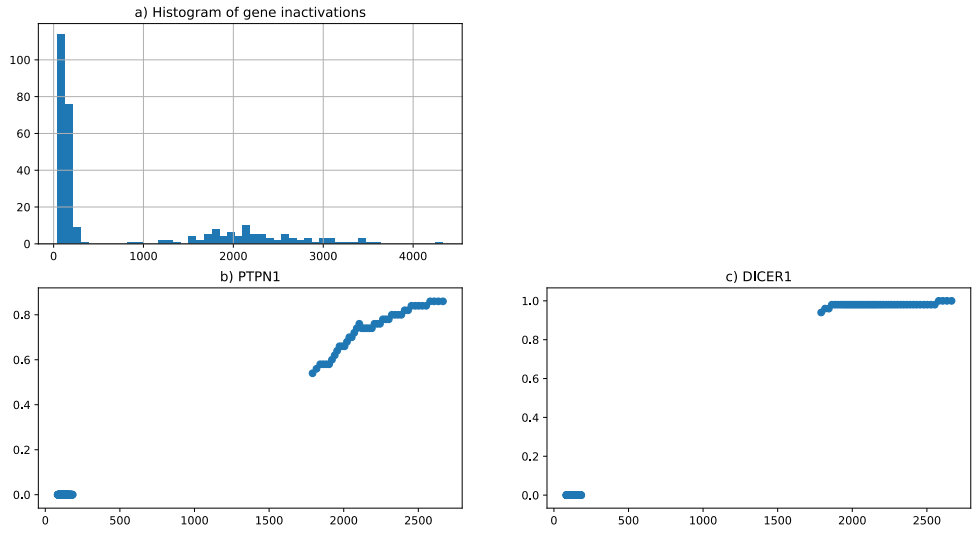


Figure 6.10: In subplot a) I plotted a histogram of the number of methylated genes in each sample. Samples generally had either less than 300 inactivations or between 1000 and 3000 inactivations as shown in a). In subplots b) and c) I plotted the mean probability of a gene being inactivated against the total number of gene inactivations due to methylation. The mean probabilities were calculated over the closest 50 samples within the same cluster. Two types of behaviours were seen. Most genes showed a strong correlation between the

probability of inactivation and the total number of inactivations as shown here by b) PTPN1, though a minority showed switching behaviour as shown by c) DICER1

There were 75 TSa genes and 410 druggable genes which had differential methylation status, giving a possible 30,750 gene pairs to analyse. 11 gene pairs (0.036%) were identified as mutually exclusive using the hypergeometric test, falling to 4 gene pairs (0.013%) using the Poisson binomial test. 21,429 gene pairs (70%) were identified as mutually exclusive using the hypergeometric test, falling to 7 gene pairs (0.023%) using the Poisson binomial test.

5.3.3.5.2 Calculations combining data types.

When I included all somatic mutation, CNV and methylation only 10 gene pairs of the possible 45,216 gene pairs (0.02%) were identified as mutually exclusive using the hypergeometric test. Using the Poisson binomial test, surprisingly, this fell further to just 6 gene pairs (0.01%) involving three of the 96 TSa genes and six of the 471 druggable genes. These are shown in figure 6.11 below. The hypergeometric test identified 17,850 gene pairs (39%) as co-occurring. This fell to 17 (0.4%) using the Poisson binomial test.

A large majority of the gene pairs predicted as co-occurring by the Poisson binomial test are also predicted as co-occurring by the hypergeometric test (85.7%). However, there is no overlap between the gene pairs predicted as mutually exclusive by the hypergeometric test and those predicted as mutually exclusive by the Poisson binomial test.

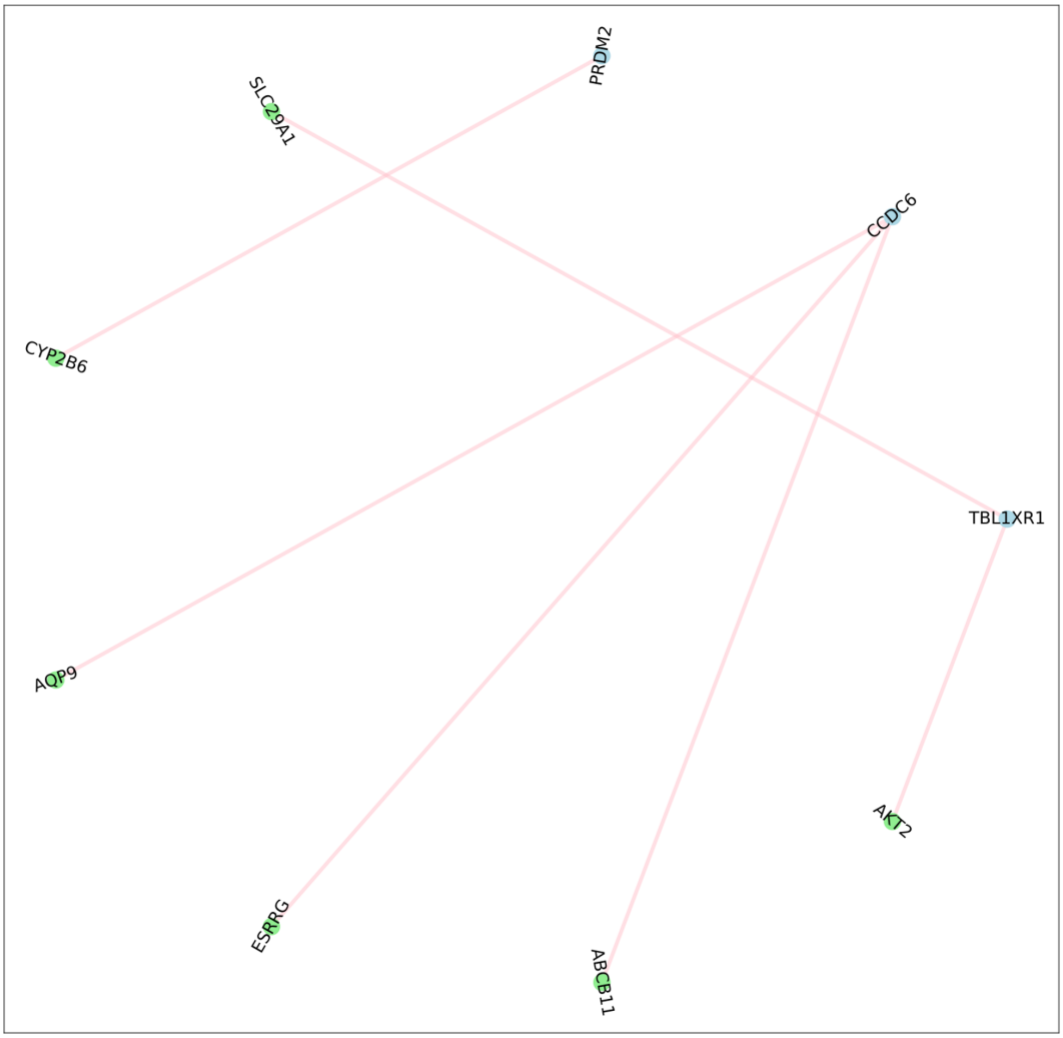


Figure 6.11: Using multi-omic data 6 gene pairs are identified as having mutually exclusive gene inactivations in samples with liver cancer. The gene pairs are shown above in blue for tumour suppressors (3) and green for druggable genes (6). In each of the pairs, TSa and druggable genes belong to different pathways.

5.3.3.5.2.1 Therapeutic Opportunities

Despite the overall small numbers of mutually exclusive pairs found, there are signs of a synthetic lethality relationship that could be exploited therapeutically. In particular, a recent study found that inhibition of AKT signalling by AKT inhibitor viii in hepatoma cells induced apoptotic cell death[296]. AKT2's mutually exclusive partner TBL1XR1 was inactivated in 16% of the samples suggesting that AKT2 could be a therapeutic strategy.

5.3.3.6 Ovarian Cancer

5.3.3.6.1 Calculations using individual data types

5.3.3.6.1.1 Mutation and CNV data

I found three TSa genes; CSMD3, NF1 and TP53, with differential somatic mutation status but no druggable genes. On the other hand, I found fifty-one druggable genes with differential CNV status, but no TSa genes. Consequently, mutational data and CNV data could not be analysed independently.

5.3.3.6.1.2 Methylation data

Methylation data alone enabled the identification of 140 genes that were differentially inactivated. Of these, 12 were target genes and 128 were druggable genes. Whilst this is less than for other cancers it still gives 1,536 gene pairs which could potentially have mutually exclusive or co-occurring interactions.

5.3.3.6.1.2.1 Analysis of the methylation samples

Compared to other cancers, the methylation samples had fewer inactivations. Analysis of the 512 ovarian cancer methylation samples suggested that the samples split into two distinct clusters with either fewer than 130 inactivations (roughly 200 samples) or between 170 and 820 inactivations (roughly 300 samples) See figure 6.12a).

To use the Poisson binomial test, I estimated the probability of each gene being inactivated in each sample as a piecewise linear function of the total number of inactivations t in the sample, to give me an inactivation function, $f(t)$, for each gene (see methods for more details). In roughly 5% of genes a switching inactivation function was seen the gene is predominantly turned on in those samples with few inactivations and off in those with many (see figure 6.12 b). For 85% of genes there was a strong positive correlation between the probability of inactivation and the total number of inactivations (see figure 6.12c).

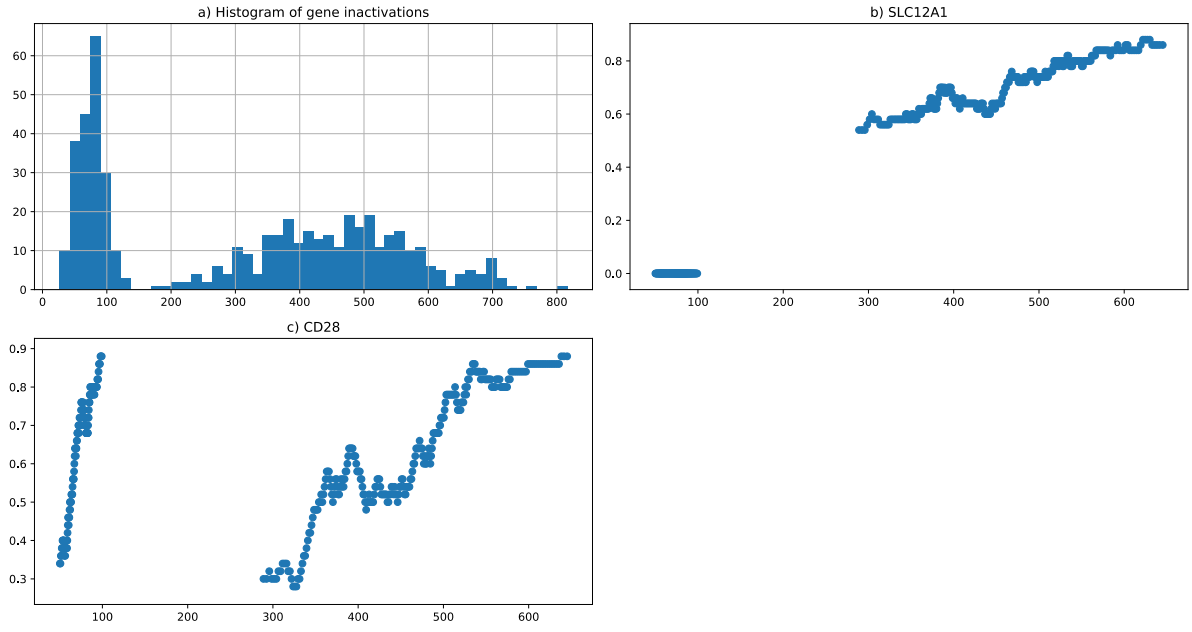


Figure 6.12: In ovarian cancer there are comparatively few genes inactivated by methylation. In subplot a) I plotted a histogram of the number of methylated genes in each

sample. There are two distinct groups of samples, generally with less than 100 or between 200-800. In subplot b-c I plotted the mean probability of a gene being inactivated against the total number of gene inactivations due to methylation. The mean probabilities were calculated over the closest 50 samples within the same cluster. Two behaviours were seen. For many genes, the gene is not inactivated in the first group, but is inactivated in the second (shown here by SLC12A1 in subplot b), but for around one third the probability of inactivation is positively correlated with the number of inactivations in both clusters (shown by CD28 in subplot c).

Most genetic inactivations occurred independently of one another. Only 4 gene pairs of the possible 1536 possible pairs (0.3%) were identified as mutually exclusive using the hypergeometric test and this rose only marginally to 5 gene pairs (0.3%) using the Poisson binomial test. These were (MGMT/AOAH, MGMT/CXCR2, RUNX1/HDC, RUNX1/MYLK2 and SOCS1/CD80) Co-occurrence of genes inactivated by methylation is reasonably common with 696 of the possible 1,526 gene pairs (45%) being identified as co-occurring using the hypergeometric test. This fell to just 14 (0.9%) using the Poisson binomial test.

5.3.3.6.2 Calculations combining data types.

When I included all somatic mutation, CNV and methylation data, I identified 4 pairs of a possible 3,600 gene pairs (0.1%) as mutually exclusive, using either the hypergeometric test or Poisson binomial test. This included 4 TSas and 3 druggable genes. ARHGEF10 was mutually exclusive with both DYRK2 and ITGB2 whilst MGMT was mutually exclusive with AOAH and RUNX1 was mutually exclusive with HDC. See figure 14. In addition, I found that

690 (19%) of a possible 3600 gene pairs co-occurred using the hypergeometric test. This dropped to 48 (1.3%) using the Poisson binomial test.

For mutually exclusive pairs, 75% were jointly captured whilst 99.9% of pairs were jointly rejected. For co-occurring pairs 100% of pairs were jointly captured by the hypergeometric test and 81.91% were jointly rejected.

5.3.3.6.2.1 Therapeutic Opportunities

Although the number of mutually exclusive gene pairs found was limited, they include the pair ARHGEF10/ IGTB2. ARHGEF10 is inactivated in 6% of the ovarian cancer samples. IGTB2 is also known as CD18 which has been shown to be inhibited in rats by methotrexate[297]. Low doses of methotrexate have been successfully trialled alongside cyclophosphamide in women with recurrent ovarian cancer[298] .

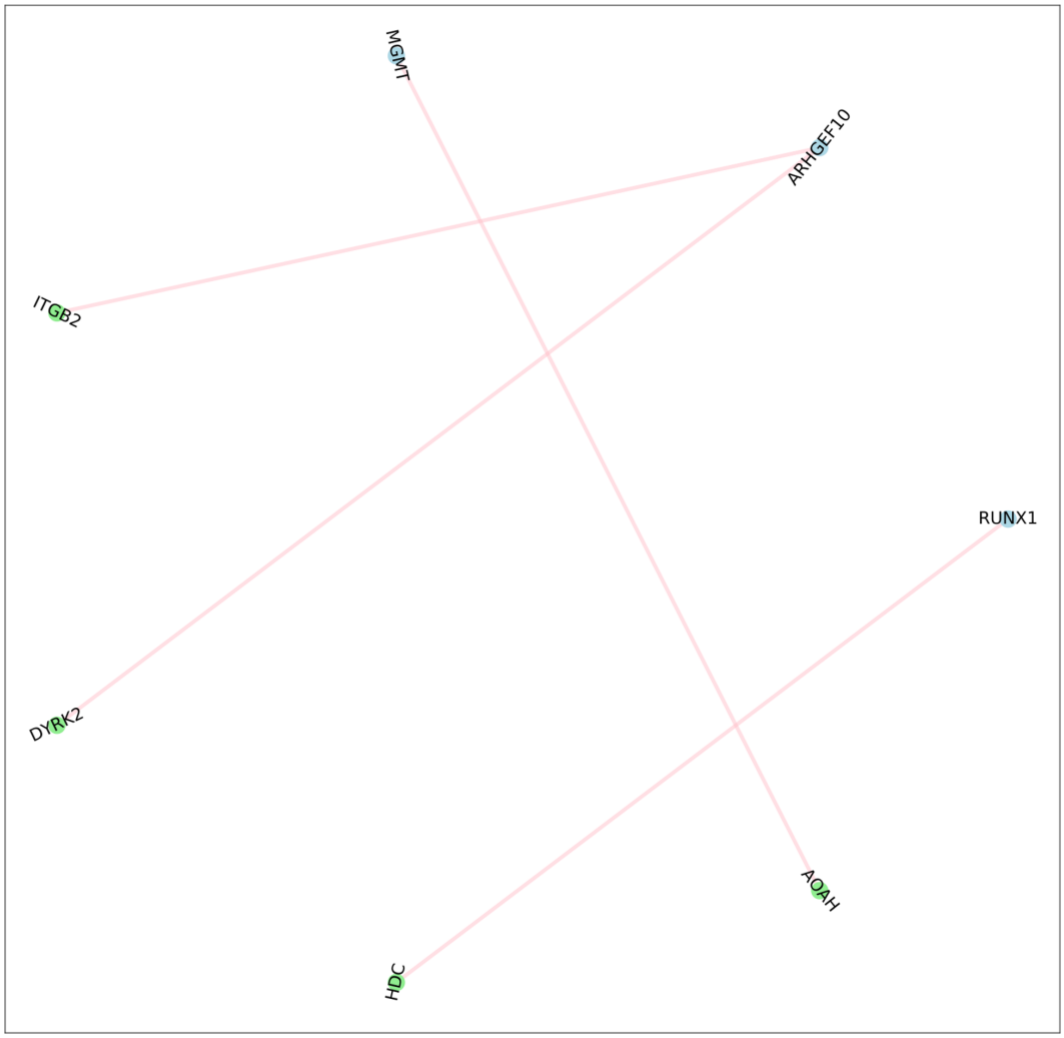


Figure 6.13: Using multi-omic data 4 gene pairs are identified as having mutually exclusive gene inactivations in samples with ovarian cancer. The gene pairs are shown above in blue for TSas (3) and green for druggable genes (4). In each of the pairs, target and druggable genes belong to different string clusters.

5.3.3.7 Prostate Cancer

5.3.3.7.1 Calculations using individual data types

5.3.3.7.1.1 Mutation and CNV data

Four TSa genes; KMT2C, LRP1B, SPOP, TP53, had differential somatic mutation status, but none of the druggable genes did, so mutational data could not be analysed independently. Forty genes had differential CNV status; the TSa gene PTEN and 39 druggable genes. No mutually exclusive gene pairs were found. Using the hypergeometric test but not the Poisson binomial test, PTEN and LIPF were found to co-occur.

5.3.3.7.1.2 Methylation

I found 97 TSa genes and 535 druggable gene with differential methylation status giving a possible 51,895 gene pairs which could be potentially mutually exclusive or have co-occurring relationships.

5.3.3.7.1.2.1 Analysis of the methylation samples

Over 95% of samples had between 2000 and 4000 genes inactivated via methylation while the remainder had just a few hundred. To use the Poisson binomial test, I estimated the probability of each gene being inactivated in each sample as a piecewise linear function of the total number of inactivations t in the sample, to give me an inactivation function, $f(t)$, for each gene (see methods for more details). As for other cancers most genes showed either strong correlation or switching behaviour, however for around 4% of genes there is a strong anticorrelation between probability of inactivation and the total number of inactivations, see figure 6.14.

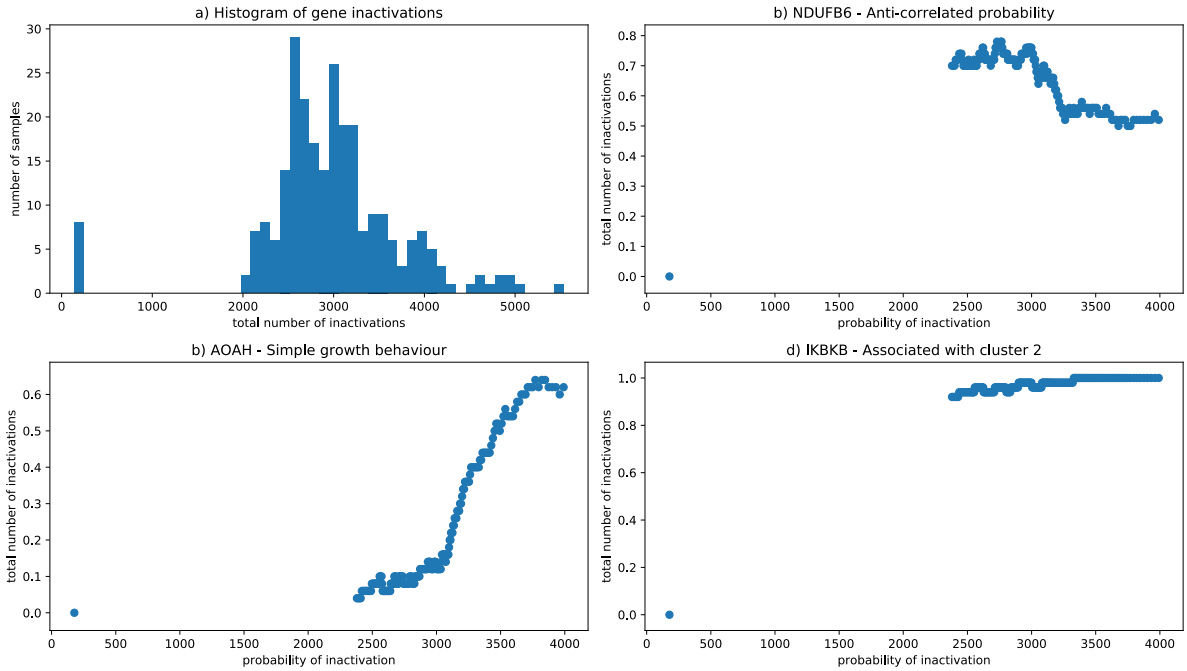


Figure 6.14: In prostate cancer most samples had around 3000 genes inactivated by methylation. In subplot a) I plotted a histogram of the number of methylated genes in each sample. There appear to be two distinct groups of samples, 5% had less than two hundred whilst the remainder had between 2000-5000. In subplot b-d I plotted the mean probability of a gene being inactivated against the total number of gene inactivations due to methylation. The mean probabilities were calculated over the closest 50 samples within the same cluster. Three main patterns of $f(t)$ emerged: b) For 4% of the genes $f(t)$ was highly anti-correlated with t shown here by NDUFB6; c) for most genes (66%) $f(t)$ was highly correlated, shown here by AOAH, and d) in 10% of genes $f(t)$ was a switch, shown here by IKBKB.

Using the hypergeometric test 234 gene pairs (0.4%) were identified as mutually exclusive. Unusually, using the Poisson binomial test this figure fell further to just 23 pairs (0.04%).

Using the hypergeometric test, co-occurrence of genes inactivated by methylation is only 26% (13,726 of the possible 51,895 gene pairs). This fell to 121 (0.2%) using the Poisson binomial test.

5.3.3.7.2 Calculations combining data types.

When I included all methylation, CNV and somatic mutation data and used the hypergeometric test 224 gene pairs (0.42%) were identified as mutually. However, this fell to just 17 pairs (0.03%) using the Poisson binomial test. 11,142 gene pairs (21.0%) co-occurred of a possible 53,025. This fell to 81 (0.15%) using the Poisson binomial test.

For mutually exclusive pairs, 76.5% of pairs found using the Poisson binomial test were jointly captured by the hypergeometric test as well, whilst 99.6% of pairs reject by the Poisson binomial test were jointly rejected by the hypergeometric test. For co-occurring pairs, 98.8% were jointly and 79.1% jointly rejected.

Nine target genes had mutually exclusive pairs from a possible 101, using fourteen of a possible 525 druggable genes (see figure 6.15).

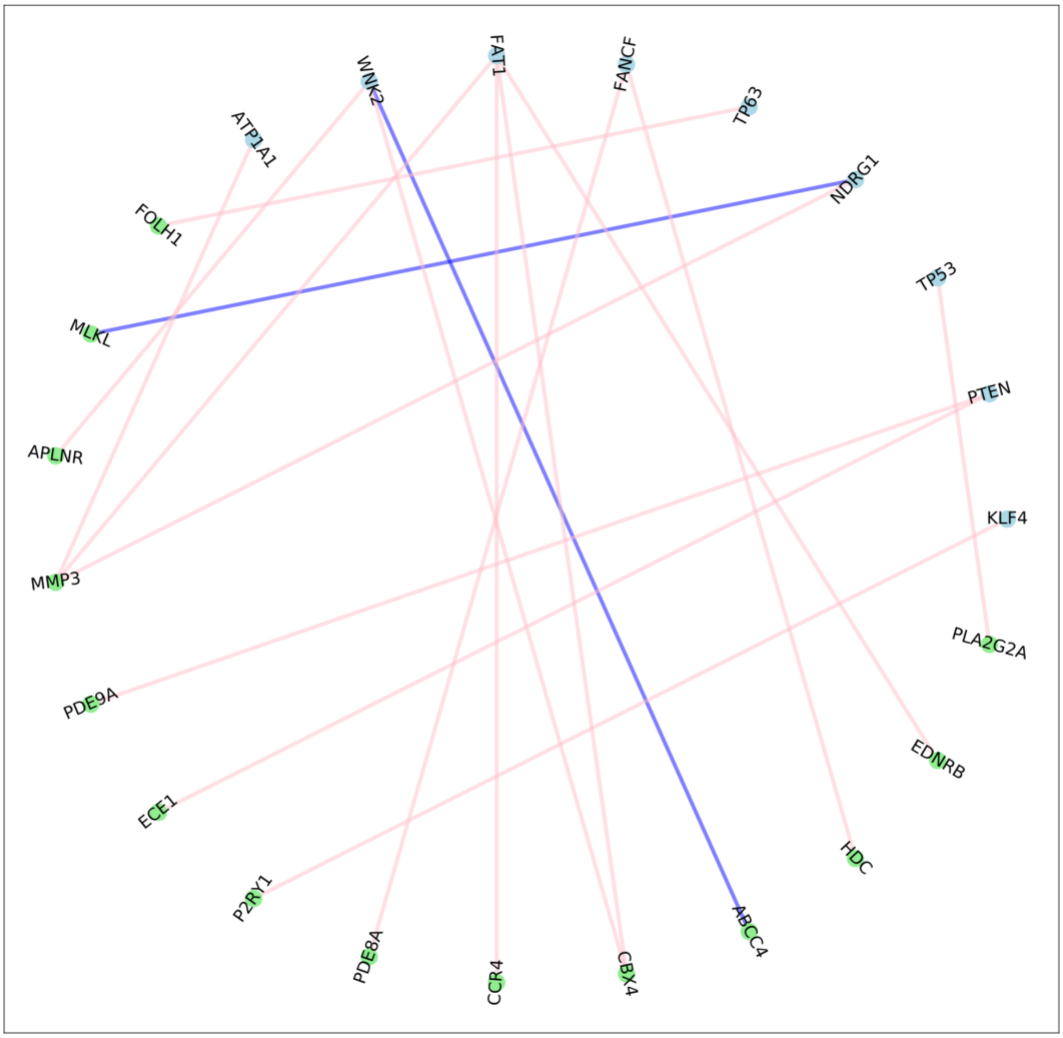


Figure 6.15: Using multi-omic data 17 gene pairs are identified as having mutually exclusive gene inactivations in samples with prostate cancer. The gene pairs are shown above in blue for tumour suppressors (9) and green for druggable genes (14). In 15 of the pairs, target and druggable genes belong to different pathways. In two pairs the genes share one or more pathway (shown in blue).

5.3.3.7.2.1 Therapeutic Opportunities

Despite the relatively small number of mutually exclusive gene pairs found, it is possible that some may be synthetically lethal. Inactivation of TP63, found in 74% of the prostate cancer samples, is mutually exclusive with the inactivation of FOLH1, also known as PSMA. PSMA is a known target and expression is a biomarker for prostate cancer, currently treated with Docetaxel [299][300].

5.3.4 Pan-Cancer Analysis

In most of the cancer types analysed, between 3-8 genes were differentially inactivated through somatic mutation, although 35 met this criteria for lung cancer. Similarly, the number of genes differentially inactivated in the CNV data tended to be modest, ranging from 4 genes in samples from the large intestine to 51 in samples from the ovaries. This means that I most tissues there was not enough data to identify large numbers of mutual exclusive gene pairs using mutational data and CNV alone.

Almost all of the differentially mutated genes were TSa genes rather than drug targets. This is not surprising as the druggable genes tend not to be cancer driver gene hence inactivating mutations in them are rare. On the other hand, in the large majority of cases genes inactivated through CNV were druggable genes rather than TSa genes. This suggests that the impact of copy number changes from the direct deletion of tumour suppressors is limited.

In contrast, methylation data was rich in differentially inactivated genes, with all of the cancers under consideration having several hundred differentially inactivated genes. However, even within the same cancer the extent of genes inactivated through methylation

varied considerably by sample. Within each cancer there appeared to be either two or three distinct clusters of samples. Clusters had either a few hundred, a few thousand or more than 10,000 inactivations.

In total I found 5403 distinct mutually exclusive pairs and 8471 co-occurring pairs. The numbers of mutually exclusive or co-occurring gene pairs are shown in tables 6.4 and 6.5 below. A full list of all the mutually exclusive gene pairs found using the Poisson binomial test is available on MexDrugs at https://users.sussex.ac.uk/~skw24/mexdrugs1/index.html .

Site	Statistical method	Data type	# ME pairs found	# target genes	# druggable genes
Breast	hypergeometric	Combined omic data	462	71	120
Breast	Poisson binomial	Combined omic data	492	36	114
Breast	hypergeometric	Just methylation data	352	74	86
Breast	Poisson binomial	Just methylation data	752	40	160
Kidney	hypergeometric	Combined omic data	532	44	229
Kidney	Poisson binomial	Combined omic data	3014	61	250
Kidney	hypergeometric	Just methylation data	691	77	245

Kidney	Poisson binomial	Just methylation data	2453	60	463
Large intestine	hypergeometric	Combined omic data	147	11	88
Large intestine	Poisson binomial	Combined omic data	1847	35	196
Large intestine	hypergeometric	Just methylation data	90	8	37
Large intestine	Poisson binomial	Just methylation data	3007	41	246
Liver	hypergeometric	Combined omic data	10	5	8
Liver	Poisson binomial	Combined omic data	6	3	6
Liver	hypergeometric	Just methylation data	11	6	9
Liver	Poisson binomial	Just methylation data	4	1	4
Lung	hypergeometric	Combined omic data	20	4	16
Lung	Poisson binomial	Combined omic data	96	91	5
Lung	hypergeometric	Just methylation data	26	6	15
Lung	Poisson binomial	Just methylation data	35	6	14
Ovary	hypergeometric	Combined omic data	4	3	4

Ovary	Poisson binomial	Combined omic data	4	3	4
Ovary	hypergeometric	Just methylation data	4	2	4
Ovary	Poisson binomial	Just methylation data	5	3	5
Prostate	hypergeometric	Combined omic data	224	39	128
Prostate	Poisson binomial	Combined omic data	17	9	14
Prostate	hypergeometric	Just methylation data	234	35	129
Prostate	Poisson binomial	Just methylation data	23	12	20

Table 6.4. Numbers of mutually exclusive inactivated gene pairs identified using combined CNV, somatic mutation and methylation data, or just methylation data, using either the hypergeometric test or Poisson binomial test.

Site	Statistical method	Data type	#co-occurring pairs found	# Tumour suppressor genes	# druggable genes
Breast	hypergeometric	Combined omic data	47762	114	606
Breast	Poisson binomial	Combined omic data	854	57	180

Breast	hypergeometric	Just methylation data	50265	107	581
Breast	Poisson binomial	Just methylation data	1037	61	224
Kidney	hypergeometric	Combined omic data	270134	235	1418
Kidney	Poisson binomial	Combined omic data	5230	76	401
Kidney	hypergeometric	Just methylation data	279727	228	1373
Kidney	Poisson binomial	Just methylation data	180681	218	1278
Large intestine	hypergeometric	Combined omic data	73754	141	753
Large intestine	Poisson binomial	Combined omic data	2116	37	196
Large intestine	hypergeometric	Just methylation data	201371	191	1150
Large intestine	Poisson binomial	Just methylation data	2369	43	221
Liver	hypergeometric	Combined omic data	17850	94	459
Liver	Poisson binomial	Combined omic data	17	9	13
Liver	hypergeometric	Just methylation data	21429	75	404
Liver	Poisson binomial	Just methylation data	7	3	7

Lung	hypergeometric	Combined omic data	279943	229	1322
Lung	Poisson binomial	Combined omic data	814	59	286
Lung	hypergeometric	Just methylation data	223814	197	1157
Lung	Poisson binomial	Just methylation data	1389	50	287
Ovary	hypergeometric	Combined omic data	690	17	140
Ovary	Poisson binomial	Combined omic data	48	10	22
Ovary	hypergeometric	Just methylation data	696	12	122
Ovary	Poisson binomial	Just methylation data	14	7	12
Prostate	hypergeometric	Combined omic data	11142	101	515
Prostate	Poisson binomial	Combined omic data	81	18	49
Prostate	hypergeometric	Just methylation data	13726	97	532
Prostate	Poisson binomial	Just methylation data	121	22	61

Table 6.5. Numbers of co-occurring inactivated gene pairs identified using combined CNV, somatic mutation and methylation data, or just methylation data, using either the hypergeometric test or Poisson binomial test.

5.3.4.1.1 Tissue specificity

The gene dependency projects from both the Broad Institute and Sanger labs found that cell dependency and synthetic lethal pairs are highly tissue specific [301][302]. My results here confirm that finding. In fact, no mutually exclusive pairs found between tissue types with the exception of cancers of the breast, kidney and large intestine. Kidney and large intestine tumours share 69 mutually exclusive gene pairs, with kidney and breast sharing 4 mutually exclusive gene pairs. Although this may sound like a low number, it is in fact higher than expected by chance. The p-value is outside the limits of the statistical test used. I therefore show the shared pairs for kidney and large intestine below in figure 6.16. The numbers of mutually exclusive gene pairs that are shared between tissue types is shown in figure 6.17.

Cooccurring gene pairs were also generally tissue specific, but with some shared between at least two of breast, kidney, lung, and large intestine. One pair PTPRC/IL10 is of particular interest because it co-occurs in five different tissue types. Both of these genes are connected with B cell proliferation. These are shown in figure 6.19 below.

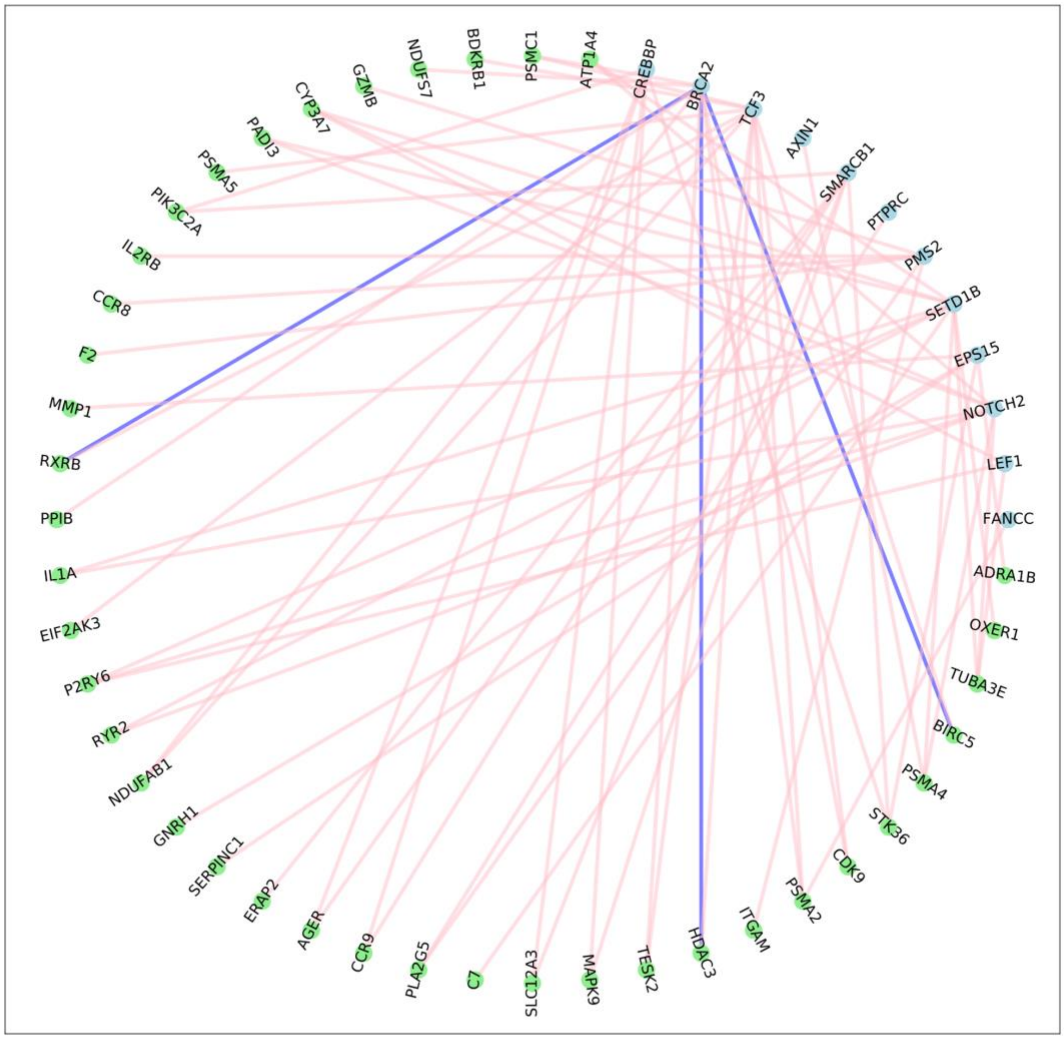


Figure 6.16: Using multi-omic data 69 gene pairs are identified as having mutually exclusive gene inactivations in samples with either cancer of the kidney or large intestine. The gene pairs are shown above in blue for tumour suppressors and green for druggable genes. The three gene pairs where both genes are in the same pathway are shown in blue.

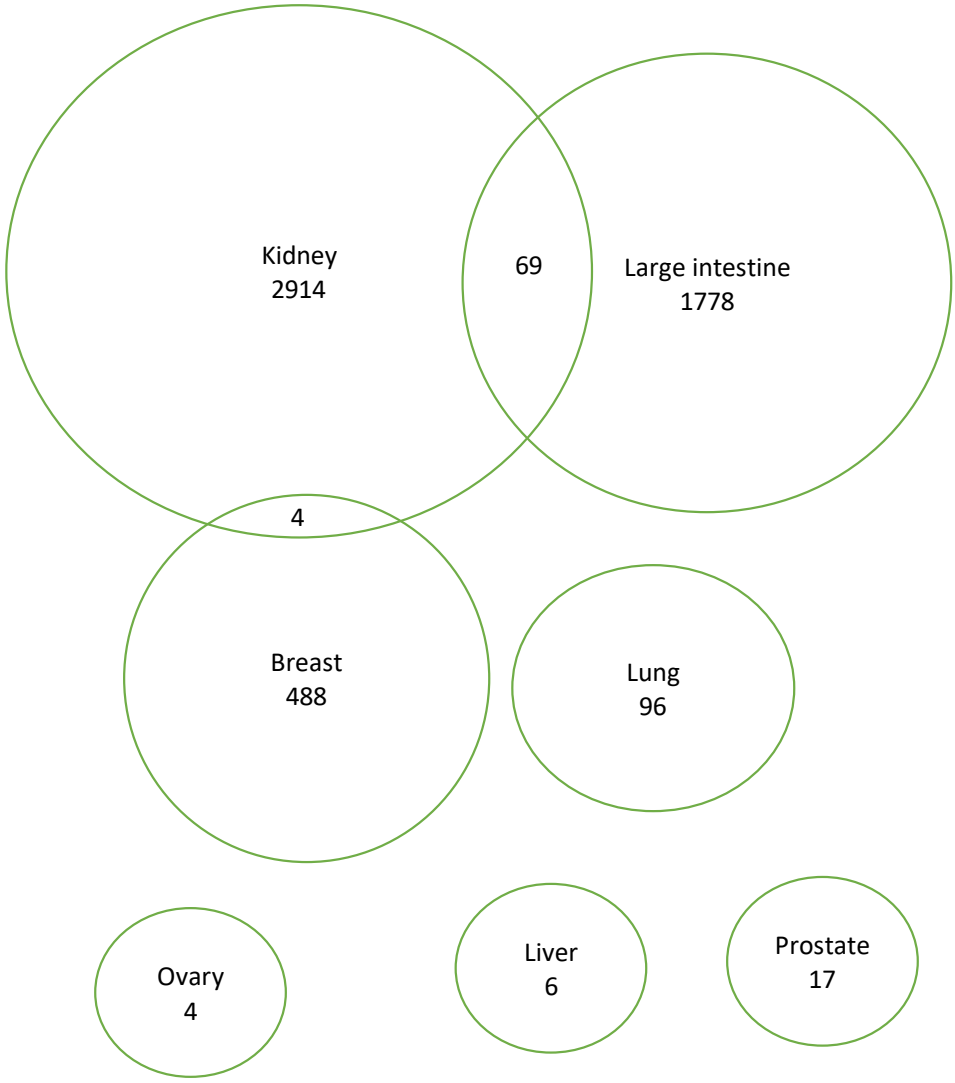


Figure 6.17: Numbers of mutually exclusive gene pairs in each of the tissue types, showing the number of mutually exclusive gene pairs that occur in more than one tissue type.

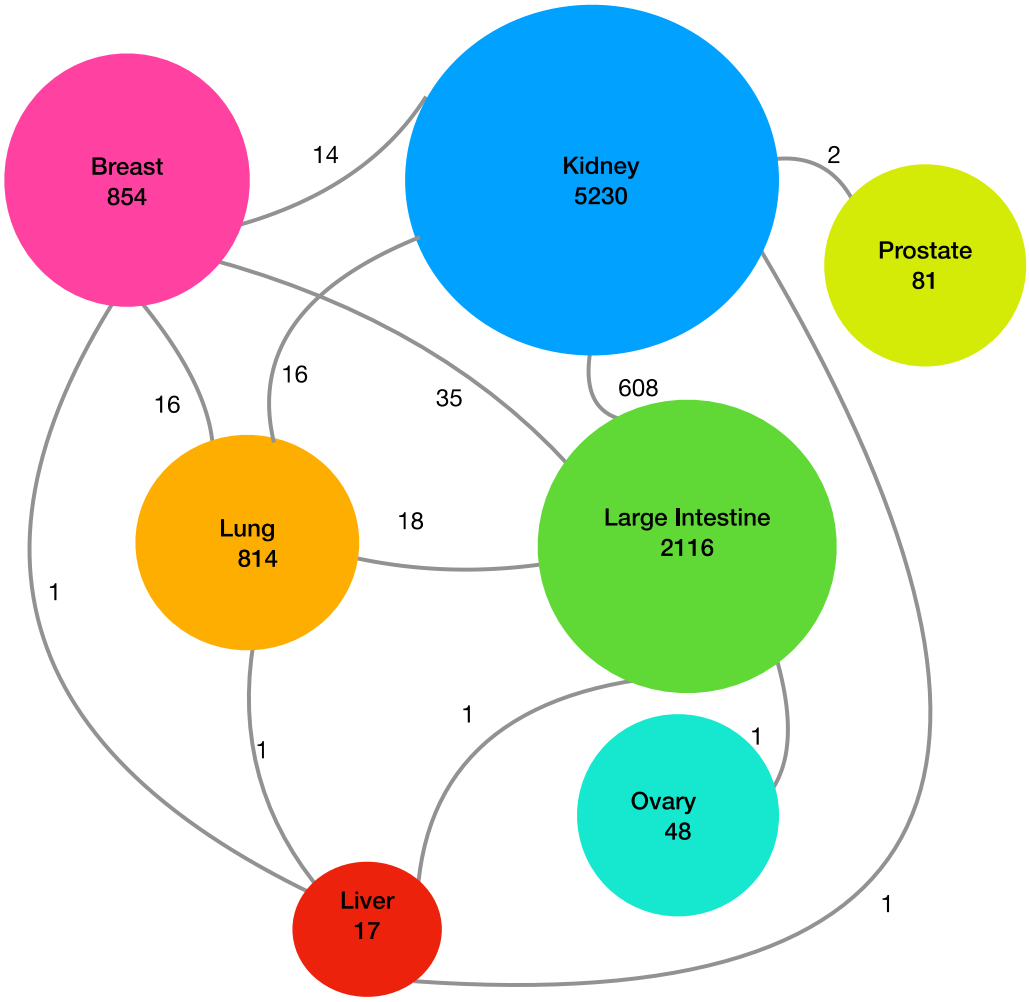


Figure 6.18: Number of co-occurring gene pairs in different tissue types. Just one gene pair is shared between five tissue types; PTPRC / IL10. Both of these genes are connected with B cell proliferation.

5.3.4.1.2 Mutually exclusive pairs tend to occur different pathways.

One possible explanation for mutual exclusivity is that the genes come from the same pathway: one of the genes being inactivates is sufficient to enable a tumorigenic phenotype, reducing or removing the selective pressure to inactivate the other gene. To test this theory, I looked at how many of the gene pairs shared at least one PPI interaction cluster. These clusters are predominantly median 12 genes. However, some of the clusters include almost all of the genes and are of no interest for pathway analysis. I therefore removed all the clusters above a 1000 threshold. Of the 5,403 gene pairs that were mutually exclusive in at least one tissue type, 5210 did not share a PPI interaction cluster pathway. Moreover, the cluster threshold was very robust. The number of gene pairs without a common PPI interaction cluster fell only to 4,495 if cluster sizes up to 4000 were included. This suggests that the mutually exclusivity seen does not generally come lack of pressure on genes in the same pathway.

By way of contrast, the 8,471 co-occurring gene pairs were all in clusters containing less than 134 genes, and the median size for the cluster was 5. This suggests that co-occurring gene pairs are generally in the same pathway.

5.4 MexDrugs

Mutual exclusivity cannot be taken as implying that two genes are synthetically lethal. However, sets of mutually exclusive gene pairs will be enriched in synthetically lethal pairs so it indicates that further investigation is of interest. I have therefore included information about the complete set of drugs associated with these mutually exclusive gene pairs in my website MexDrugs.

MexDrugs is an online database of the mutually exclusive interactions that were found using the methods in this chapter and the Poisson binomial test. Screen shots are shown in figure 6.19 below. It can be accessed at https://users.sussex.ac.uk/~skw24/mexdrugs1. The website is organised by cancer. For each cancer it lists the drugs that are identified in the Druggable genome databank [148] as inhibiting a gene which has mutually exclusive TSa partners where these are inactivated in more than 5% of samples. The mutually exclusive relationships are portrayed as a network and a list of the relevant mutually exclusive pairs, together with the percentage of samples affected can be downloaded for each drug. In addition, all the mutually exclusive results found using the Poisson binomial test can be downloaded from the title page.

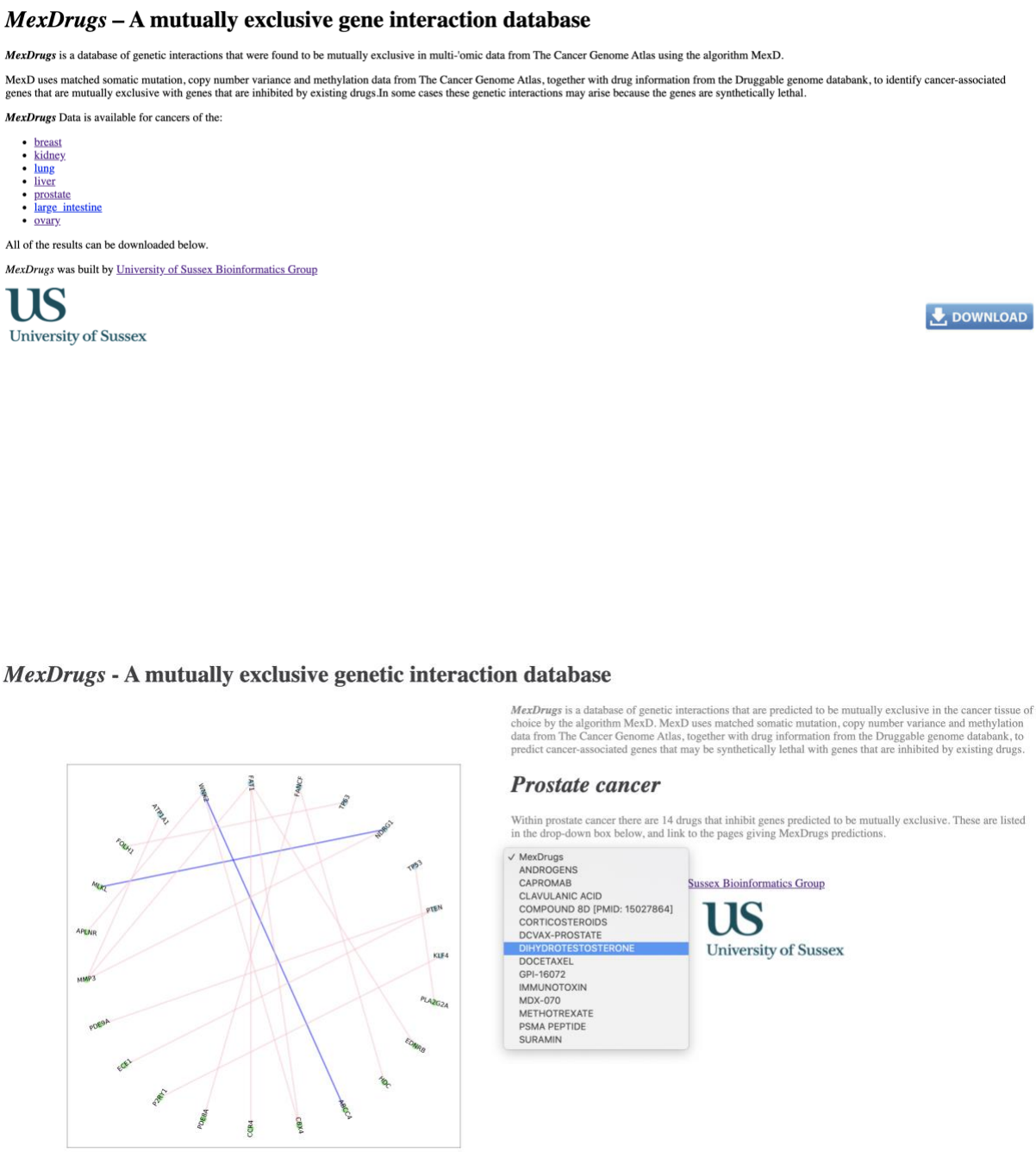


Figure 6.19: Screenshots of the website MexDrugs available at

https://users.sussex.ac.uk/~skw24/mexdrugs1.

5.5 Conclusion and discussion

Mutual exclusivity between inactivations in a pair of genes may indicate either: that any selective advantage resulting from the inactivation of one gene is not increased by inactivating both the genes, or alternatively that inactivating both the genes confers a distinct disadvantage to the tumour. That is, the combination is synthetically lethal.

Many of the existing approaches to find mutually exclusive pairs or groups of genes are hoping that by doing so they can find genes within the same protein pathway, as mutations that damage an already defunct pathway are considered unlikely to confer additional selective advantage to the tumour [278][280]. In contrast, in this work, I looked specifically at genes that were unlikely to be part of the same pathway, as these are more likely to reveal genuinely synthetically lethal genetic interactions.

Using the hypergeometric test, most gene pairs co-occurred in almost all cancers, and there were very few mutually exclusive pairs. However, I showed that the probability of a specific gene being inactivated has a strong dependency on the total number of inactivations. Since gene inactivations are not independently and identically distributed, the hypergeometric distribution is not the appropriate statistical distribution to use as a baseline against which to measure co-occurrence and mutual exclusivity. I turned instead to the Poisson binomial distribution which enabled me to use the approximate probabilities of each gene being inactivated in each sample, when forming a baseline.

By simulating data sets I showed that the Poisson binomial test improved the overall accuracy of prediction finding more of the mutually exclusive pairs and fewer false positive co-occurring pairs. For most cancers, this is also the case. The Poisson binomial test gives a fairly dramatic reduction in the number of co-occurring pairs and a substantial increase in the number of mutually exclusive gene pairs

In order to assess the probability of mutual exclusion or co-occurrence using the Poisson binomial distribution I had first to identify the probability of a specific gene g_i being is inactivated in a specific sample s_j (that is P_{ij}) as a function of the total number of inactivations in that sample t_j, i.e. $P_{ij} = f_i(t_j)$. I estimated these inactivation function, f_i , as a piecewise, linear average of the number of inactivations of g_i in each of roughly 10 bins with samples of similar t_j in them. In most cases, $f(t)$ was either a stepwise function, that is the probability of inactivation of a gene is almost zero for one cluster of samples and almost one for another, or it increased sharply as t increases. In either case $f(t)$ normally shows distinct differences between the different sample clusters. It is worth noting two other features of $f(t)$; firstly, that on occasion $f(t)$ decreases as t increases and that secondly, other functions are possible, including intriguingly two bowl shaped functions. These are reasonably inactivation functions for genes in kidney cancers. The biological significance of these functions is not currently known.

In breast cancers and cancer of the large intestine the methylation data alone provided an excellent approximation of the mutually exclusive gene pairs found using all the data, with agreement on over 90% of the gene pairs found. However, in others, the CNV and mutation data was sufficient to substantially change the mutually exclusive pairs predicted.

Personalised therapies for cancers provide a way of identifying the people who are most likely to benefit from a given drug, and exclude those for whom the therapy is unlikely to hold much promise. In view of the severe side-effects associated with many of current cancer drugs this is an important consideration. However, a personalised medicine has by definition a reduced target audience which can mean that the research needed to bring a new drug onto the market is not considered cost-effective. By focusing on drug repurposing, these costs, and the associated time needed to bring them to market is considerably reduced.

Demonstrating mutual exclusivity is just one of the steps on the way to showing that two genes are synthetically lethal. There is no substitute for experimental evidence. However, the many hundreds of mutually exclusive inactivations found in breast, kidney and colorectal cancers suggests that these are rich avenues to explore further.

Moreover, if the TCGA sample set is typical, then a high percentage of samples could stand to benefit. For example, 22 of the drugs considered inhibit genes where the mutually exclusive gene partner is inactivated in over 70% of the kidney samples considered. Unfortunately, I cannot conclude that 70% of patients could benefit: the mutually exclusive tests that I use are soft tests. They demonstrate only that less joint inactivations took place than expected. Some cells were able to carry inactivations in both genes at the same time, and more work is needed to understand what characterises these samples.

Finally, Canisius et al. found that there is hardly any co-occurrence in somatic mutations not explained by chance alone[277]. This is not true with multi-omic data. Although the reduction in co-occurrences is dramatic, many still remain. This suggests that groups of genes may be sometimes switched off together by hyper-methylated during tumorigenesis, as my findings are dominated by methylation data.

6 Discussion

6.1 Overview of major findings

Mutations and other genetic and epigenetic changes occur as a result of stochastic processes. The ones that survive in the genetic record reflect not only the endogenous and exogenous processes that cause the damage and the cells ability to repair the damage done, but also by the nucleotides in the neighbourhood of the damage and by the evolutionary advantage or disadvantage conferred by the changes. That is, we see only those changes that do not kill the cell, and we see a higher proportion of the changes that enable some form of advantage. These changes are not independent of one another: there are large numbers of genetic interactions that alter the impact of individual changes leading to synthetically lethal and co-occurring genetic inactivations.

In this thesis, I have sought to identify the contribution that each of these factors makes to the overall picture of mutations and other forms of genetic inactivations.

In chapter 2 I look at the impact that neighbouring nucleotides have on the patterns of mutations and in particular the pattern of indels. I find that that for medium length indels the frequency of indel is well described by a power law. However, I find an excess of inframe indels. Inframe indels occur disproportionately at the site of a repeat of the indel in the DNA. Whilst this suggests that replication slippage is the cause of the excess of inframe indels, this explanation is not sufficient to account for the excess. Looking just at those indels which are not next to a repeat, I show that the proportion of di-nucleotide indels to tri-nucleotide indels is less in the exome than in the non-protein-coding region. Since there

is far stronger selective pressure in the exome than else-where this surfeit of inframe indels suggests evidence of negative selection against frameshift indels.

In chapter 3 I use two different methods to look at associations between the processes that cause DNA damage (and the resulting mutational fingerprints/signatures) and the prevalence of mutations in cancer-associated genes. I find sixteen examples of genes in specific tissue types where the samples with a pathogenic mutation in the gene have a statistically significantly different mutational signature to those that do not. The mutational signatures provide a record of the mutational damage that is replicated in cancer cells and as such reflects both the damage done, and the cell's ability to repair the damage. It may be that the genes identified impact of the DDR pathways in some way. However, it is also possible the changes in the mutational signatures predispose the cell to gain mutations in the identified genes, or that the genetic mutations co-occur with other types of unidentified genetic changes such as hyper-methylation.

I find that the majority of cancer associated genes are mutated more frequently in those tissue types where the mutational fingerprints are conducive to creating the pathogenic mutations seen. There are some interesting exceptions to this, including BRAF V600E which occurs far more frequently in skin cancers than is expected from the makeup of the gene and mutational fingerprints of mutated samples.

Whilst chapter 3 makes extensive use of existing mutational signatures in cancer, in chapter 4 I apply this technique to identify novel bacterial signatures. In order to ensure that I use recent evolutionary mutations rather than historical fixed mutations to generate the

signatures I first clustered the strains using the sequence identity of genes that are shared by all the bacterial strains. I then identified unique substitutions, which are silent and thus not selected for. By normalising the signatures using the distribution of codons that could give rise to silent mutations in the bacterial subspecies, I am able to provide a cross-species comparison between different bacteria. Encouragingly I found that many of these signatures are very similar for different clusters of the same bacterial species. I also found that some of the signatures are statistically significantly similar to signatures derived from human cells . In some cases, there is a known aetiology for the human mutational signatures. This leads me to propose that some of the bacterial signatures could also have been created as a result of action by alkylating agents, failure of DNA MMR, and /or the error prone polymerase POLE.

In chapter 5 I present our published review of bioinformatics in translational drug discovery. This looks at the use of bioinformatics in analysing cancer data to enable a more personalized approach to cancer therapy, but also at the use of bioinformatics to identify how our genetic makeup affects our likelihood of developing a wide range of diseases, our responses to a variety of drug treatments and the progression of many infectious diseases. It then goes on to look at whether a particular target gene is likely to be druggable; how well the corresponding protein will bind small drug-like molecules, as well as whether the drug-like molecules have the right chemical properties to be successful as drugs, and whether we can predict translational properties of drugs by looking at protein-protein interaction data. This chapter touches on the nature of synthetic lethal genetic interactions, which I go on to explore in the next chapter.

Finally, in chapter 6, I look for mutually exclusive and co-occurring gene inactivations in cancer samples: bringing together data on mutations, on copy number variance and also on inactivations as a result of hypermethylation. Given the cost of bringing new drugs to market, and the potentially small markets for personalised therapies, I am particularly interested in translational drug repurposing. I therefore look at whether it might be possible to use existing drugs that have already been licensed for different purposes to target cancers that are driven by inactive tumour suppressors. To do this I look for mutually exclusive gene pairs where one of the genes is a tumour suppressor and the other is a gene whose protein product can be inhibited via a known drug. Using the hypergeometric distribution to test for such pairs leads to disappointingly few such predictions. However, I find that using the Poisson binomial distribution instead of the hypergeometric distribution improves the accuracy of results, and also has the effect of reducing the number of co-occurring gene inactivations and increasing the number of mutually exclusive ones.

6.2 Limitations

Data for this thesis is derived from the COSMIC data bank, from other Genomic Data Commons data on TCGA patients and from ensembl. These sources act as repositories for data from many different studies, often employing different standards and conventions. Improvements continue to be made to the standard of most of the data used within this thesis. For example, although the first human genome was completed in 2000, the genomic coordinate system GRCh37/Hg19 was significantly improved in 2013 using data from the 1000 Genome Project to provide GRCh38/Hg38. Data derived before that data, including the TCGA data has been converted to the later genomic coordinate system to improve both

methylation and mutation data. However, GRCh38/Hg38 will not be the final coordinate system. A further human reference assembly is in development, and new models are being evaluated[303].

I also rely heavily on information from the cancer gene census to identify candidate genes for tumour suppression [105]. This is a work in progress, and is updated as new evidence comes to light.

Annotation techniques are also improving greatly, but are still not ideal. In particular, COSMIC's mutation database v90 was introduced in September 2019. The new version reannotates the cosmic data introducing new mutation identifiers and a more standardised representation of the variants. This helps reduce problems with multiple copies of the same mutation. However, many still remain. A particular problem is that many mutations are mapped to multiple transcripts. For example, sample PD4107a in breast cancer has 174 entries at the same genomic location, mapped to 87 different transcripts. The TCGA data also has some accompanying basic medical data, but much of it is not complete, and this too is an area where standardisation would be very helpful.

This thesis relies heavily on a number of other statistical methods, sometimes explicit, such as non-negative matrix factorisation and at other times embedded within experimental methods such as mutation calling. These are not perfect. For example, the methods for calling indels are not always reliable, particularly next to poly-nucleotide repeats. This raises the possibility that some of the excess of inframe indels next to poly-nucleotide repeats may be an artefact of indel-calling[304]. However, this should not affect the evidence of selective

pressure against frame-shift indels as this work specifically looked at those indels not next to repeats. More importantly, multi-nucleotide indels are a relatively uncommon form of mutation. As the number of cancer samples grows so it should be possible to repeat this analysis on the different tissue types.

I have used the FATHMM column in order to assess which mutations are damaging. Mutations are classified as Pathogenic in COSMIC if the FATHMM score is 0.7 or over. However, FATHMM is itself a predictive tool and scores over 0.5 may be deleterious. Calling deleterious mutations is an unsolved problem, with many alternative options available including SIFT[305], PolyPhen[217], MutationAssessor[306], MutationTaster2[307] and Provean[308]. Although the methods vary they all depend on the ground truth of variants explored by experimental methods, included in collections such as 1000 Genomes[309], ClinVar[307], and Humsavar[310]. These databases are a really useful resource draw together many different experimental methods but can be assumed to be better at identifying seriously deleterious variants than at identifying those that are marginal. It is disappointing that there is no equivalent to the Protein Data Bank for providing such information.

Similarly, I have used the Broad Institute's analysis of correlations between methylation levels and gene expression to predict which genes are switched off. Again, this is an indirect method of predicting gene inactivation. This work would be much improved by data that more directly enabled identification of stable reductions in protein expression levels. Finally, there is a lack of experimental evidence associating changes in bacterial DNA with specific agents.

6.3 Future work

In chapter 4, I identified potential bacterial signatures on the basis of variations in the genes seen in multiple bacterial strains and proposed potential aetiologies derived by comparison with existing mutational signatures from cancer samples or samples that have been derived from exposure to known environmental mutagens. However, no experimental work has been done to confirm these signatures. I would like to collaborate with experimentalists to take the same approach to the bacterial signatures, systematically exposing bacteria to known mutagenic stimuli and then identifying the mutagenic signatures. Doing this, together with an analysis of the DNA damage repair genes present in different bacteria, would enable comparisons of the impact of the same environmental mutagens in both human and bacterial cells providing an insight into the impact of the different DNA damage repair systems at work.

I would also like to follow up work in chapter 6 to identify potential therapeutic drugs whose use could potentially be extended to provide therapies for patients with inactive tumour suppressor genes. The three main bioinformatic areas to this work I see as being: limiting the consideration of drugs to those with highly specific gene targets in order to reduce predictable side-effects; further analysis of the likely impact of missense mutations in order to distinguish between those that might cause loss of function or gain of function, and refinement of methylation correlations with gene expression in order to ignore those where correlations are likely to be spurious. With sufficient data is should also be able to look at the potential impact of drug combinations, by looking for groups of 3 genes that together have mutually exclusive inactivations.

Bibliography

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA. Cancer J. Clin.*, 2018.

[2] S. Bin Zaman, M. A. Hussain, R. Nye, V. Mehta, K. T. Mamun, and N. Hossain, "A Review on Antibiotic Resistance: Alarm Bells are Ringing," *Cureus*, 2017.

[3] R. Guigo, *Genetic Databases - Biological Technique Series*. 1997.

[4] S. Nik-Zainal *et al.*, "Mutational processes molding the genomes of 21 breast cancers," *Cell*, 2012.

[5] J. D. Watson and F. H. C. Crick, "Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid," *Nature*, 1953.

[6] M. E. Aldrup-Macdonald and B. A. Sullivan, "The past, present, and future of human centromere genomics.," *Genes (Basel).*, 2014.

[7] M. Méchali, "Eukaryotic DNA replication origins: Many choices for appropriate answers," *Nature Reviews Molecular Cell Biology*. 2010.

[8] J. R. Raab and R. T. Kamakaka, "Insulators and promoters: Closer than we think," *Nature Reviews Genetics*. 2010.

[9] R. De Bont and N. van Larebeke, "Endogenous DNA damage in humans: A review of quantitative data," *Mutagenesis*. 2004.

[10] G. Scherer *et al.*, "Determination of methyl-, 2-hydroxyethyl- and 2-cyanoethylmercapturic acids as biomarkers of exposure to alkylating agents in cigarette smoke," *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.*, 2010.

[11] E. C. Minca and D. Kowalski, "Replication fork stalling by bulky DNA damage:

Localization at active origins and checkpoint modulation," *Nucleic Acids Res.*, 2011.

[12] S. Maloy and K. Hughes, *Brenner's Encyclopedia of Genetics: Second Edition*. 2013.

[13] M. Ye, J. Beach, J. W. Martin, and A. Senthilselvan, "Occupational pesticide exposures and respiratory health," *Int. J. Environ. Res. Public Health*, 2013.

[14] R. P. Rastogi, Richa, A. Kumar, M. B. Tyagi, and R. P. Sinha, "Molecular mechanisms of ultraviolet radiation-induced DNA damage and repair," *Journal of Nucleic Acids*. 2010.

[15] G. Borrego-Soto, R. Ortiz-López, and A. Rojas-Martínez, "Ionizing radiation-induced DNA injury and damage detection in patients with breast cancer," *Genetics and Molecular Biology*. 2015.

[16] Z. Msiska, M. Pacurari, A. Mishra, S. S. Leonard, V. Castranova, and V. Vallyathan, "DNA double-strand breaks by asbestos, silica, and titanium dioxide: Possible biomarker of carcinogenic potential?," *Am. J. Respir. Cell Mol. Biol.*, 2010.

[17] P. Rous, "A sarcoma of the fowl transmissible by an agent separable from the tumor cells," *J. Exp. Med.*, 1911.

[18] C. Münz, "Latency and lytic replication in Epstein–Barr virus-associated oncogenesis," *Nature Reviews Microbiology*. 2019.

[19] K. K. Aneja and Y. Yuan, "Reactivation and lytic replication of Kaposi's sarcoma-associated herpesvirus: An update," *Frontiers in Microbiology*. 2017.

[20] P. Rous, "A transmissible avian neoplasm. (sarcoma of the common fowl.)," *J. Exp. Med.*, 1910.

[21] J. J. Champoux, "DNA topoisomerases: Structure, function, and mechanism," *Annual Review of Biochemistry*. 2001.

[22] L. F. Liu, C. C. Liu, and B. M. Alberts, "Type II DNA topoisomerases: Enzymes that can unknot a topologically knotted DNA molecule via a reversible double-strand break,"

Cell, 1980.

[23] B. Alberts *et al.*, *Molecular Biology of the Cell (Sixth Edition)*. Garland Science, 2015.

[24] J. E. Barrick *et al.*, “Genome evolution and adaptation in a long-term experiment with Escherichia coli,” *Nature*, 2009.

[25] R. G. EAGON, “Pseudomonas natriegens, a marine bacterium with a generation time of less than 10 minutes.,” *J. Bacteriol.*, 1962.

[26] L. B. Alexandrov *et al.*, “Signatures of mutational processes in human cancer,” *Nature*, 2013.

[27] Y. Mishina, E. M. Duguid, and C. He, “Direct reversal of DNA alkylation damage,” *Chemical Reviews*. 2006.

[28] N. Goosen and G. F. Moolenaar, “Repair of UV damage in bacteria,” *DNA Repair*. 2008.

[29] C. Kisker, J. Kuper, and B. Van Houten, “Prokaryotic nucleotide excision repair,” *Cold Spring Harb. Perspect. Biol.*, 2013.

[30] G. M. Li, “Mechanisms and functions of DNA mismatch repair,” *Cell Research*. 2008.

[31] I. Georgakopoulos-Soares, G. Koh, S. E. Momen, J. Jiricny, M. Hemberg, and S. Nik-Zainal, “Transcription-coupled repair and mismatch repair contribute towards preserving genome integrity at mononucleotide repeat tracts,” *Nat. Commun.*, 2020.

[32] J. R. Chapman, M. R. G. Taylor, and S. J. Boulton, “Playing the End Game: DNA Double-Strand Break Repair Pathway Choice,” *Molecular Cell*. 2012.

[33] E. P. C. Rocha, E. Cornet, and B. Michel, “Comparative and evolutionary analysis of the bacterial homologous recombination systems,” *PLoS Genetics*. 2005.

[34] B. J. Sishc and A. J. Davis, “The role of the core non-homologous end joining factors in carcinogenesis and cancer,” *Cancers*. 2017.

[35] M. McVey and S. E. Lee, “MMEJ repair of double-strand breaks (director’s cut): deleted sequences and alternative endings,” *Trends in Genetics*. 2008.

[36] G. J. McKenzie, R. S. Harris, P. L. Lee, and S. M. Rosenberg, “The SOS response regulates adaptive mutation,” *Proc. Natl. Acad. Sci. U. S. A.*, 2000.

[37] D. Žgur-Bertok, “DNA Damage Repair and Bacterial Pathogens,” *PLoS Pathog.*, 2013.

[38] T. D. Halazonetis, V. G. Gorgoulis, and J. Bartek, “An oncogene-induced DNA damage model for cancer development,” *Science*. 2008.

[39] L. H. Pearl, A. C. Schierz, S. E. Ward, B. Al-Lazikani, and F. M. G. Pearl, “Therapeutic opportunities within the DNA damage response,” *Nat. Rev. Cancer*, 2015.

[40] N. V. Volkova *et al.*, “Mutational signatures are jointly shaped by DNA damage and repair,” *Nat. Commun.*, 2020.

[41] P. Von Morgen and J. Maciejowski, “The ins and outs of telomere crisis in cancer,” *Genome Medicine*. 2018.

[42] C. P. Wild, “The global cancer burden: necessity is the mother of prevention,” *Nature Reviews Cancer*. 2019.

[43] S. Hossen, M. K. Hossain, M. K. Basher, M. N. H. Mia, M. T. Rahman, and M. J. Uddin, “Smart nanocarrier-based drug delivery systems for cancer therapy and toxicity studies: A review,” *Journal of Advanced Research*. 2019.

[44] N. Iqbal and N. Iqbal, “Imatinib: A Breakthrough of Targeted Therapy in Cancer,” *Chemother. Res. Pract.*, 2014.

[45] Q. Jiao, L. Bi, Y. Ren, S. Song, Q. Wang, and Y. shan Wang, “Advances in studies of tyrosine kinase inhibitors and their acquired resistance,” *Molecular Cancer*. 2018.

[46] A. Jameera Begam, S. Jubie, and M. J. Nanjan, “Estrogen receptor agonists/antagonists in breast cancer therapy: A critical review,” *Bioorganic*

Chemistry. 2017.

[47] J. E. Delmore *et al.*, "BET bromodomain inhibition as a therapeutic strategy to target c-Myc," *Cell*, 2011.

[48] S. J. Oiseth and M. S. Aziz, "Cancer immunotherapy: a brief review of the history, possibilities, and challenges ahead," *J. Cancer Metastasis Treat.*, 2017.

[49] I. Martincorena *et al.*, "High burden and pervasive positive selection of somatic mutations in normal human skin," *Science.*, 2015.

[50] C. Tomasetti and B. Vogelstein, "Variation in cancer risk among tissues can be explained by the number of stem cell divisions," *Science.*, 2015.

[51] R. A. Weinberg, *Biology of the Cancer*. 2014.

[52] E. S. Lander *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, 2001.

[53] D. Hanahan and R. A. Weinberg, "The hallmarks of cancer," *Cell*. 2000.

[54] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: The next generation," *Cell*. 2011.

[55] M. A. Davies and Y. Samuels, "Analysis of the genome to personalize therapy for melanoma," *Oncogene*. 2010.

[56] B. H. Jiang and L. Z. Liu, "PI3K/PTEN signaling in tumorigenesis and angiogenesis," *Biochimica et Biophysica Acta - Proteins and Proteomics*. 2008.

[57] D. L. Burkhart and J. Sage, "Cellular mechanisms of tumour suppression by the retinoblastoma gene," *Nature Reviews Cancer*. 2008.

[58] G. M. Wahl, S. P. Linke, T. G. Paulson, and L. C. Huang, "Maintaining genetic stability through TP53 mediated checkpoint control," *Cancer Surveys*. 1997.

[59] C. Giacinti and A. Giordano, "RB and cell cycle progression," *Oncogene*. 2006.

[60] J. T. Zilfou and S. W. Lowe, “Tumor suppressive functions of p53.,” *Cold Spring Harbor perspectives in biology*. 2009.

[61] S. Heerboth *et al.*, “EMT and tumor metastasis,” *Clin. Transl. Med.*, 2015.

[62] Y. Yang, H. Zheng, Y. Zhan, and S. Fan, “An emerging tumor invasion mechanism about the collective cell migration,” *American Journal of Translational Research*. 2019.

[63] P. Friedl and K. Wolf, “Tumour-cell invasion and migration: Diversity and escape mechanisms,” *Nature Reviews Cancer*. 2003.

[64] M. A. Jafri, S. A. Ansari, M. H. Alqahtani, and J. W. Shay, “Roles of telomeres and telomerase in cancer, and advances in telomerase-targeted therapies,” *Genome Medicine*. 2016.

[65] M. R. Junttila and G. I. Evan, “P53 a Jack of all trades but master of none,” *Nature Reviews Cancer*. 2009.

[66] S. W. Lowe, E. Cepero, and G. Evan, “Intrinsic tumour suppression,” *Nature*. 2004.

[67] Y. Shan *et al.*, “Targeting HIBCH to reprogram valine metabolism for the treatment of colorectal cancer,” *Cell Death Dis.*, 2019.

[68] C. W. Yun and S. H. Lee, “The roles of autophagy in cancer,” *International Journal of Molecular Sciences*. 2018.

[69] S. Y. Lee *et al.*, “Regulation of Tumor Progression by Programmed Necrosis,” *Oxidative Medicine and Cellular Longevity*. 2018.

[70] P. Carmeliet and R. K. Jain, “Angiogenesis in cancer and other diseases,” *Nature*. 2000.

[71] M. G. V. Heiden, L. C. Cantley, and C. B. Thompson, “Understanding the warburg effect: The metabolic requirements of cell proliferation,” *Science*. 2009.

[72] M. Murata, “Inflammation and cancer,” *Environmental Health and Preventive Medicine*. 2018.

[73] L. H. Pearl, A. C. Schierz, S. E. Ward, B. Al-Lazikani, and F. M. G. Pearl, “Therapeutic opportunities within the DNA damage response,” *Nat. Rev. Cancer*, vol. 15, no. 3, pp. 166–180, 2015.

[74] V. Thorsson *et al.*, “The Immune Landscape of Cancer,” *Immunity*, 2018.

[75] A. Tubbs and A. Nussenzweig, “Endogenous DNA Damage as a Source of Genomic Instability in Cancer,” *Cell*. 2017.

[76] T. J. Mitchell *et al.*, “Timing the Landmark Events in the Evolution of Clear Cell Renal Cell Cancer: TRACERx Renal,” *Cell*, 2018.

[77] X. Dai *et al.*, “Breast cancer intrinsic subtype classification, clinical use and future trends,” *American Journal of Cancer Research*. 2015.

[78] C. Kandoth *et al.*, “Mutational landscape and significance across 12 major cancer types,” *Nature*, 2013.

[79] B. Vogelstein and K. W. Kinzler, “The multistep nature of cancer,” *Trends in Genetics*. 1993.

[80] M. Uhlén *et al.*, “Tissue-based map of the human proteome,” *Science.*, 2015.

[81] A. Tsherniak *et al.*, “Defining a Cancer Dependency Map,” *Cell*, vol. 170, no. 3, pp. 564-576.e16, 2017.

[82] R. G. H. Lindeboom, F. Supek, and B. Lehner, “The rules and impact of nonsense-mediated mRNA decay in human cancers,” *Nat. Genet.*, 2016.

[83] P. A. Ascierto *et al.*, “The role of BRAF V600 mutation in melanoma,” *Journal of Translational Medicine*. 2012.

[84] C. Kandoth *et al.*, “Mutational landscape and significance across 12 major cancer

types," *Nature*, 2013.

[85] P. A. Jones, J. P. J. Issa, and S. Baylin, "Targeting the cancer epigenome for therapy," *Nature Reviews Genetics*. 2016.

[86] S. Seisenberger, J. R. Peat, T. A. Hore, F. Santos, W. Dean, and W. Reik, "Reprogramming DNA methylation in the mammalian life cycle: Building and breaking epigenetic barriers," *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2013.

[87] E. Schafer *et al.*, "Promoter hypermethylation in MLL-r infant acute lymphoblastic leukemia: Biology and therapeutic targeting," *Blood*, 2010.

[88] Z. Sondka, S. Bamford, C. G. Cole, S. A. Ward, I. Dunham, and S. A. Forbes, "The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers," *Nature Reviews Cancer*. 2018.

[89] M. A. Haidar *et al.*, "ATM gene deletion in patients with adult acute lymphoblastic leukemia," *Cancer*, 2000.

[90] A. Maréchal and L. Zou, "DNA damage sensing by the ATM and ATR kinases," *Cold Spring Harb. Perspect. Biol.*, 2013.

[91] J. Li *et al.*, "PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer," *Science.*, 1997.

[92] G. M. Cooper and R. E. Hausman, *The Cell: A Molecular Approach 2nd Edition*. 2007.

[93] J. S. Ross *et al.*, "Comprehensive genomic profiling of epithelial ovarian cancer by next generation sequencing-based diagnostic assay reveals new routes to targeted therapies," *Gynecol. Oncol.*, 2013.

[94] M. Ehrlich, "DNA methylation in cancer: Too much, but also too little," *Oncogene*. 2002.

[95] S. Wong and O. N. Witte, “The BCR-ABL story: Bench to bedside and back,” *Annual Review of Immunology*. 2004.

[96] N. Iqbal and N. Iqbal, “Human Epidermal Growth Factor Receptor 2 (HER2) in Cancers: Overexpression and Therapeutic Implications,” *Mol. Biol. Int.*, 2014.

[97] *Holland-Frei Cancer Medicine*. 2016.

[98] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler, “Cancer genome landscapes,” *Science*, vol. 340, no. 6127. pp. 1546–1558, 2013.

[99] T. Soussi and K. G. Wiman, “TP53: An oncogene in disguise,” *Cell Death and Differentiation*. 2015.

[100] S. Nik-Zainal and S. Morganella, “Mutational signatures in breast cancer: The problem at the DNA level,” *Clin. Cancer Res.*, 2017.

[101] L. B. Alexandrov and M. R. Stratton, “Mutational signatures: The patterns of somatic mutations hidden in cancer genomes,” *Current Opinion in Genetics and Development*. 2014.

[102] L. B. Alexandrov *et al.*, “Signatures of mutational processes in human cancer,” *Nature*, vol. 500, no. 7463, pp. 415–421, 2013.

[103] S. Nik-Zainal *et al.*, “Landscape of somatic mutations in 560 breast cancer whole-genome sequences,” *Nature*, 2016.

[104] L. B. Alexandrov *et al.*, “Clock-like mutational processes in human somatic cells,” *Nat. Genet.*, 2015.

[105] Z. Sondka, S. Bamford, C. G. Cole, S. A. Ward, I. Dunham, and S. A. Forbes, “The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers,” *Nature Reviews Cancer*. 2018.

[106] R. Mani, R. P. St. Onge, J. L. Hartman IV, G. Giaever, and F. P. Roth, “Defining genetic

interaction," *Proc. Natl. Acad. Sci. U. S. A.*, 2008.

[107] A. Huang, L. A. Garraway, A. Ashworth, and B. Weber, "Synthetic lethality as an engine for cancer drug target discovery," *Nature Reviews Drug Discovery*. 2020.

[108] C. Underhill, M. Toulmonde, and H. Bonnefoi, "A review of PARP inhibitors: From bench to bedside," *Annals of Oncology*. 2011.

[109] S. V. Gordon and T. Parish, "Microbe profile: Mycobacterium tuberculosis: Humanity's deadly microbial foe," *Microbiol. (United Kingdom)*, 2018.

[110] E. Tacconelli *et al.*, "Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis," *Lancet Infect. Dis.*, 2018.

[111] A. D. Yates *et al.*, "Ensembl 2020," *Nucleic Acids Res.*, 2020.

[112] Y. Uncu, G. Uncu, A. Esmer, and N. Bilgel, "Should asymptomatic bacteriuria be screened in pregnancy?," *Clin. Exp. Obstet. Gynecol.*, 2002.

[113] E. Nagy, "What do we know about the diagnostics, treatment and epidemiology of Clostridioides (Clostridium) difficile infection in Europe?," *Journal of Infection and Chemotherapy*. 2018.

[114] S. V. Lynch, "The lung microbiome and airway disease," in *Annals of the American Thoracic Society*, 2016.

[115] D. Bogaert, R. De Groot, and P. W. M. Hermans, "Streptococcus pneumoniae colonisation: The key to pneumococcal disease," *Lancet Infectious Diseases*. 2004.

[116] D. A. Caugant *et al.*, "Asymptomatic carriage of Neisseria meningitidis in a randomly sampled population," *J. Clin. Microbiol.*, 1994.

[117] S. A. Rahman *et al.*, "Comparative Analyses of Nonpathogenic, Opportunistic, and Totally Pathogenic Mycobacteria Reveal Genomic and Biochemical Variabilities and

Highlight the Survival Attributes of Mycobacterium tuberculosis," *MBio*, 2014.

[118] D. M. Tobiason and H. S. Seifert, "The obligate human pathogen, Neisseria gonorrhoeae, is polyploid," *PLoS Biol.*, 2006.

[119] L. P. Stenfors Arnesen, A. Fagerlund, and P. E. Granum, "From soil to gut: Bacillus cereus and its food poisoning toxins," *FEMS Microbiology Reviews*. 2008.

[120] J. C. Low and W. Donachie, "A review of Listeria monocytogenes and listeriosis," *Veterinary Journal*. 1997.

[121] M. M. Johnson and J. A. Odell, "Nontuberculous mycobacterial pulmonary infections," *Journal of Thoracic Disease*. 2014.

[122] T. M. Pham, M. Kretzschmar, X. Bertrand, and M. Bootsma, "Tracking Pseudomonas aeruginosa transmissions due to environmental contamination after discharge in ICUs using mathematical models," *PLoS Comput. Biol.*, 2019.

[123] K. C. Malcolm *et al.*, "Mycobacterium abscessus displays fitness for fomite transmission," *Appl. Environ. Microbiol.*, 2017.

[124] W. J. Wiersinga *et al.*, "Melioidosis," *Nat. Rev. Dis. Prim.*, 2018.

[125] A. Kramer, I. Schwebke, and G. Kampf, "How long do nosocomial pathogens persist on inanimate surfaces? A systematic review," *BMC Infectious Diseases*. 2006.

[126] C. Smillie, M. P. Garcillán-Barcia, M. V. Francia, E. P. C. Rocha, and F. de la Cruz, "Mobility of Plasmids," *Microbiol. Mol. Biol. Rev.*, 2010.

[127] M. Land *et al.*, "Insights from 20 years of bacterial genome sequencing," *Functional and Integrative Genomics*. 2015.

[128] F. D. Ciccarelli, T. Doerks, C. Von Mering, C. J. Creevey, B. Snel, and P. Bork, "Toward automatic reconstruction of a highly resolved tree of life," *Science.*, 2006.

[129] K. L. Howe *et al.*, "Ensembl Genomes 2020-enabling non-vertebrate genomic

research," *Nucleic Acids Res.*, 2020.

[130] Y. Fang *et al.*, "Complete genome sequence of Acinetobacter baumannii XH386 (ST208), a multi-drug resistant bacteria isolated from pediatric hospital in China," *Genomics Data*, 2016.

[131] N. Ivanova *et al.*, "Genome sequence of Bacillus cereus and comparative analysis with Bacillus anthracis," *Nature*, 2003.

[132] O. O. Bochkareva, E. V. Moroz, I. I. Davydov, and M. S. Gelfand, "Genome rearrangements and selection in multi-chromosome bacteria Burkholderia spp.," *BMC Genomics*, 2018.

[133] E. S. Egan, M. A. Fogel, and M. K. Waldor, "Divided genomes: Negotiating the cell cycle in prokaryotes with multiple chromosomes," *Molecular Microbiology*. 2005.

[134] D. R. Knight, B. Elliott, B. J. Chang, T. T. Perkins, and T. V. Riley, "Diversity and evolution in the genome of clostridium difficile," *Clin. Microbiol. Rev.*, 2015.

[135] R. W. Purbojati *et al.*, "Complete Genome Sequence of Enterococcus faecalis Strain SGAir0397, Isolated from a Tropical Air Sample Collected in Singapore," *Microbiol. Resour. Announc.*, 2019.

[136] M. M. C. Lam *et al.*, "Comparative analysis of the first complete Enterococcus faecium genome," *J. Bacteriol.*, 2012.

[137] R. Reyes-Lamothe, X. Wang, and D. Sherratt, "Escherichia coli and its chromosome," *Trends in Microbiology*. 2008.

[138] A. C. Lin, T. L. Liao, Y. C. Lin, Y. C. Lai, M. C. Lu, and Y. T. Chen, "Complete genome sequence of Klebsiella pneumoniae 1084, a hypermucoviscosity-negative K1 clinical strain," *Journal of Bacteriology*. 2012.

[139] E. Michel and P. Cossart, "Physical map of the Listeria monocytogenes chromosome,"

J. Bacteriol., 1992.

[140] M. Sassi and M. Drancourt, “Genome analysis reveals three genomospecies in Mycobacterium abscessus,” *BMC Genomics*, 2014.

[141] I. Smith, “Mycobacterium tuberculosis pathogenesis and molecular determinants of virulence,” *Clinical Microbiology Reviews*. 2003.

[142] C. Schoen, H. Tettelin, J. Parkhill, and M. Frosch, “Genome flexibility in Neisseria meningitidis,” *Vaccine*, 2009.

[143] J. Klockgether, N. Cramer, L. Wiehlmann, C. F. Davenport, and B. Tümmler, “Pseudomonas aeruginosa genomic structure and diversity,” *Front. Microbiol.*, 2011.

[144] S. Baker and G. Dougan, “The genome of Salmonella enterica serovar typhi,” *Clinical Infectious Diseases*. 2007.

[145] J. Hoskins *et al.*, “Genome of the bacterium Streptococcus pneumoniae strain R6,” *J. Bacteriol.*, 2001.

[146] GDAC, “Correlation between mRNA expression and DNA methylation.,” *Harvard, Broad Institute of MIT and*, 2016. [Online]. Available: http://firebrowse.org/?cohort=THCA#.

[147] S. A. Forbes *et al.*, “COSMIC: Somatic cancer genetics at high-resolution,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D777–D783, 2017.

[148] K. C. Cotto *et al.*, “DGIdb 3.0: a redesign and expansion of the drug–gene interaction database,” *Nucleic Acids Res.*, 2017.

[149] S. E. Hunt *et al.*, “Ensembl variation resources,” *Database (Oxford).*, 2018.

[150] C. Printz, “Genomic Data Commons ushers in new era for information sharing,” *Cancer*, vol. 122, no. 18. pp. 2777–2778, 2016.

[151] D. Szklarczyk *et al.*, “STRING v11: Protein-protein association networks with increased

coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic Acids Res.*, 2019.

[152] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, P. J. Campbell, and M. R. Stratton, "Deciphering Signatures of Mutational Processes Operative in Human Cancer," *Cell Rep.*, 2013.

[153] M. Steinegger and J. Söding, "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets," *Nature Biotechnology*. 2017.

[154] F. Sievers *et al.*, "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega," *Mol. Syst. Biol.*, 2011.

[155] M. McVey and S. E. Lee, "MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings," *Trends in Genetics*, vol. 24, no. 11. pp. 529–538, 2008.

[156] J. M. Chen, N. Chuzhanova, P. D. Stenson, C. Férec, and D. N. Cooper, "Meta-analysis of gross insertions causing human genetic disease: Novel mutational mechanisms and the role of replication slippage," *Hum. Mutat.*, 2005.

[157] V. Marx, "Cancer genomes: Discerning drivers from passengers," *Nat. Methods*, vol. 11, no. 4, pp. 375–379, 2014.

[158] C. D. McFarland, K. S. Korolev, G. V. Kryukov, S. R. Sunyaev, and L. A. Mirny, "Impact of deleterious passenger mutations on cancer progression," *Proc. Natl. Acad. Sci.*, vol. 110, no. 8, pp. 2910–2915, 2013.

[159] S. F. Bakhoum and D. A. Landau, "Cancer Evolution: No Room for Negative Selection," *Cell*, vol. 171, no. 5. pp. 987–989, 2017.

[160] S. Nik-Zainal *et al.*, "The genome as a record of environmental exposure," *Mutagenesis*, vol. 30, no. 6, pp. 763–770, 2015.

[161] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, P. J. Campbell, and M. R. Stratton, "Deciphering Signatures of Mutational Processes Operative in Human Cancer," *Cell Rep.*, vol. 3, no. 1, pp. 246–259, 2013.

[162] B. Hang, "Formation and Repair of Tobacco Carcinogen-Derived Bulky DNA Adducts," *J. Nucleic Acids*, vol. 2010, pp. 1–29, 2010.

[163] G. Waris and H. Ahsan, "Reactive oxygen species: Role in the development of cancer and various chronic conditions," *Journal of Carcinogenesis*, vol. 5. 2006.

[164] P. Iengar, "An analysis of substitution, deletion and insertion mutations in cancer genes," *Nucleic Acids Res.*, vol. 40, no. 14, pp. 6401–6413, 2012.

[165] D. Glodzik *et al.*, "A somatic-mutational process recurrently duplicates germline susceptibility loci and tissue-specific super-enhancers in breast cancers," *Nat. Genet.*, vol. 49, no. 3, pp. 341–348, 2017.

[166] L. B. Alexandrov *et al.*, "Clock-like mutational processes in human somatic cells," *Nat. Genet.*, vol. 47, no. 12, pp. 1402–1407, 2015.

[167] S. Nik-Zainal *et al.*, "Mutational processes molding the genomes of 21 breast cancers," *Cell*, vol. 149, no. 5, pp. 979–993, 2012.

[168] N. Nagarajan *et al.*, "Whole-genome reconstruction and mutational signatures in gastric cancer," *Genome Biol.*, 2012.

[169] L. B. Alexandrov *et al.*, "Mutational signatures associated with tobacco smoking in human cancer," *Science.*, 2016.

[170] E. Viguera, D. Canceill, and S. D. Ehrlich, "Replication slippage involves DNA polymerase pausing and dissociation," *EMBO J.*, vol. 20, no. 10, pp. 2587–2595, 2001.

[171] S. Clancy, "DNA Damage & Repair: Mechanisms for Maintaining DNA Integrity," *Nat. Educ.*, vol. 1, no. 1, p. 103, 2008.

[172] L. E. Maquat, “When cells stop making sense: effects of nonsense codons on RNA metabolism in vertebrate cells.,” *Rna*, 1995.

[173] T. Helleday, S. Eshtad, and S. Nik-Zainal, “Mechanisms underlying mutational signatures in human cancers,” *Nature Reviews Genetics*. 2014.

[174] D. R. Zerbino *et al.*, “Ensembl 2018,” *Nucleic Acids Res.*, 2017.

[175] P. J. A. Cock *et al.*, “Biopython: Freely available Python tools for computational molecular biology and bioinformatics,” *Bioinformatics*, 2009.

[176] E. Jones, T. Oliphant, P. Peterson, and others, “SciPy: Open source scientific tools for Python,” *Computing in Science and Engineering*. 2007.

[177] Y.-F. Chang, J. S. Imam, and M. F. Wilkinson, “The Nonsense-Mediated Decay RNA Surveillance Pathway,” *Annu. Rev. Biochem.*, vol. 76, no. 1, pp. 51–74, 2007.

[178] D. Martin *et al.*, “The head and neck cancer cell oncogenome : a platform for the development of precision molecular therapies,” *Oncotarget*, 2014.

[179] L. B. Alexandrov, S. Nik-Zainal, H. C. Siu, S. Y. Leung, and M. R. Stratton, “A mutational signature in gastric cancer suggests therapeutic strategies,” *Nat. Commun.*, 2015.

[180] C. J. Walker *et al.*, “MonoSeq Variant Caller Reveals Novel Mononucleotide Run Indel Mutations in Tumors with Defective DNA Mismatch Repair,” *Hum. Mutat.*, 2016.

[181] M. Spies and R. Fishel, “Mismatch repair during homologous and homeologous recombination,” *Cold Spring Harb. Perspect. Biol.*, 2015.

[182] M. O. Raeker, J. Pierre-Charles, and J. M. Carethers, “Abstract 3368: Insertion and deletion frameshift rates and mutational spectra of tetranucleotide microsatellites in DNA mismatch repair-deficient human cells,” *Cancer Res.*, 2018.

[183] H. Gragg, B. D. Harfe, and S. Jinks-Robertson, “Base composition of mononucleotide runs affects DNA polymerase slippage and removal of frameshift intermediates by

mismatch repair in Saccharomyces cerevisiae.," *Mol. Cell. Biol.*, vol. 22, no. 24, pp. 8756–8762, 2002.

[184] S. Matsumura, Y. Fujita, M. Yamane, O. Morita, and H. Honda, "A genome-wide mutation analysis method enabling high-throughput identification of chemical mutagen signatures," *Sci. Rep.*, 2018.

[185] L. B. Alexandrov *et al.*, "The repertoire of mutational signatures in human cancer," *Nature*, 2020.

[186] K. Chang *et al.*, "The Cancer Genome Atlas Pan-Cancer analysis project," *Nat. Genet.*, vol. 45, no. 10, pp. 1113–1120, 2013.

[187] P. J. Campbell *et al.*, "Pan-cancer analysis of whole genomes," *Nature*, 2020.

[188] K. A. Hoadley *et al.*, "Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin," *Cell*, 2014.

[189] Y. Chen *et al.*, "Identification of druggable cancer driver genes amplified across TCGA datasets," *PLoS One*, 2014.

[190] J. J. Bianchi, X. Zhao, J. C. Mays, and T. Davoli, "Not all cancers are created equal: Tissue specificity in cancer genes and pathways," *Current Opinion in Cell Biology*. 2020.

[191] S. Carreira *et al.*, "Tumor clone dynamics in lethal prostate cancer," *Sci. Transl. Med.*, 2014.

[192] M. Olivier, M. Hollstein, and P. Hainaut, "TP53 mutations in human cancers: origins, consequences, and clinical use.," *Cold Spring Harbor perspectives in biology*. 2010.

[193] L. Schmidt *et al.*, "Germline and somatic mutations in the tyrosine kinase domain of the MET proto-oncogene in papillary renal carcinomas," *Nat. Genet.*, 1997.

[194] Z. Yao *et al.*, "BRAF Mutants Evade ERK-Dependent Feedback by Different

Mechanisms that Determine Their Sensitivity to Pharmacologic Inhibition," *Cancer Cell*, 2015.

[195] A. Hodgkinson, Y. Chen, and A. Eyre-Walker, "The large-scale distribution of somatic mutations in cancer genomes," *Hum. Mutat.*, 2012.

[196] S. Shooter, J. Czarnecki, and S. Nik-Zainal, "Signal: The home page of mutational signatures," *Ann. Oncol.*, 2019.

[197] J. E. Kucab *et al.*, "A Compendium of Mutational Signatures of Environmental Agents," *Cell*, 2019.

[198] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, P. J. Campbell, and M. R. Stratton, "Deciphering Signatures of Mutational Processes Operative in Human Cancer," *Cell Rep.*, 2013.

[199] L. B. Alexandrov *et al.*, "Signatures of mutational processes in human cancer," *Nature*, 2013.

[200] P. A. Futreal *et al.*, "A census of human cancer genes," *Nature Reviews Cancer*, vol. 4, no. 3. pp. 177–183, 2004.

[201] H. A. Shihab *et al.*, "Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models.," *Predict. Funct. Mol. phenotypic consequences Amin. acid substitutions using hidden Markov Model.*, vol. 34, no. 1, pp. 57–65, 2013.

[202] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *J. R. Stat. Soc. Ser. B*, 1995.

[203] D. E. Brash, "UV signature mutations," *Photochemistry and Photobiology*. 2015.

[204] C. L. Chen *et al.*, "Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes," *Genome Res.*, 2010.

[205] F. Blokzijl *et al.*, “Tissue-specific mutation accumulation in human adult stem cells during life,” *Nature*, 2016.

[206] S. S. Wong *et al.*, “Genomic landscape and genetic heterogeneity in gastric adenocarcinoma revealed by whole-genome sequencing,” *Nat. Commun.*, 2014.

[207] P. Liu *et al.*, “Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing,” *Carcinogenesis*, 2012.

[208] A. J. Levine, “The many faces of p53: Something for everyone,” *Journal of Molecular Cell Biology*. 2019.

[209] M. Periyasamy *et al.*, “P53 controls expression of the DNA deaminase APOBEC3B to limit its potential mutagenic activity in cancer cells,” *Nucleic Acids Res.*, 2017.

[210] S. Henderson, A. Chakravarthy, X. Su, C. Boshoff, and T. R. Fenton, “APOBEC-Mediated Cytosine Deamination Links PIK3CA Helical Domain Mutations to Human Papillomavirus-Driven Tumor Development,” *Cell Rep.*, 2014.

[211] V. Menon and L. Povirk, “Involvement of p53 in the repair of DNA double strand breaks: Multifaceted roles of p53 in homologous recombination repair (HRR) and non-homologous end joining (NHEJ),” *Subcell. Biochem.*, 2014.

[212] D. Xiao, F. Li, H. Pan, H. Liang, K. Wu, and J. He, “Integrative analysis of genomic sequencing data reveals higher prevalence of LRP1B mutations in lung adenocarcinoma patients with COPD,” *Sci. Rep.*, 2017.

[213] P. Brachova, K. W. Thiel, and K. K. Leslie, “The consequence of oncomorphic TP53 mutations in ovarian cancer,” *International Journal of Molecular Sciences*. 2013.

[214] L. M. Cortez *et al.*, “APOBEC3A is a prominent cytidine deaminase in breast cancer,” *PLoS Genet.*, 2019.

[215] F. J. Núñez *et al.*, “IDH1-R132H acts as a tumor suppressor in glioma via epigenetic

up-regulation of the DNA damage response," *Sci. Transl. Med.*, 2019.

[216] H. A. Shihab *et al.*, "Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models," *Hum. Mutat.*, 2013.

[217] I. Adzhubei, D. M. Jordan, and S. R. Sunyaev, "Predicting functional effect of human missense mutations using PolyPhen-2," *Curr. Protoc. Hum. Genet.*, no. SUPPL.76, 2013.

[218] M. L. Turski *et al.*, "Genomically driven tumors and actionability across histologies: BRAF-mutant cancers as a paradigm," *Molecular Cancer Therapeutics*. 2016.

[219] M. G. Lepre, S. I. Omar, G. Grasso, U. Morbiducci, M. A. Deriu, and J. A. Tuszynski, "Insights into the effect of the G245S single point mutation on the structure of p53 and the binding of the protein to DNA," *Molecules*, 2017.

[220] M. B. Olszewski, M. Pruszko, E. Snaar-Jagalska, A. Zylicz, and M. Zylicz, "Diverse and cancer type-specific roles of the p53 R248Q gain-of-function mutation in cancer migration and invasiveness," *Int. J. Oncol.*, 2019.

[221] R. C. Poulos, Y. T. Wong, R. Ryan, H. Pang, and J. W. H. Wong, "Analysis of 7,815 cancer exomes reveals associations between mutational processes and somatic driver mutations," *PLoS Genet.*, 2018.

[222] D. Temko, I. P. M. Tomlinson, S. Severini, B. Schuster-Böckler, and T. A. Graham, "The effects of mutational processes and selection on driver mutations across cancer types," *Nat. Commun.*, 2018.

[223] A. Karimian *et al.*, "Crosstalk between Phosphoinositide 3-kinase/Akt signaling pathway with DNA damage response and oxidative stress in cancer," *J. Cell. Biochem.*, 2019.

[224] S. Inoue *et al.*, “Mutant IDH1 Downregulates ATM and Alters DNA Repair and Sensitivity to DNA Damage Independent of TET2,” *Cancer Cell*, 2016.

[225] M. Petljak and L. B. Alexandrov, “Understanding mutagenesis through delineation of mutational signatures in human cancer,” *Carcinogenesis*, 2016.

[226] L. B. Alexandrov and M. R. Stratton, “Mutational signatures: The patterns of somatic mutations hidden in cancer genomes,” *Current Opinion in Genetics and Development*. 2014.

[227] L. B. Alexandrov *et al.*, “Clock-like mutational processes in human somatic cells,” *Nat. Genet.*, 2015.

[228] X. Didelot and M. C. J. Maiden, “Impact of recombination on bacterial evolution,” *Trends in Microbiology*. 2010.

[229] O. H. Ambur *et al.*, “Genome dynamics in major bacterial pathogens,” in *FEMS Microbiology Reviews*, 2009.

[230] D. R. Zeigler, “Gene sequences useful for predicting relatedness of whole genomes in bacteria,” *Int. J. Syst. Evol. Microbiol.*, 2003.

[231] J. H. Ward, “Hierarchical Grouping to Optimize an Objective Function,” *J. Am. Stat. Assoc.*, 1963.

[232] S. R. Santos and H. Ochman, “Identification and phylogenetic sorting of bacterial lineages with universally conserved genes and proteins,” *Environ. Microbiol.*, 2004.

[233] F. Marcille *et al.*, “Distribution of genes encoding the trypsin-dependent lantibiotic ruminococcin A among bacteria isolated from human fecal microbiota,” *Appl. Environ. Microbiol.*, 2002.

[234] N. Datta *et al.*, “Distribution of genes for trimethoprim and gentamicin resistance in bacteria and their plasmids in a general hospital,” *J. Gen. Microbiol.*, 1980.

[235] G. Lina, A. Quaglia, M. E. Reverdy, R. Leclercq, F. Vandenesch, and J. Etienne, "Distribution of genes encoding resistance to macrolides, lincosamides, and streptogramins among staphylococci," *Antimicrob. Agents Chemother.*, 1999.

[236] E. V. Koonin, K. S. Makarova, and L. Aravind, "Horizontal Gene Transfer in Prokaryotes: Quantification and Classification," *Annu. Rev. Microbiol.*, 2001.

[237] E. Stackebrandt and B. M. Goebel, "Taxonomic note: A place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology," *International Journal of Systematic Bacteriology*. 1994.

[238] S. Mignard and J. P. Flandrois, "16S rRNA sequencing in routine bacterial identification: A 30-month experiment," *J. Microbiol. Methods*, 2006.

[239] J. Chun and F. A. Rainey, "Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea," *International Journal of Systematic and Evolutionary Microbiology*. 2014.

[240] M. Kimura, "Evolutionary rate at the molecular level," *Nature*, 1968.

[241] D. Fu, J. A. Calvo, and L. D. Samson, "Balancing repair and tolerance of DNA damage caused by alkylating agents," *Nature Reviews Cancer*. 2012.

[242] S. D. Richardson, M. J. Plewa, E. D. Wagner, R. Schoeny, and D. M. DeMarini, "Occurrence, genotoxicity, and carcinogenicity of regulated and emerging disinfection by-products in drinking water: A review and roadmap for research," *Mutation Research - Reviews in Mutation Research*. 2007.

[243] M. Stiborová *et al.*, "Ellipticine cytotoxicity to cancer cell lines-a comparative study," *Interdiscip. Toxicol.*, 2011.

[244] S. Kim *et al.*, "PubChem 2019 update: Improved access to chemical data," *Nucleic Acids Res.*, 2019.

[245] S. M. Paul *et al.*, “How to improve RD productivity: The pharmaceutical industry’s grand challenge,” *Nature Reviews Drug Discovery*, vol. 9, no. 3. pp. 203–214, 2010.

[246] S. K. Wooller, G. Benstead-Hume, X. Chen, Y. Ali, and F. M. G. Pearl, “Bioinformatics in translational drug discovery,” *Biosci. Rep.*, vol. 37, no. 4, p. BSR20160180, 2017.

[247] S. Martino *et al.*, “The Role of Selective Estrogen Receptor Modulators in the Prevention of Breast Cancer: Comparison of the Clinical Trials,” *Oncologist*, 2004.

[248] A. Sharma, S. R. Shah, H. Illum, and J. Dowell, “Vemurafenib: Targeted inhibition of mutated BRAF for treatment of advanced melanoma and its potential in other malignancies,” *Drugs*. 2012.

[249] A. E. Maennling *et al.*, “Molecular targeting therapy against egfr family in breast cancer: Progress and future potentials,” *Cancers (Basel).*, 2019.

[250] M. Murtaza *et al.*, “Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA,” *Nature*, 2013.

[251] C. J. Lord and A. Ashworth, “PARP inhibitors: Synthetic lethality in the clinic,” *Science*, vol. 355, no. 6330. pp. 1152–1158, 2017.

[252]. -L Jpo@benevolent Ai *et al.*, “A comprehensive map of molecular drug targets,” *Nat. Rev. | DRUG Discov.*, vol. 16, 2017.

[253] C. J. Lord and A. Ashworth, “PARP inhibitors: Synthetic lethality in the clinic,” *Science*. 2017.

[254] N. J. O’Neil, M. L. Bailey, and P. Hieter, “Synthetic lethality and cancer,” *Nature Reviews Genetics*. 2017.

[255] A. Chatr-Aryamontri *et al.*, “The BioGRID interaction database: 2017 update,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D369–D379, 2017.

[256] a. H. Y. Tong, “Systematic Genetic Analysis with Ordered Arrays of Yeast Deletion

Mutants," *Science.*, vol. 294, no. 5550, pp. 2364–2368, 2001.

[257] A. H. Y. Tong and C. Boone, "Synthetic genetic array analysis in Saccharomyces cerevisiae.," *Methods Mol. Biol.*, 2006.

[258] B. Housden, H. Nicholson, and N. Perrimon, "Synthetic Lethality Screens Using RNAi in Combination with CRISPR-based Knockout in Drosophila Cells," *BIO-PROTOCOL*, 2017.

[259] S. M. B. Nijman, "Synthetic lethality: General principles, utility and detection using genetic screens in human cells," *FEBS Letters*. 2011.

[260] M. M. Martins *et al.*, "Linking tumor mutations to drug responses via a quantitative chemical–genetic interaction map," *Cancer Discov.*, 2015.

[261] A. S. L. Wong *et al.*, "Multiplexed barcoded CRISPR-Cas9 screening enabled by CombiGEM," *Proc. Natl. Acad. Sci. U. S. A.*, 2016.

[262] A. Dhoonmoon, E. M. Schleicher, K. E. Clements, C. M. Nicolae, and G. L. Moldovan, "Genome-wide CRISPR synthetic lethality screen identifies a role for the ADP-ribosyltransferase PARP14 in DNA replication dynamics controlled by ATR," *Nucleic Acids Res.*, 2020.

[263] F. M. Behan *et al.*, "Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens," *Nature*, 2019.

[264] J. L. Wilding and W. F. Bodmer, "Cancer cell lines for drug discovery and development," *Cancer Research*, vol. 74, no. 9. pp. 2377–2384, 2014.

[265] K. C. Chipman and A. K. Singh, "Predicting genetic interactions with random walks on biological networks," *BMC Bioinformatics*, 2009.

[266] G. Benstead-Hume, X. Chen, S. R. Hopkins, K. A. Lane, J. A. Downs, and F. M. G. Pearl, "Predicting synthetic lethal interactions using conserved patterns in protein interaction networks," *PLoS Comput. Biol.*, 2019.

[267] T. Kranthi, S. B. Rao, and P. Manimaran, “Identification of synthetic lethal pairs in biological systems through network information centrality,” *Mol. Biosyst.*, 2013.

[268] N. Conde-Pueyo, A. Munteanu, R. V. Solé, and C. Rodríguez-Caso, “Human synthetic lethal inference as potential anti-cancer target gene detection,” *BMC Syst. Biol.*, 2009.

[269] B. Vandersluis *et al.*, “Genetic interactions reveal the evolutionary trajectories of duplicate genes,” *Mol. Syst. Biol.*, 2010.

[270] E. N. Koch *et al.*, “Conserved rules govern genetic interaction degree across species,” *Genome Biol.*, 2012.

[271] X. Lu, P. R. Kensche, M. A. Huynen, and R. A. Notebaart, “Genome evolution predicts genetic interactions in protein complexes and reveals cancer drug targets,” *Nat. Commun.*, 2013.

[272] C. Pesquita, D. Faria, A. O. Falcão, P. Lord, and F. M. Couto, “Semantic similarity in biomedical ontologies,” *PLoS Computational Biology*. 2009.

[273] R. Hoehndorf, N. W. Hardy, D. Osumi-Sutherland, S. Tweedie, P. N. Schofield, and G. V. Gkoutos, “Systematic Analysis of Experimental Phenotype Data Reveals Gene Functions,” *PLoS One*, 2013.

[274] E. Remy, S. Rebouissou, C. Chaouiya, A. Zinovyev, F. Radvanyi, and L. Calzone, “A modeling approach to explain mutually exclusive and co-occurring genetic alterations in bladder tumorigenesis,” *Cancer Res.*, 2015.

[275] C. M. Blakely *et al.*, “Evolution and clinical impact of co-occurring genetic alterations in advanced-stage EGFR-mutant lung cancers,” *Nat. Genet.*, 2017.

[276] S. Srihari *et al.*, “Inferring synthetic lethal interactions from mutual exclusivity of genetic events in cancer.”

[277] S. Canisius, J. W. M. Martens, and L. F. A. Wessels, “A novel independence test for

somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence," *Genome Biol.*, 2016.

[278] M. D. M. Leiserson, H. T. Wu, F. Vandin, and B. J. Raphael, "CoMEt: A statistical approach to identify combinations of mutually exclusive alterations in cancer," *Genome Biol.*, vol. 16, no. 1, 2015.

[279] Ö. Babur *et al.*, "Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations," *Genome Biol.*, 2015.

[280] G. Ciriello, E. Cerami, C. Sander, and N. Schultz, "Mutual exclusivity analysis identifies oncogenic network modules," *Genome Res.*, 2012.

[281] J. R. Bradley and D. L. Farnsworth, "Testing for mutual exclusivity," *J. Appl. Stat.*, 2009.

[282] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge," *Wspolczesna Onkologia*. 2015.

[283] D. Sproul and R. R. Meehan, "Genomic insights into cancer-associated aberrant CpG island hypermethylation," *Brief. Funct. Genomics*, 2013.

[284] X. Ma, Y. W. Wang, M. Q. Zhang, and A. F. Gazdar, "DNA methylation data analysis and its application to cancer research," *Epigenomics*. 2013.

[285] S. B. Baylin and J. G. Herman, "DNA hypermethylation in tumorigenesis: Epigenetics joins genetics," *Trends in Genetics*. 2000.

[286] Y. Hong, "On computing the distribution function for the Poisson binomial distribution," *Comput. Stat. Data Anal.*, 2013.

[287] Z. Ji, C. Huo, and P. Yang, "Genistein inhibited the proliferation of kidney cancer cells via CDKN2a hypomethylation: role of abnormal apoptosis," *Int. Urol. Nephrol.*, 2020.

[288] Mei Pharma 2018 "ME-344 information," available at

https://www.meipharma.com/our-programs/me-344. (Accessed October 2020).

[289] S. C. Lim, K. T. Carey, and M. McKenzie, “Anti-cancer analogues ME-143 and ME-344 exert toxicity by directly inhibiting mitochondrial NADH: Ubiquinone oxidoreductase (Complex I),” *Am. J. Cancer Res.*, 2015.

[290] J. L. Kirkland and T. Tchkonia, “Senolytic Drugs: From Discovery to Translation,” *J. Intern. Med.*, 2020.

[291] Y. Zheng *et al.*, “Coupling the near-infrared fluorescent dye IR-780 with cabazitaxel makes renal cell carcinoma chemotherapy possible,” *Biomed. Pharmacother.*, 2019.

[292] B. I. Rini *et al.*, “Atezolizumab plus bevacizumab versus sunitinib in patients with previously untreated metastatic renal cell carcinoma (IMmotion151): a multicentre, open-label, phase 3, randomised controlled trial,” *Lancet*, 2019.

[293] A. Wienecke and G. Bacher, “Indibulin, a novel microtubule inhibitor, discriminates between mature neuronal and nonneuronal tubulin,” *Cancer Res.*, 2009.

[294] S. Kapoor, S. Srivastava, and D. Panda, “Indibulin dampens microtubule dynamics and produces synergistic antiproliferative effect with vinblastine in MCF-7 cells: Implications in cancer chemotherapy,” *Sci. Rep.*, 2018.

[295] M. Sisay and D. Edessa, “PARP inhibitors as potential therapeutic agents for various cancers: focus on niraparib and its first global approval for maintenance therapy of gynecologic cancers,” *Gynecol. Oncol. Res. Pract.*, 2017.

[296] F. Buontempo *et al.*, “Inhibition of Akt signaling in hepatoma cells induces apoptotic cell death independent of Akt activation status,” *Invest. New Drugs*, 2011.

[297] C. J. Ciesielski, J. Mei, and L. A. Piccinini, “Effects of cyclosporine A and methotrexate on CD18 expression in recipients of rat cardiac allografts,” *Transpl. Immunol.*, 1998.

[298] R. Scheusan, S. Curescu, D. Stanculeanu, and P. Curescu, “Low-dose methotrexate

and cyclophosphamide in recurrent ovarian cancer," *J. Clin. Oncol.*, 2009.

[299] A. A. A. Elgamal *et al.*, "Prostate-specific membrane antigen (PSMA): Current benefits and future value," *Seminars in Surgical Oncology*. 2000.

[300] M. Michalska, S. Schultze-Seemann, L. Bogatyreva, D. Hauschke, U. Wetterauer, and P. Wolf, "In vitro and in vivo effects of a recombinant anti-PSMA immunotoxin in combination with docetaxel against prostate cancer," *Oncotarget*, 2016.

[301] J. M. Dempster *et al.*, "Agreement between two large pan-cancer CRISPR-Cas9 gene dependency data sets," *Nat. Commun.*, 2019.

[302] X. Deng, S. Das, K. Valdez, K. Camphausen, and U. Shankavaram, "SL-BioDP: Multi-cancer interactive tool for prediction of synthetic lethality and response to cancer treatment," *Cancers (Basel).*, 2019.

[303] Genome Reference Consortium "Human Genome Overview," Available at /www.ncbi.nlm.nih.gov/grc/human. (Accessed November 2020.)

[304] T. C. A. Smith, A. M. Carr, and A. C. Eyre-Walker, "Are sites with multiple single nucleotide variants in cancer genomes a consequence of drivers, hypermutable sites or sequencing errors?," *PeerJ*, vol. 4, p. e2391, 2016.

[305] N. L. Sim, P. Kumar, J. Hu, S. Henikoff, G. Schneider, and P. C. Ng, "SIFT web server: Predicting effects of amino acid substitutions on proteins," *Nucleic Acids Res.*, 2012.

[306] B. Reva, Y. Antipin, and C. Sander, "Predicting the functional impact of protein mutations: Application to cancer genomics," *Nucleic Acids Res.*, vol. 39, no. 17, 2011.

[307] E. Lubeck, A. F. Coskun, T. Zhiyentayev, M. Ahmad, and L. Cai, "MutationTaster2: mutation prediction for the deep-sequencing age," *Nat. Methods*, 2012.

[308] Y. Choi and A. P. Chan, "PROVEAN web server: A tool to predict the functional effect of amino acid substitutions and indels," *Bioinformatics*, 2015.

[309] A. Auton and T. Salcedo, “The 1000 genomes project,” in *Assessing Rare Variation in Complex Traits: Design and Analysis of Genetic Studies*, 2015.

[310] UniProt - Swiss-Prot Protein Knowledgebase, updated September 2020 “Humsavar,” available at *https://www.uniprot.org/docs/humsavar*. (Accessed October 2020).

[311] A. Bateman, “UniProt: A worldwide hub of protein knowledge,” *Nucleic Acids Res.*, 2019.

Appendix 1 – Contribution to other work

During my studies I have also contributed to other work as set out below.

A draft proteomics identification database for *Lymnaea stagnalis*

This work is a collaboration between Murat Eravci, Sarah K. Wooller, Aikaterini Anagnostopoulou, Michael Crossley, Paul Benjamin, Frances Pearl, Ildiko Kemenes and George Kemenes. The work is in progress.

Contribution

I provided bioinformatics support in order to help build a draft proteomics identification database for the snail, *Lymnaea stagnalis* from amino acid fastas provided by the experimental team. To do this I used the program MMSeqs[153] to 'blast' the sequences against genomes from all molluscs available via Uniprot[311], in order to identify the closest protein which is currently annotated, and suggest a draft annotation. The work has not yet been published.

Biological network topology features predict gene dependencies in cancer cell lines

This work is a collaboration between Graeme Benstead-Hume**,** Sarah K. Wooller**,** Samantha Dias**,** Lisa Woodbine**,** Anthony M. Carr**,** and Frances M. G. Pearl.

The paper is currently available on bioxiv at doi: https://doi.org/10.1101.751776

Abstract

In this paper we explore computational approaches that enable us to identify genes that have become essential in individual cancer cell lines. Using recently published experimental

cancer cell line gene essentiality data, human protein-protein interaction (PPI) network data and individual cell-line genomic alteration data we have built a range of machine learning classification models to predict cell line specific acquired essential genes. Genetic alterations found in each individual cell line were modelled by removing protein nodes to reflect loss of function mutations and changing the weights of edges in each PPI to reflect gain of function mutations and gene expression changes. We found that PPI networks can be used to successfully classify human cell line specific acquired essential genes within individual cell lines and between cell lines, even across tissue types with AUC ROC scores of between 0.75 and 0.85. Our novel perturbed PPI network models further improved prediction power compared to the base PPI model and are shown to be more sensitive to genes on which the cell becomes dependent as a result of other changes. These improvements offer opportunities for personalised therapy with each individual's cancer cell dependencies presenting a potential tailored drug target. The overriding motivation for predicting cancer cell line specific acquired essential genes is to provide a low-cost approach to identifying personalised cancer drug targets without the cost of exhaustive loss of function screening.

Contribution

I contributed to the main idea behind the paper and provided bioinformatics as well as editing. Specifically, using mutation and gene expression data from COSMIC[105] and DepMap[81], I calculated the weights of the PPI network edges.

Defining Signatures of Arm-Wise Copy Number Change and Their Associated Drivers in Kidney Cancers

This work is a contribution between Graeme Benstead-Hume, Sarah K. Wooller, Jessica A Downs, and Frances M. G. Pearl. It was published in Nov 2019 and is available at doi: 10.3390/ijms20225762.

Abstract

Using pan-cancer data from The Cancer Genome Atlas (TCGA), we investigated how patterns in copy number alterations in cancer cells vary both by tissue type and as a function of genetic alteration. We find that patterns in both chromosomal ploidy and individual arm copy number are dependent on tumour type. We highlight for example, the significant losses in chromosome arm 3p and the gain of ploidy in 5q in kidney clear cell renal cell carcinoma tissue samples. We find that specific gene mutations are associated with genome-wide copy number changes. Using signatures derived from non-negative matrix factorisation (NMF), we also find gene mutations that are associated with particular patterns of ploidy change. Finally, utilising a set of machine learning classifiers, we successfully predicted the presence of mutated genes in a sample using arm-wise copy number patterns as features. This demonstrates that mutations in specific genes are correlated and may lead to specific patterns of ploidy loss and gain across chromosome arms. Using these same classifiers, we highlight which arms are most predictive of commonly mutated genes in kidney renal clear cell carcinoma (KIRC).

Contribution

I provided analysis and statistical support and helped edit the paper.

Repression of Transcription at DNA Breaks Requires Cohesin throughout Interphase and Prevents Genome Instability

This work is a collaboration between Cornelia Meisenberg, Sarah I Pinder, Suzanna R Hopkins, Sarah K Wooller, Graeme Benstead-Hume, Frances M G Pearl, Penny A Jeggo, and Jessica A Downs. It was published in November 2018 and is available at DOI: 10.1016/j.molcel.2018.11.001.

Abstract

Cohesin subunits are frequently mutated in cancer, but how they function as tumour suppressors is unknown. Cohesin mediates sister chromatid cohesion, but this is not always perturbed in cancer cells. Here, we identify a previously unknown role for cohesin. We find that cohesin is required to repress transcription at DNA double-strand breaks (DSBs). Notably, cohesin represses transcription at DSBs throughout interphase, indicating that this is distinct from its known role in mediating DNA repair through sister chromatid cohesion. We identified a cancer-associated SA2 mutation that supports sister chromatid cohesion but is unable to repress transcription at DSBs. We further show that failure to repress transcription at DSBs leads to large-scale genome rearrangements. Cancer samples lacking SA2 display mutational patterns consistent with loss of this pathway. These findings uncover a new function for cohesin that provides insights into its frequent loss in cancer.

Contribution

I provided statistical and bioinformatics support for this paper. Specifically, I demonstrated that there is a link between large-scale chromosomal alterations and changes in CNV using

COSMIC CNA data for an exome screen on breast cancers where information on large structural changes were available. I then used COSMIC copy number variance data and matching mutation data to demonstrate that samples with STAG2 mutations had statistically significantly more genes exhibiting either a loss or gain due to copy number variation than those without a STAG2 mutation.

'Big data' approaches for novel anti-cancer drug discovery

This work is a collaboration between Graeme Benstead-Hume, Sarah K Wooller, Frances M G Pearl . It was published in June 2017 and is available at

DOI: 10.1080/17460441.2017.1319356.

Abstract

The development of improved cancer therapies is frequently cited as an urgent unmet medical need. Recent advances in platform technologies and the increasing availability of biological 'big data' are providing an unparalleled opportunity to systematically identify the key genes and pathways involved in tumorigenesis. The discoveries made using these new technologies may lead to novel therapeutic interventions. Areas covered: The authors discuss the current approaches that use 'big data' to identify cancer drivers. These approaches include the analysis of genomic sequencing data, pathway data, multi-platform data, identifying genetic interactions such as synthetic lethality and using cell line data. They review how big data is being used to identify novel drug targets. The authors then provide an overview of the available data repositories and tools being used at the forefront of cancer drug discovery. Expert opinion: Targeted therapies based on the genomic events driving the tumour will eventually inform treatment protocols. However, using a tailored

approach to treat all tumour patients may require developing a large repertoire of targeted drugs.

Contribution

I helped research and write the paper.

Appendix 2 – Chapter 3 Supplementary information

The tumour suppressor associated genes with a pathogenic mutation in at least 4% of samples in at least one tumour were:

PTEN, APC, CSMD3, LRP1B, FAT4, ARID1A, PTPRT, GRIN2A, VHL, PTPRD, CNTNAP2, PIK3R1, KMT2C, CTCF, PTPRB, PTPRK, PBRM1, FAT1, DNMT3A, RB1, FBXW7, NF1, ZFHX3, ARID2, CDH10, KEAP1, EP300, ROBO2, STAG2, MED12, BCOR, ATR, NCOR2, ATM, SPEN, CAMTA1, DICER1, MYH9, NCOR1, PPP2R1A, POLE, ELF3, CDH1, CDKN2A, GPC5, RANBP2, KAT6B, SMAD4, SPOP, ERCC2, PTCH1, SETD2, NRG1, TSC2, SMARCA4, CDH11, CHEK2, PTPN13, CHD2, STAG1, PTPRC, ATRX, CBLB, BRCA2, PRDM1, WNK2, IKZF1, AMER1, TSC1, ARID1B, ARHGEF12, ZBTB16, CLTC, ZMYM3, TET2, TRIM33, STK11, DDX3X, CDK12, KDM5C, EBF1, PPP6C, DROSHA, RNF43, CASP8, ATP2B3, MSH6, SMC1A, ARHGEF10, CLTCL1, DNM2, ARHGEF10L, POLG, ASXL2, BAZ1A, LZTR1, LATS2, BAP1, CYLD, MRTFA, ERCC5, PRDM2, EXT1, SLC34A2, PER1, CNOT3, SETD1B, ERCC4, CDC73, LRIG3, EXT2, PMS2, CARS, BRIP1, ASXL1, CUL3, LATS1, N4BP2, AXIN1, ARHGAP26, FANCD2, HNF1A, LARP4B, ACVR2A, TENT5C, WRN, FBLN2, ZNRF3, MAX, RBM10, USP44, PTPN6, CEBPA, AXIN2, KNL1, SMAD2, PHF6, FH, PPARG, TRAF7, ETV6, ERCC3, MSH2, MLH1, EED, SDHA, POLD1, TPM3, CCDC6, RSPO2, CPEB3, FANCA, TGFBR2, DDX10, EIF3E, ABI1, IGF2BP2, BLM, SMARCD1 and NF2.

The oncogene associated genes with a pathogenic mutation in at least 4% of samples in at least one tumour were:

BRAF, PIK3CA, KRAS, CTNNB1, IDH1, NRAS, PREX2, MECOM, BIRC6, TRRAP, KMT2A, GRM3, CTNND2, ROS1, MTOR, KDR, CHD4, FGFR2, EGFR, SETBP1, NTRK3, ALK, TNC, FGFR3, CACNA1D, CARD11, ZEB1, CTNNA2, BCL11A, A1CF, FLT4, PRDM16, ERBB2, ERBB3, ZNF521, KAT6A, BCL9, FLT3, IDH2, AFF3, NUP98, MET, NCOA2, SF3B1, PDGFRA, KDM5A, USP6, ARHGAP5, PLCG1, FGFR1, TCF7L2, FOXP1, AFF4, PDGFRB, KIT, ABL2, HIP1, SALL4, RET, BRD4, RAC1, ABL1, MYB, MACC1, GATA2, DDR2, FGFR4, XPO1, PIK3CB, NFATC2, RAF1, STAT3, KCNJ5, PLAG1, LCK, BCL6, GLI1, IKBKB, SND1, PTPN11, MAPK1, HIF1A, TSHR, STIL, ERG, EWSR1, MN1, BRD3, GNAS, IL7R, TEC, JAK2, MLLT10, ETV1, MAP2K1, MAML2, PBX1, RAP1GDS1, AR, CCND1, PSIP1, DGCR8, KAT7, ARAF, CDH17, SYK, SMO, POU2AF1, AKT3, SGK1, SIX2, ACVR1, IL6ST, NUTM1, FLI1, HRAS, PAX3, USP8, WAS, TAL1, FUBP1, CREB3L2, MYC, STAT6, MUC16, H3F3A, PPM1D, SIX1 and MITF.

The genes that were identified in the Cancer Gene Census as having both tumour suppressor and oncogenic activity, and have a pathogenic mutation in at least 4% of samples in at least one tumour were:

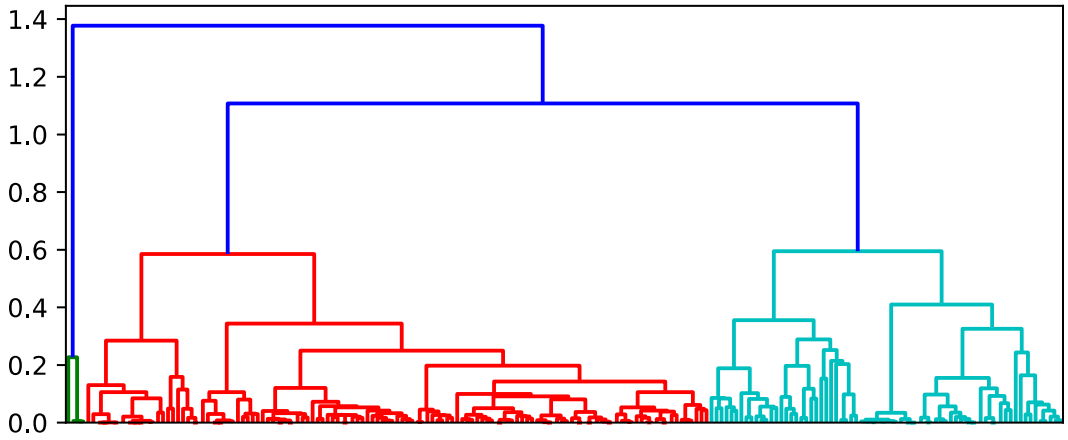
TP53, KMT2D, KDM6A, TP63, ERBB4, NOTCH1, CREBBP, RUNX1T1, CUX1, MAP3K1, BCL9L, BCL11B, RUNX1, GATA3, CIC, JAK1, POLQ, BTK, CDKN1A, BCORL1, NTRK1, WT1, MAP3K13, NOTCH2, IRS4, NFE2L2, ARNT, EZH2, CBL, QKI, TET1, EPAS1, STAT5B, SUZ12, TBL1XR1, PABPC1, ATP1A1, RHOA, TRIM24, PRKAR1A, LEF1, MAP2K4, RAD21, NFKB2, FOXO1, FES, PAX5, ESR1, TCF3 and GPC3.

Appendix 3 – Chapter 4 supplementary figures

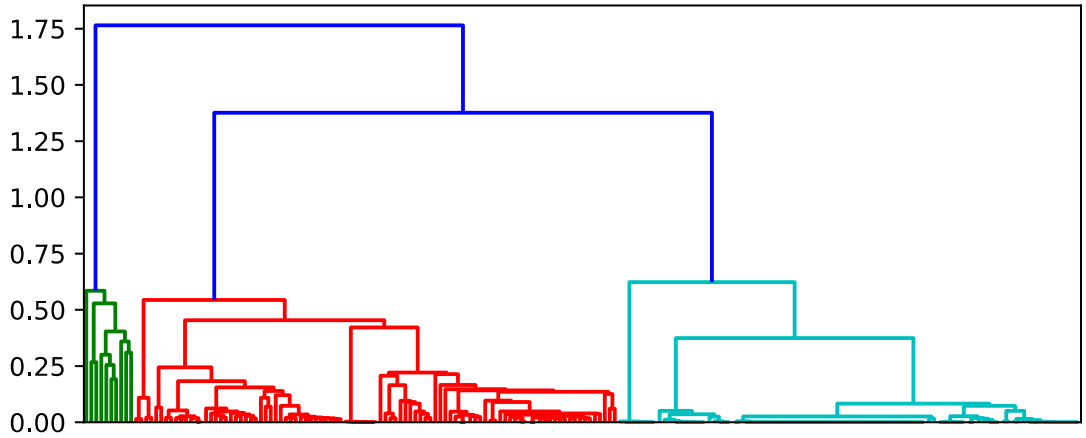
6.3.1 Supplementary figure 4.1

Supplementary figure 1a legend. Bacterial strain clusters across conserved genes using sequence identity from the consensus gene as a measure of distance and using a cophrenetic distance of 0.7 to distinguish sub-species. This example shows *Acinetobacter baumannii*.

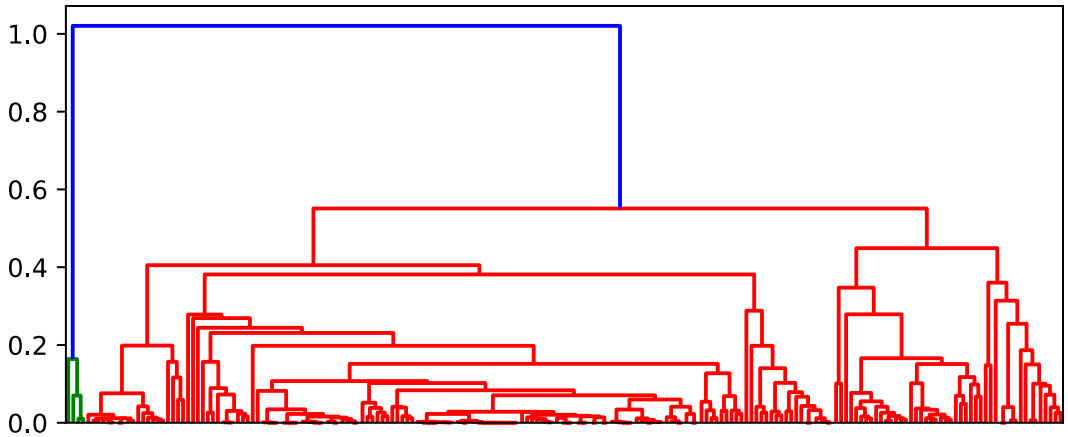
Supplementary figure 1b legend. Bacterial strain clusters across conserved genes using sequence identity from the consensus gene as a measure of distance and using a cophrenetic distance of 0.7 to distinguish sub-species. This example shows *Bacillus cereus.*



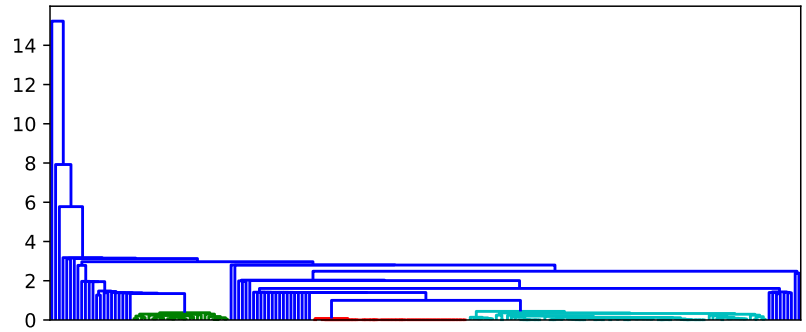
Supplementary figure 1c legend. Bacterial strain clusters across conserved genes using sequence identity from the consensus gene as a measure of distance and using a cophrenetic distance of 0.7 to distinguish sub-species. This example shows *Burkholderia pseudomallei.*

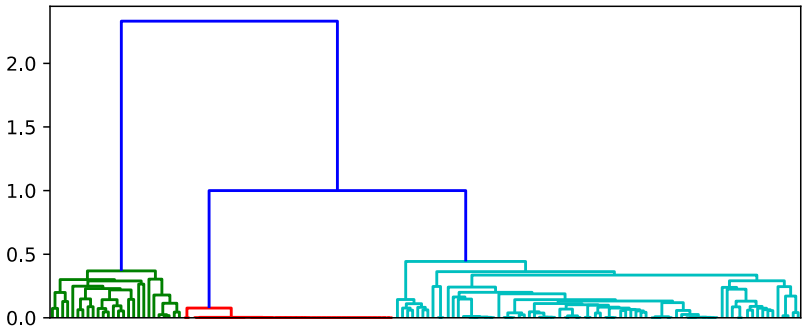


Supplementary figure 1d legend. Bacterial strain clusters across conserved genes using sequence identity from the consensus gene as a measure of distance and using a cophrenetic distance of 0.7 to distinguish sub-species. This example shows *Clostridioides difficile*.

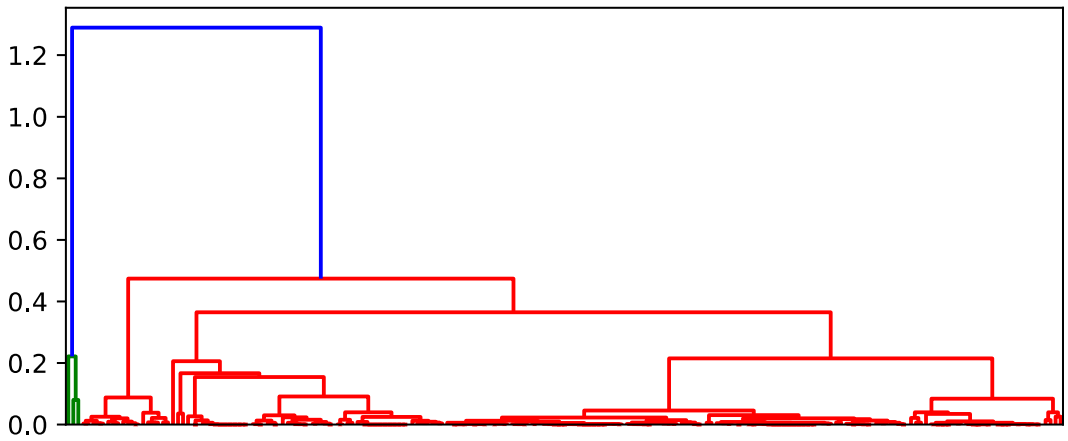


Supplementary figure 1e legend. Bacterial strain clusters across conserved genes using sequence identity from the consensus gene as a measure of distance and using a cophrenetic distance of 0.7 to distinguish sub-species. This example shows *Enterococcus faecalis* . Two versions are presented: the first shows the raw data with many outlying strains that have large differences from each other. The second figure shows those subspecies that lie within the 0.7 cut-off.

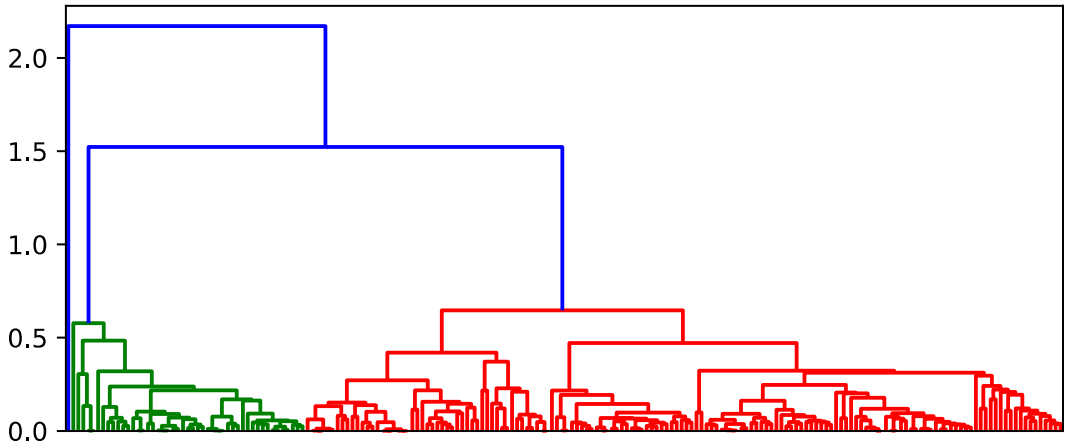




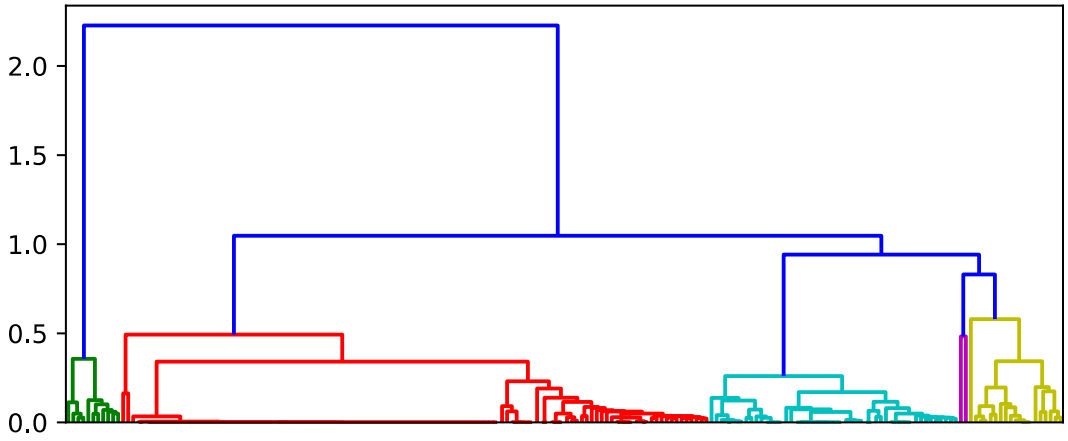
Supplementary figure 1f legend. Bacterial strain clusters across conserved genes using sequence identity from the consensus gene as a measure of distance and using a cophrenetic distance of 0.7 to distinguish sub-species. This example shows *Enterococcus faecium*



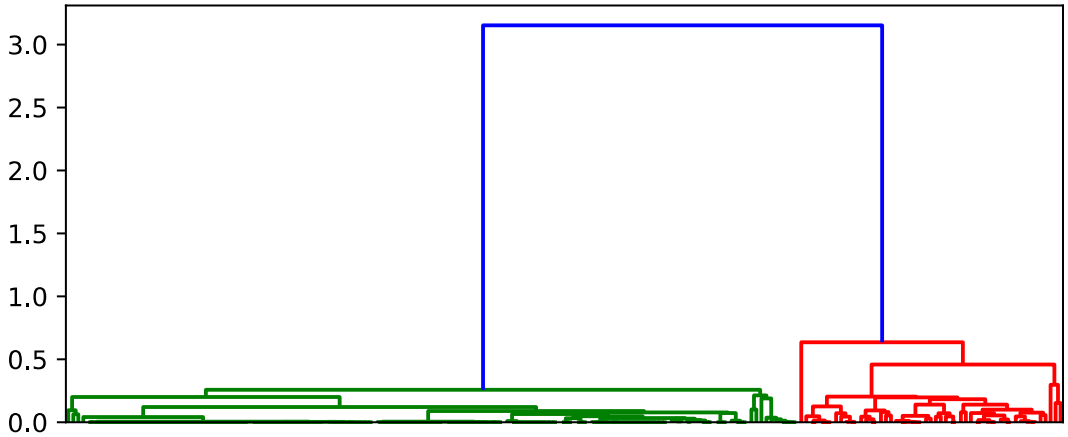
Supplementary figure 1g legend. Bacterial strain clusters across conserved genes using sequence identity from the consensus gene as a measure of distance and using a cophrenetic distance of 0.7 to distinguish sub-species. This example shows *Escherichia coli.*



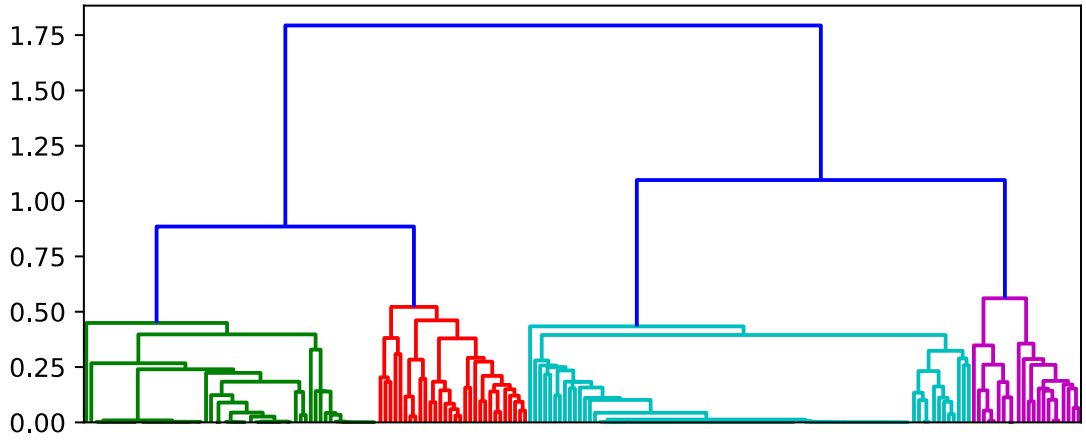
Supplementary figure 1h legend. Bacterial strain clusters across conserved genes using sequence identity from the consensus gene as a measure of distance and using a cophrenetic distance of 0.7 to distinguish sub-species. This example shows *Klebsiella pneumoniae*.



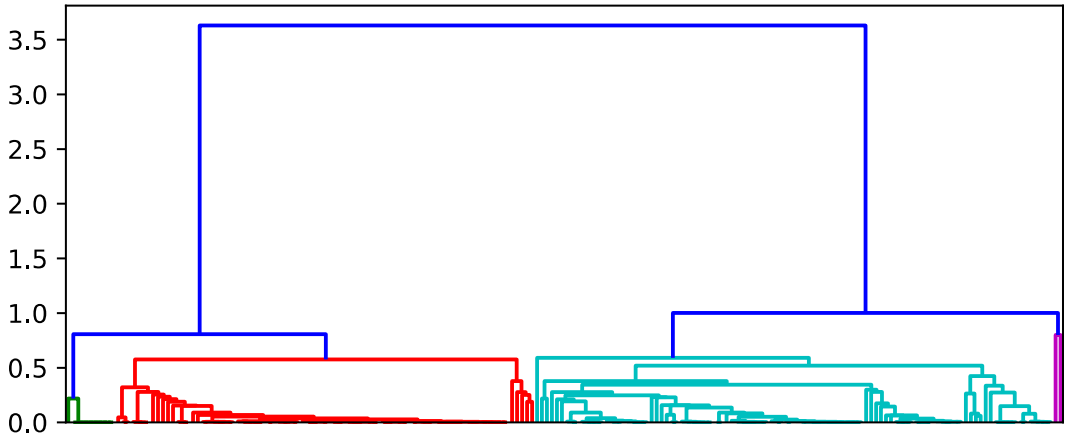
Supplementary figure 1g legend. Bacterial strain clusters across conserved genes using sequence identity from the consensus gene as a measure of distance and using a cophrenetic distance of 0.7 to distinguish sub-species. This example shows *Listeria monocytogenes*



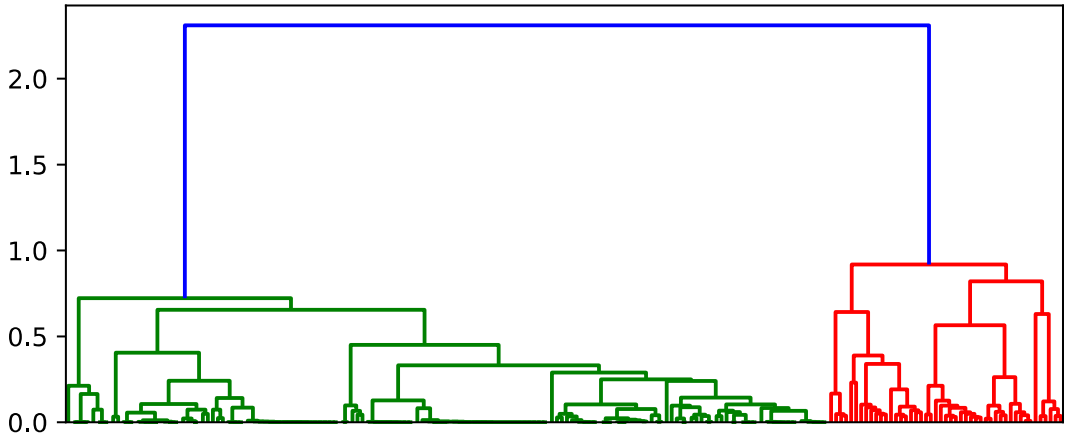
Supplementary figure 1i legend. Bacterial strain clusters across conserved genes using sequence identity from the consensus gene as a measure of distance and using a cophrenetic distance of 0.7 to distinguish sub-species. This example shows *Mycobacterium abscessus.*



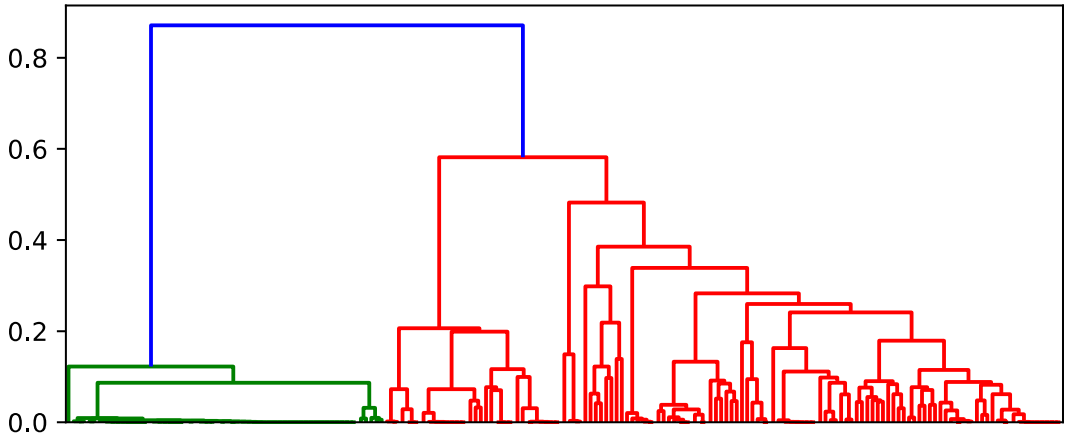
Supplementary figure 1j legend. Bacterial strain clusters across conserved genes using sequence identity from the consensus gene as a measure of distance and using a cophrenetic distance of 0.7 to distinguish sub-species. This example shows *Mycobacterium tuberculosis.*



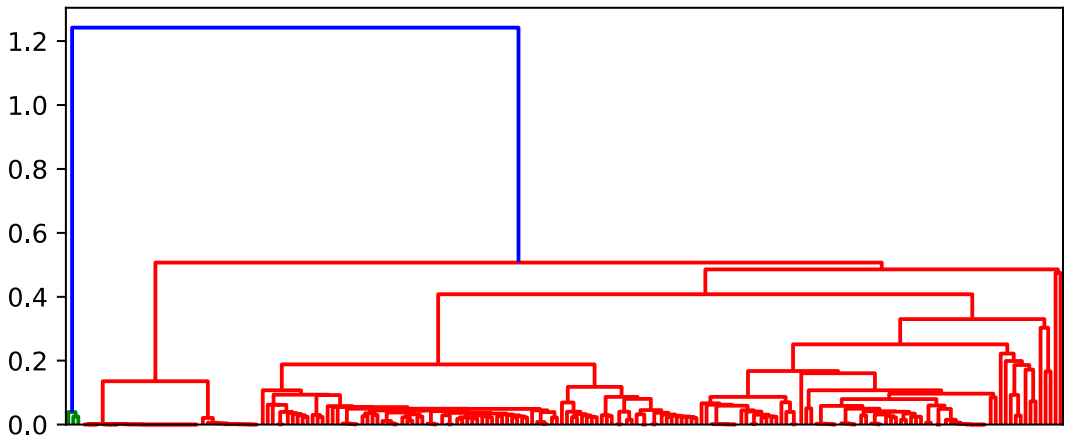
Supplementary figure 1k legend. Bacterial strain clusters across conserved genes using sequence identity from the consensus gene as a measure of distance and using a cophrenetic distance of 0.7 to distinguish sub-species. This example shows *Neisseria gonorrhoeae.*



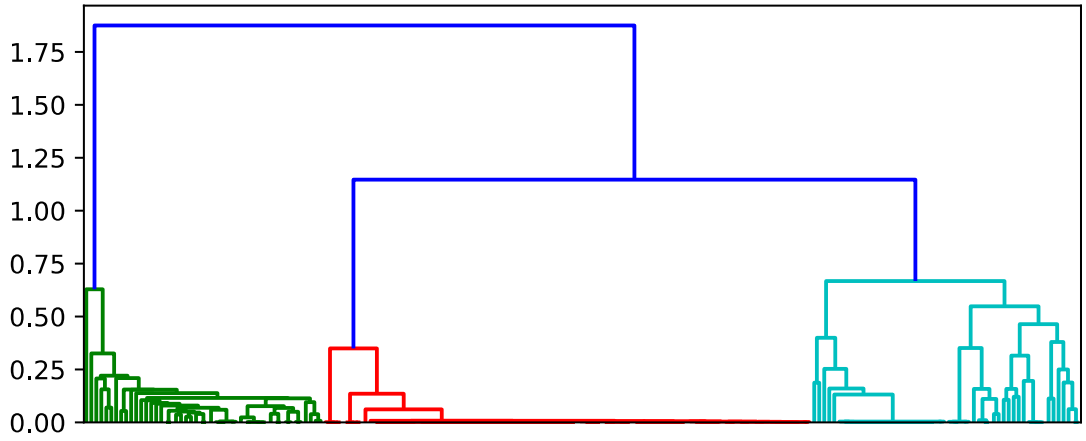
Supplementary figure 1l legend. Bacterial strain clusters across conserved genes using sequence identity from the consensus gene as a measure of distance and using a cophrenetic distance of 0.7 to distinguish sub-species. This example shows *Neisseria meningitidis*.



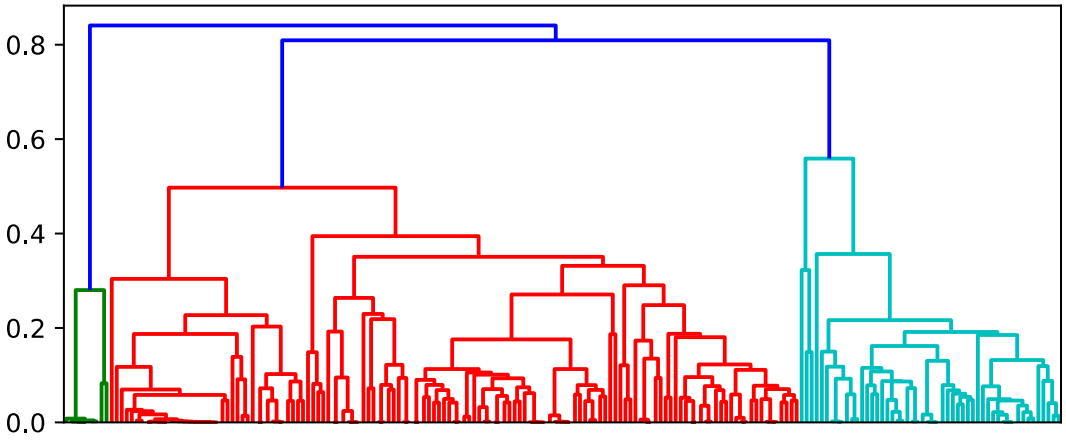
Supplementary figure 1m legend. Bacterial strain clusters across conserved genes using sequence identity from the consensus gene as a measure of distance and using a cophrenetic distance of 0.7 to distinguish sub-species. This example shows *Pseudomonas aeruginosa.*



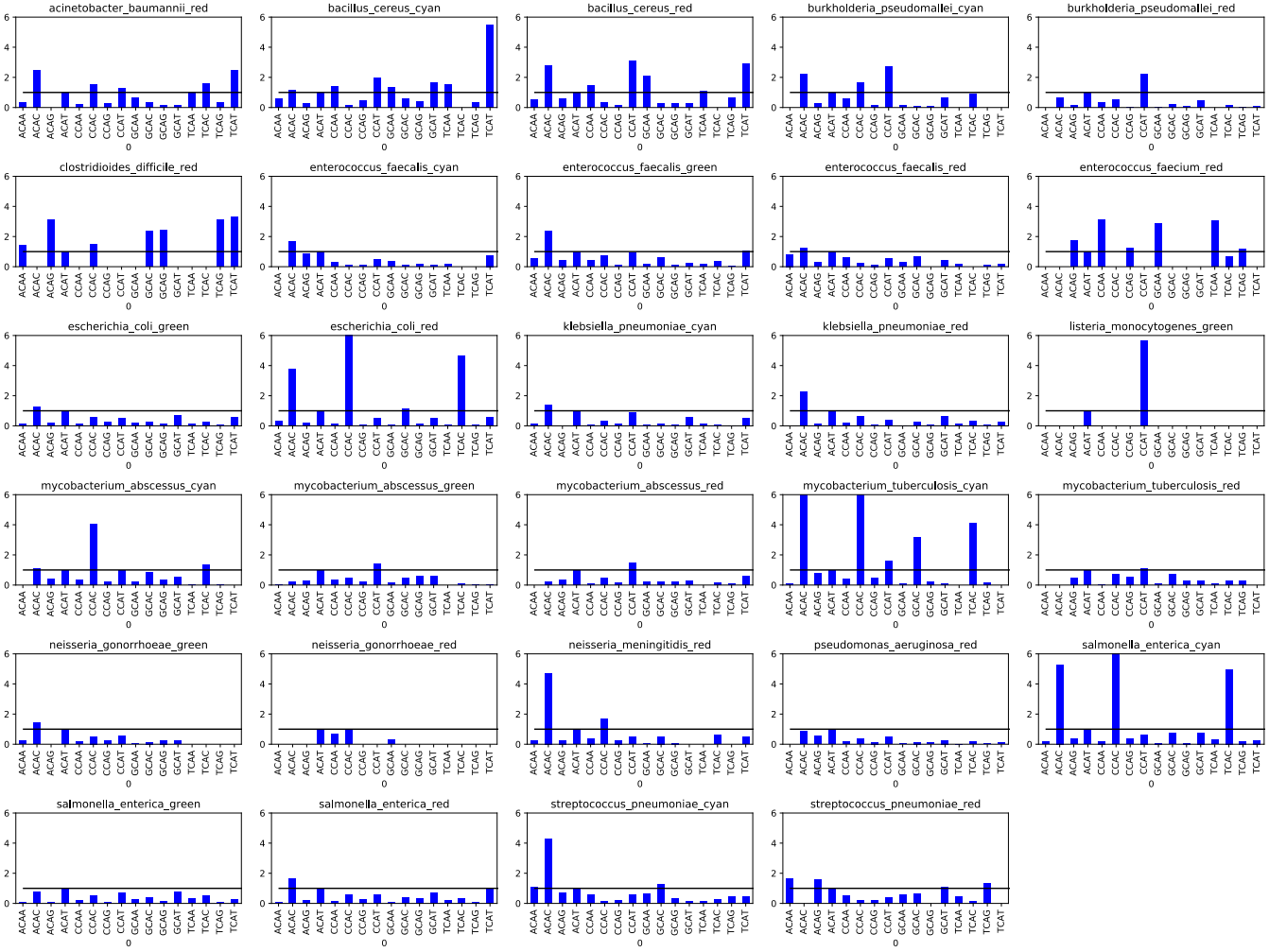
Supplementary figure 1n legend. Bacterial strain clusters across conserved genes using sequence identity from the consensus gene as a measure of distance and using a cophrenetic distance of 0.7 to distinguish sub-species. This example shows *Salmonella enterica.*



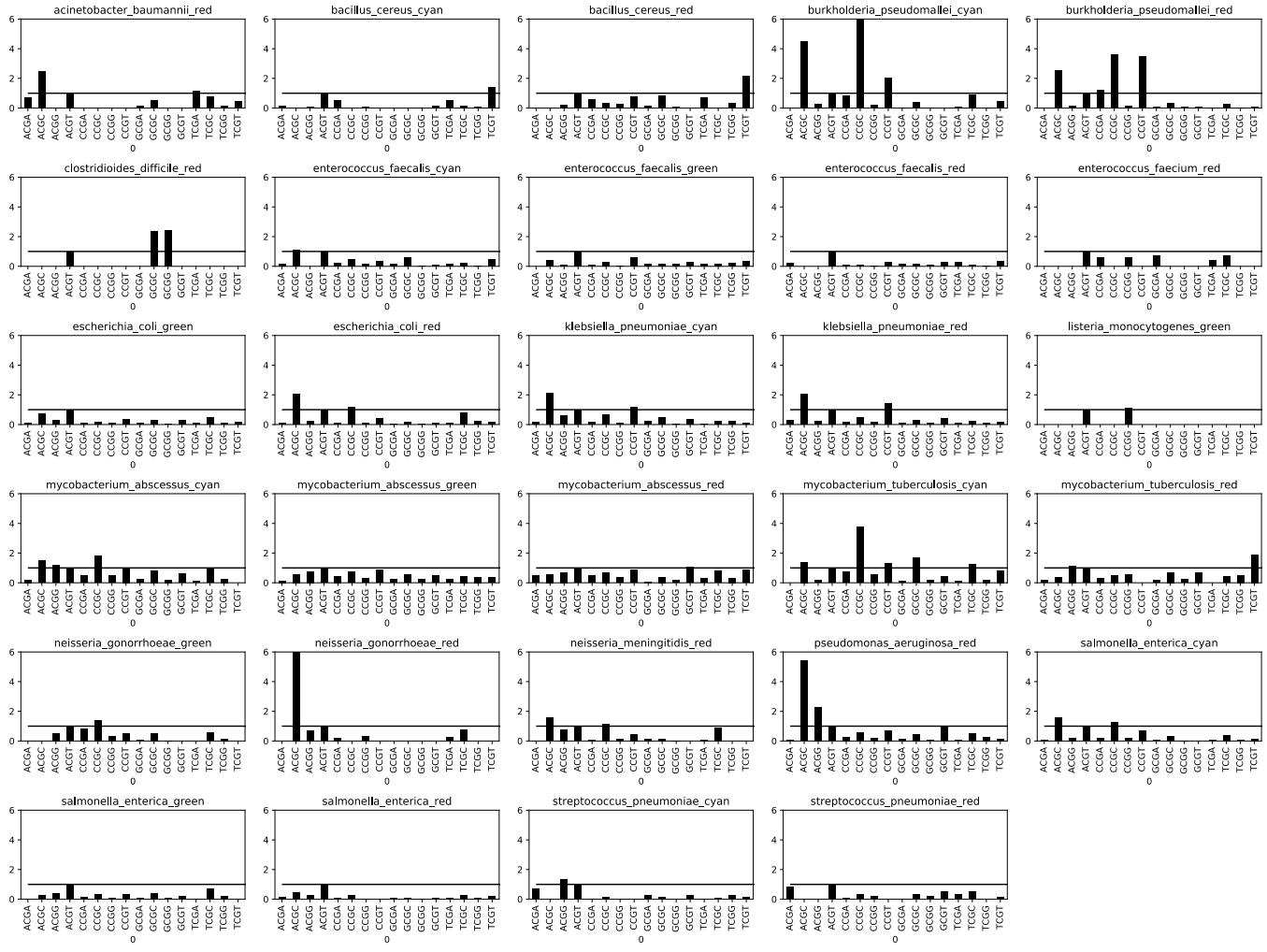
Supplementary figure 1o legend. Bacterial strain clusters across conserved genes using sequence identity from the consensus gene as a measure of distance and using a cophrenetic distance of 0.7 to distinguish sub-species. This example shows *Streptococcus pneumoniae*.



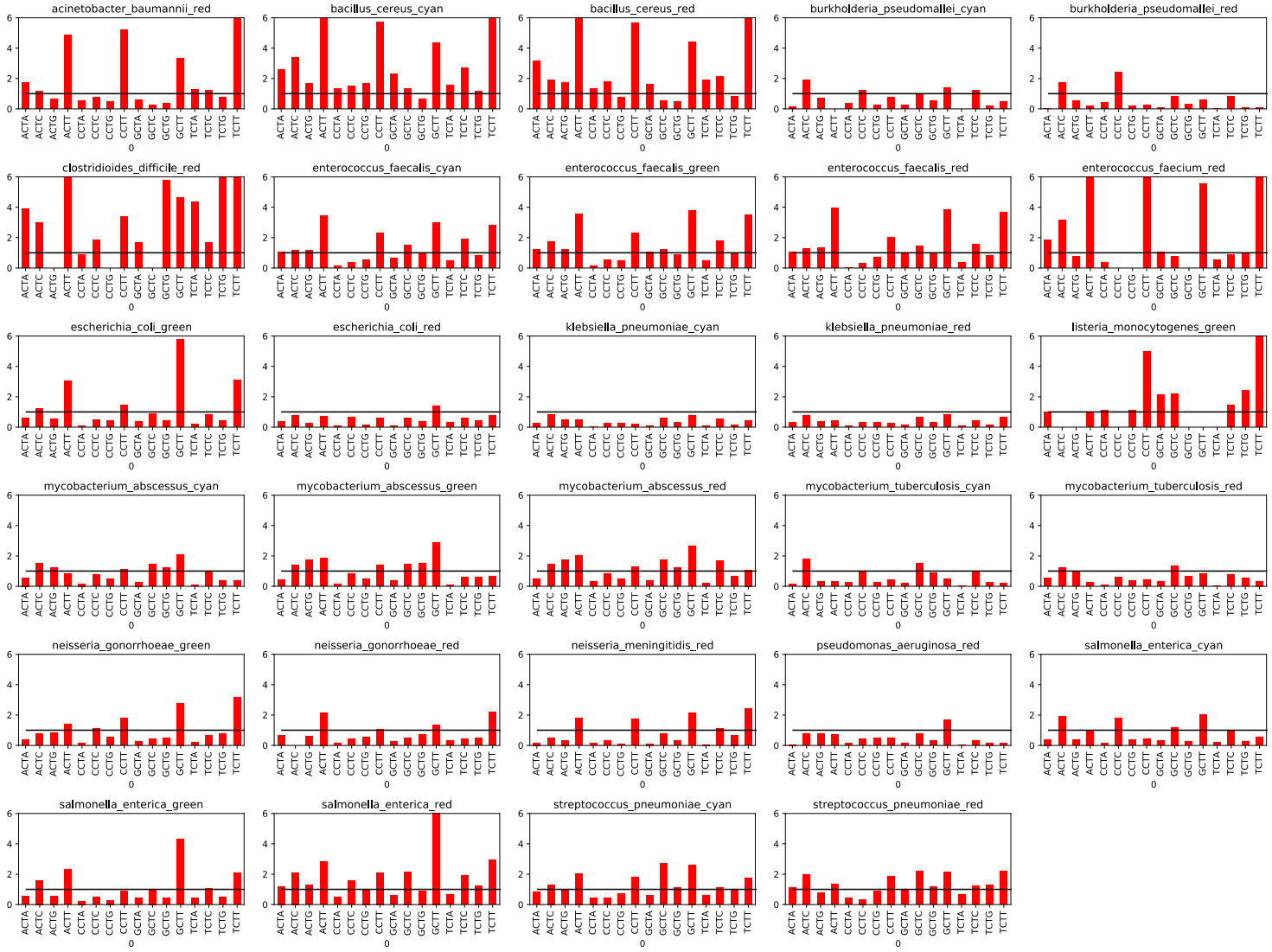
6.3.2 *Supplementary figure 4.2*



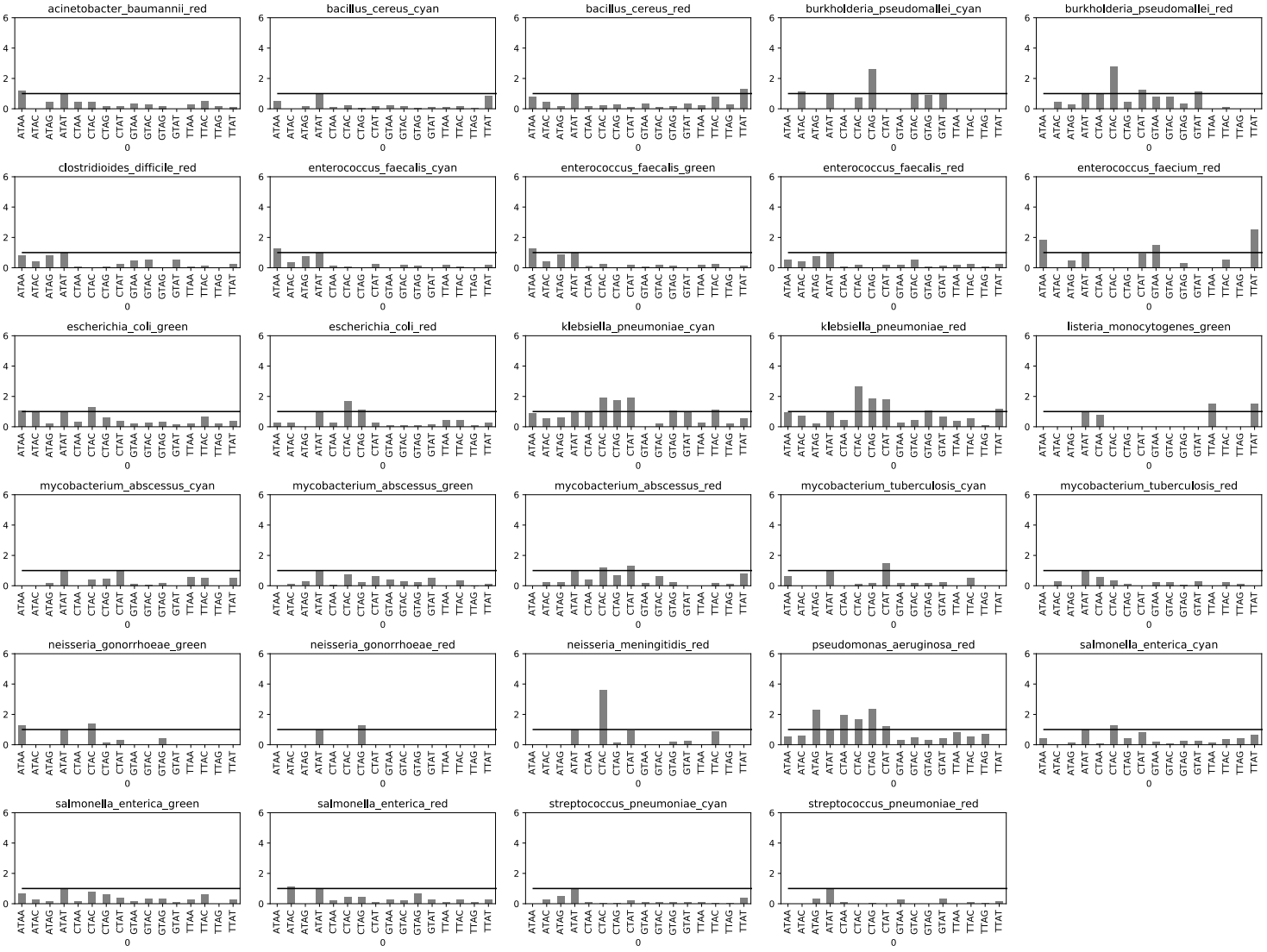
Supplementary Figure 2a: Relative frequency of unique silent C>A mutations



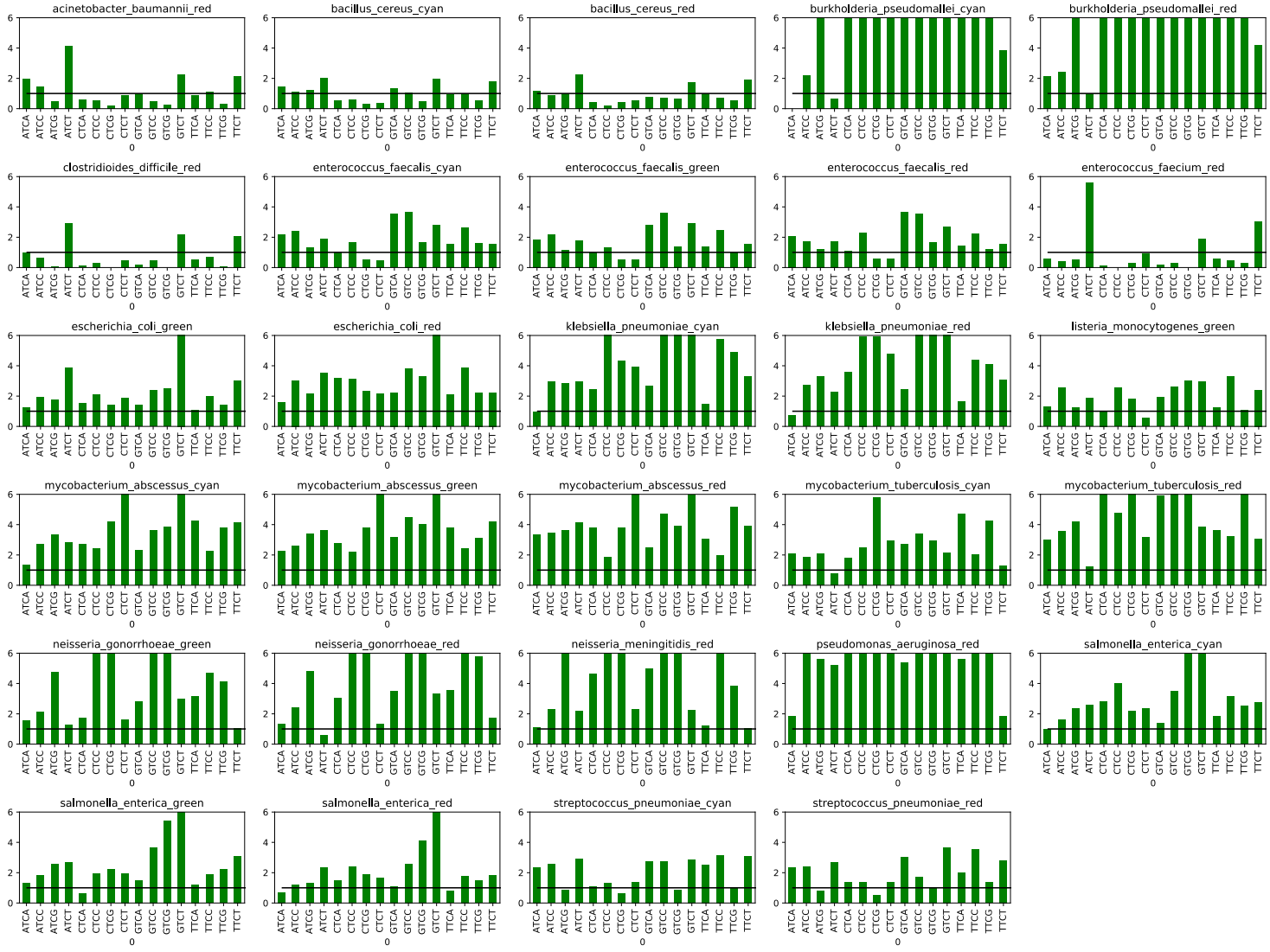
Supplementary Figure 2b: Relative frequency of unique silent C>G mutations



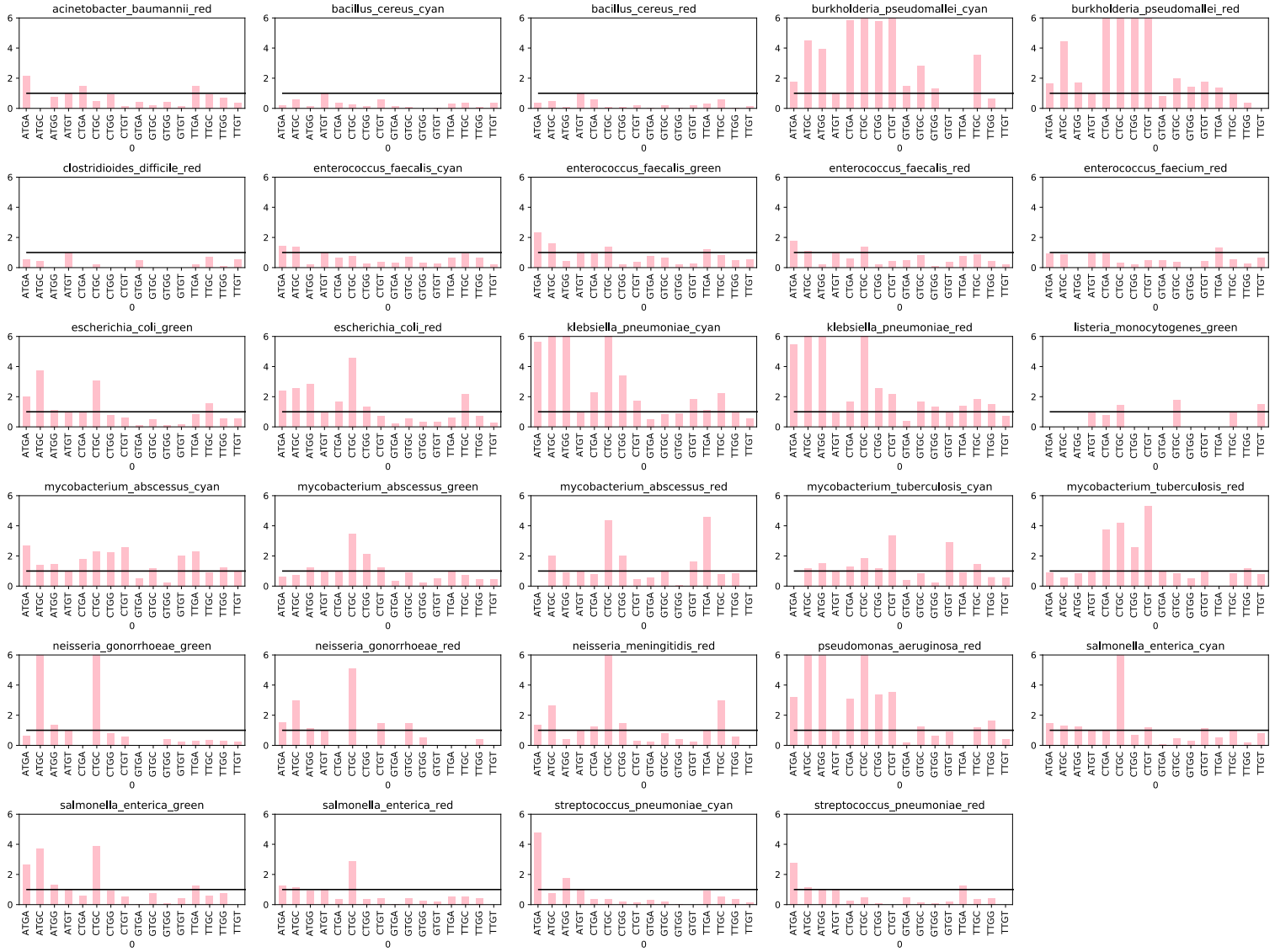
Supplementary Figure 2c: Relative frequency of unique silent C>T mutations



Supplementary Figure 2d: Relative frequency of unique silent T>A mutations



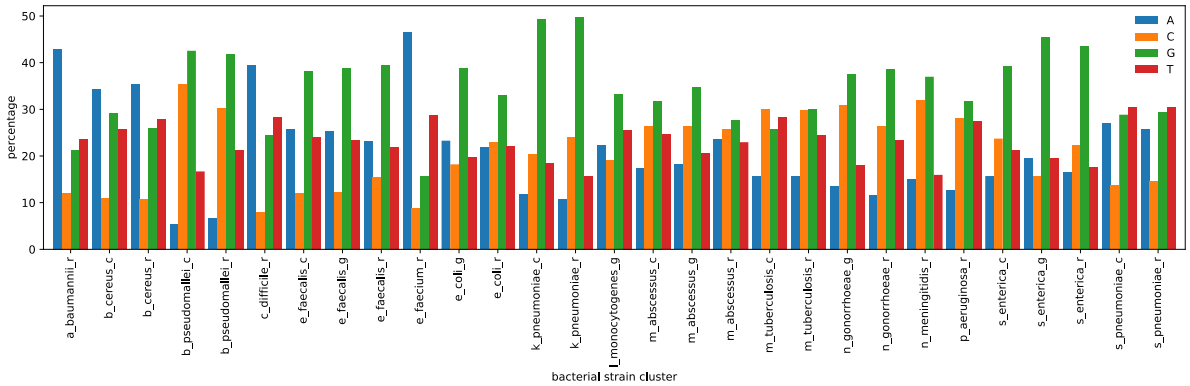
Supplementary Figure 2e: Relative frequency of unique silent T>C mutations

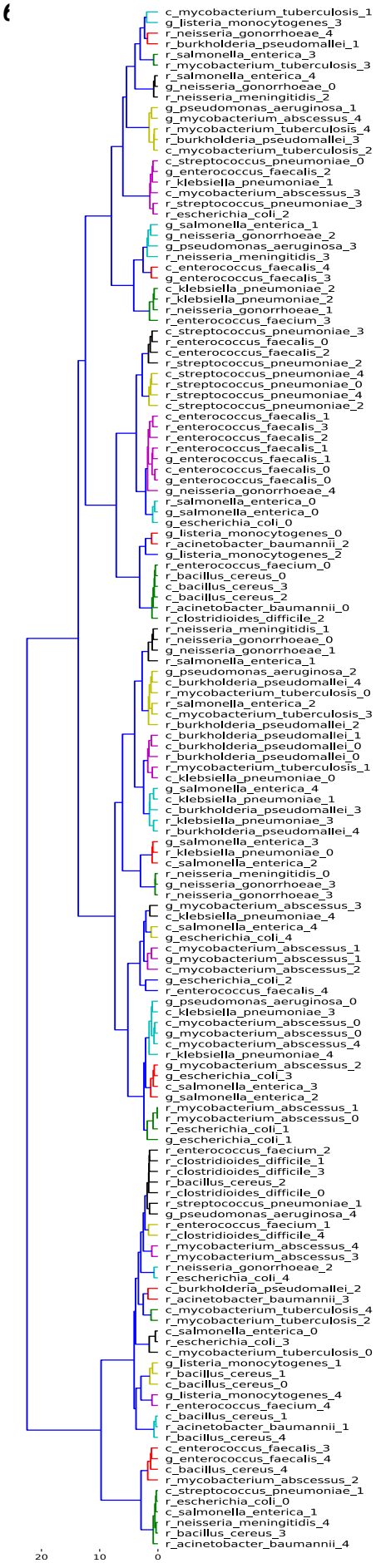


Supplementary Figure 2f: Relative frequency of unique silent T>G mutations

6.3.3 Supplementary figure 4.3

Percentage of unique silent T>C mutations that are preceded by A,C,G,T nucleotides respectively for different bacterial strain clusters. Suffixes _r, _g, _c refer to the colours (red, green, cyan) of the bacterial strain cluster in supplementary figures X. Most bacterial strain clusters are dominated by T>C mutations. For most bacterial species these are disproportionately found after a G nucleotide, i.e. they take the form GT>GC.

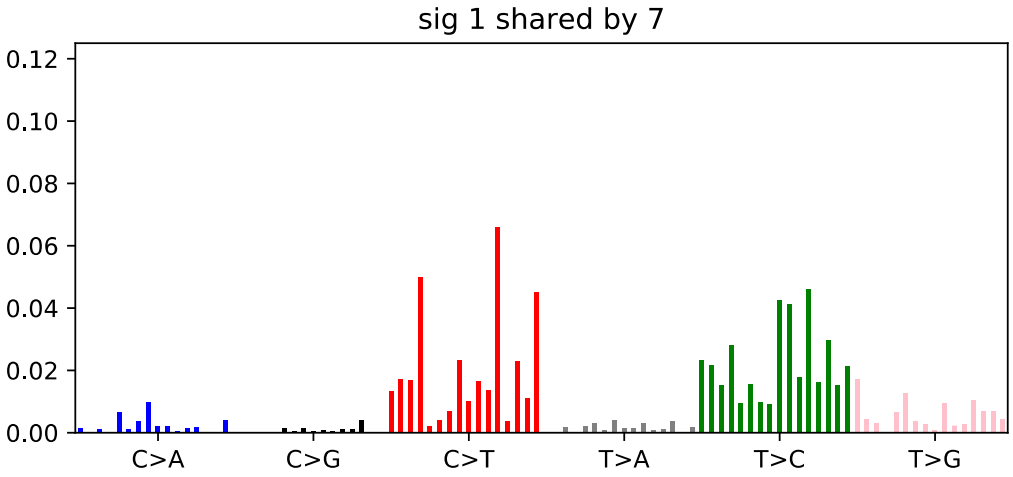



c_mycobacterium_tuberculosis_1
g_listeria_monocytogenes_3
r_neisseria_gonorrhoeae_4
r_burkholderia_pseudomallei_1
r_salmonella_enterica_3
r_mycobacterium_tuberculosis_3
r_salmonella_enterica_4
g_neisseria_gonorrhoeae_0
r_neisseria_meningitidis_2
g_pseudomonas_aeruginosa_1
g_mycobacterium_abscessus_4
r_mycobacterium_tuberculosis_4
r_burkholderia_pseudomallei_3
c_mycobacterium_tuberculosis_2
c_streptococcus_pneumoniae_0
g_enterococcus_faecalis_2
r_klebsiella_pneumoniae_1
c_mycobacterium_abscessus_3
r_streptococcus_pneumoniae_3
r_escherichia_coli_2
g_salmonella_enterica_1
g_neisseria_gonorrhoeae_2
g_pseudomonas_aeruginosa_3
r_neisseria_meningitidis_3
c_enterococcus_faecalis_4
g_enterococcus_faecalis_3
c_klebsiella_pneumoniae_2
r_klebsiella_pneumoniae_2
r_neisseria_gonorrhoeae_1
r_enterococcus_faecium_3
c_streptococcus_pneumoniae_3
r_enterococcus_faecalis_0
c_enterococcus_faecalis_2
r_streptococcus_pneumoniae_2
c_streptococcus_pneumoniae_4
r_streptococcus_pneumoniae_0
r_streptococcus_pneumoniae_4
c_streptococcus_pneumoniae_2
c_enterococcus_faecalis_1
r_enterococcus_faecalis_3
r_enterococcus_faecalis_2
r_enterococcus_faecalis_1
g_enterococcus_faecalis_1
c_enterococcus_faecalis_0
g_enterococcus_faecalis_0
g_neisseria_gonorrhoeae_4
r_salmonella_enterica_0
g_salmonella_enterica_0
g_escherichia_coli_0
g_listeria_monocytogenes_0
r_acinetobacter_baumannii_2
g_listeria_monocytogenes_2
r_enterococcus_faecium_0
r_bacillus_cereus_0
c_bacillus_cereus_3
c_bacillus_cereus_2
r_acinetobacter_baumannii_0
r_clostridioides_difficile_2
r_neisseria_meningitidis_1
r_neisseria_gonorrhoeae_0
g_neisseria_gonorrhoeae_1
r_salmonella_enterica_1
g_pseudomonas_aeruginosa_2
c_burkholderia_pseudomallei_4
r_mycobacterium_tuberculosis_0
r_salmonella_enterica_2
c_mycobacterium_tuberculosis_3
r_burkholderia_pseudomallei_2
c_burkholderia_pseudomallei_1
c_burkholderia_pseudomallei_0
r_burkholderia_pseudomallei_0
r_mycobacterium_tuberculosis_1
c_klebsiella_pneumoniae_0
g_salmonella_enterica_4
c_klebsiella_pneumoniae_1
c_burkholderia_pseudomallei_3
r_klebsiella_pneumoniae_3
r_burkholderia_pseudomallei_4
g_salmonella_enterica_3
r_klebsiella_pneumoniae_0
c_salmonella_enterica_2
r_neisseria_meningitidis_0
g_neisseria_gonorrhoeae_3
r_neisseria_gonorrhoeae_3
g_mycobacterium_abscessus_3
c_klebsiella_pneumoniae_4
c_salmonella_enterica_4
g_escherichia_coli_4
c_mycobacterium_abscessus_1
g_mycobacterium_abscessus_1
c_mycobacterium_abscessus_2
g_escherichia_coli_2
r_enterococcus_faecalis_4
g_pseudomonas_aeruginosa_0
c_klebsiella_pneumoniae_3
c_mycobacterium_abscessus_0
g_mycobacterium_abscessus_0
c_mycobacterium_abscessus_4
r_klebsiella_pneumoniae_4
g_mycobacterium_abscessus_2
g_escherichia_coli_3
c_salmonella_enterica_3
g_salmonella_enterica_2
r_mycobacterium_abscessus_1
r_mycobacterium_abscessus_0
r_escherichia_coli_1
g_escherichia_coli_1
r_enterococcus_faecium_2
r_clostridioides_difficile_1
r_clostridioides_difficile_3
r_bacillus_cereus_2
r_clostridioides_difficile_0
r_streptococcus_pneumoniae_1
g_pseudomonas_aeruginosa_4
r_enterococcus_faecium_1
r_clostridioides_difficile_4
r_mycobacterium_abscessus_4
r_mycobacterium_abscessus_3
r_neisseria_gonorrhoeae_2
r_escherichia_coli_4
c_burkholderia_pseudomallei_2
r_acinetobacter_baumannii_3
c_mycobacterium_tuberculosis_4
r_mycobacterium_tuberculosis_2
c_salmonella_enterica_0
r_escherichia_coli_3
c_mycobacterium_tuberculosis_0
g_listeria_monocytogenes_1
r_bacillus_cereus_1
c_bacillus_cereus_0
g_listeria_monocytogenes_4
r_enterococcus_faecium_4
c_bacillus_cereus_1
r_acinetobacter_baumannii_1
r_bacillus_cereus_4
c_enterococcus_faecalis_3
g_enterococcus_faecalis_4
c_bacillus_cereus_4
r_mycobacterium_abscessus_2
c_streptococcus_pneumoniae_1
r_escherichia_coli_0
c_salmonella_enterica_1
r_neisseria_meningitidis_4
r_bacillus_cereus_3
r_acinetobacter_baumannii_4
20
10
0

The mutational signatures formed from the mean mutational fingerprints can be clustered into 40 distinct clusters, 24 of which are shared between 3 or more bacterial subspecies.

6.3.5 Supplementary figure 4.5

In total 24 signatures were found that were shared by 3 or more bacterial subspecies. These are shown below. The prefix r_, g_, c_ refers to the colour of the corresponding cluster of bacterial strains in Supplementary figure 4.1.



Signature 1 was the dominant signature in:

56% of r_enterococcus_faecalis strains

48% of g_enterococcus_faecalis strains

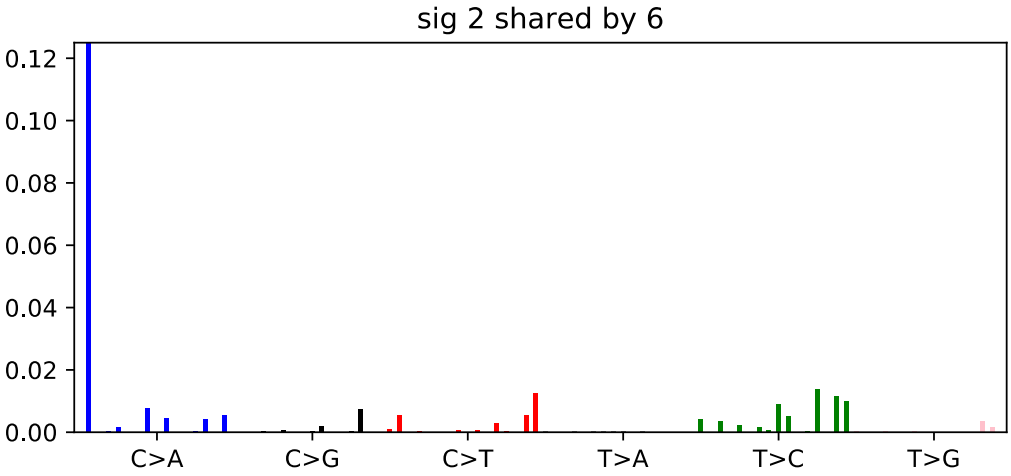
44% of c_enterococcus_faecalis strains

27% of g_neisseria_gonorrhoeae strains

15% of r_salmonella_enterica strains

6% of g_salmonella_enterica strains

2% of g_escherichia_coli strains



Signature 2 was the dominant signature in:

8% of c_salmonella_enterica strains

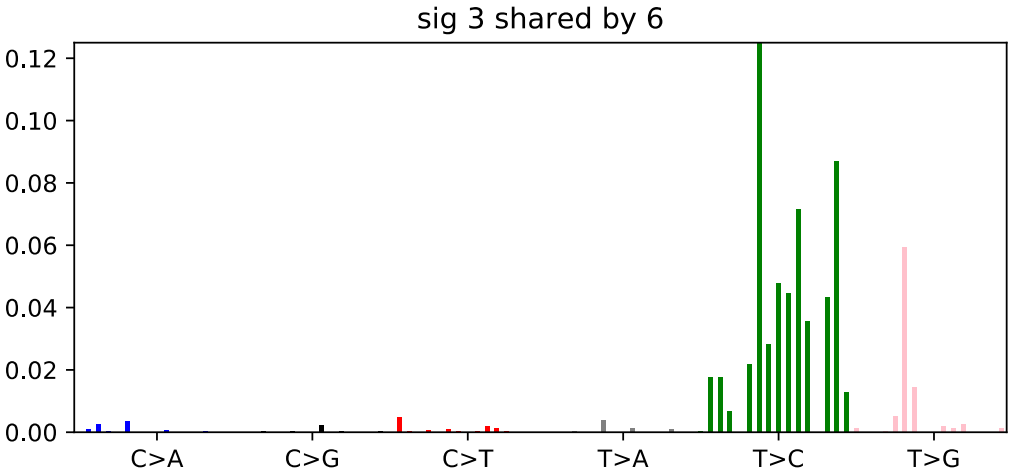
4% of r_bacillus_cereus strains

4% of c_streptococcus_pneumoniae strains

3% of r_neisseria_meningitidis strains

2% of r_escherichia_coli strains

1% of r_acinetobacter_baumannii strains



Signature 3 was the dominant signature in:

52% of c_burkholderia_pseudomallei strains

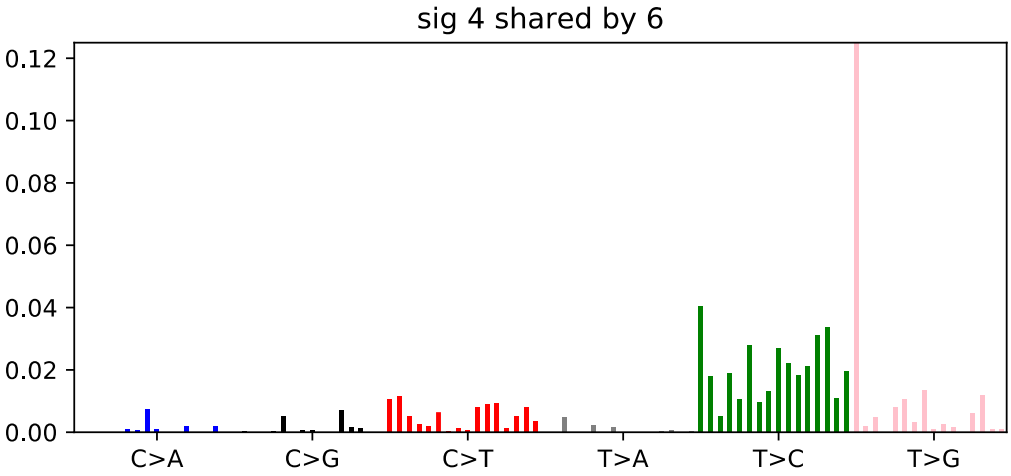
46% of g_pseudomonas_aeruginosa strains

40% of c_mycobacterium_tuberculosis strains

27% of r_mycobacterium_tuberculosis strains

27% of r_burkholderia_pseudomallei strains

26% of r_salmonella_enterica strains



Signature 4 was the dominant signature in:

17% of c_mycobacterium_abscessus strains

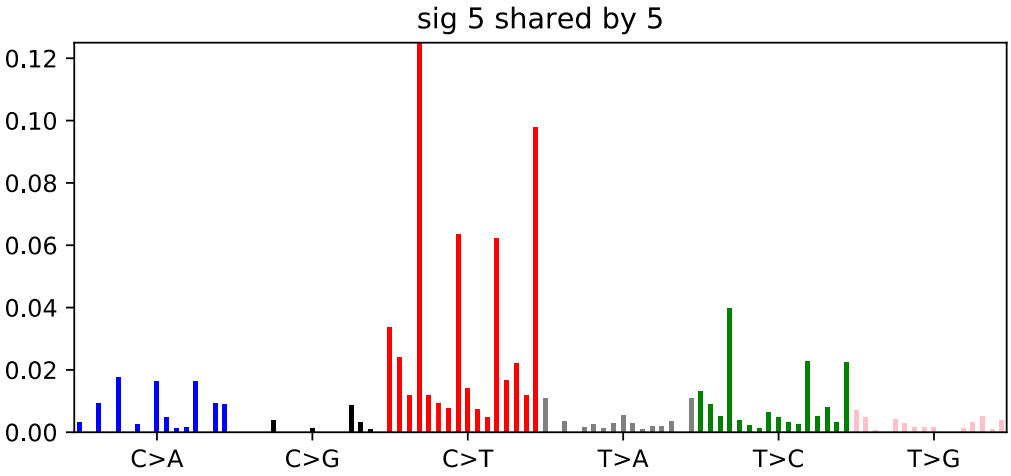
12% of g_enterococcus_faecalis strains

9% of r_streptococcus_pneumoniae strains

9% of c_streptococcus_pneumoniae strains

6% of r_klebsiella_pneumoniae strains

5% of r_escherichia_coli strains



Signature 5 was the dominant signature in:

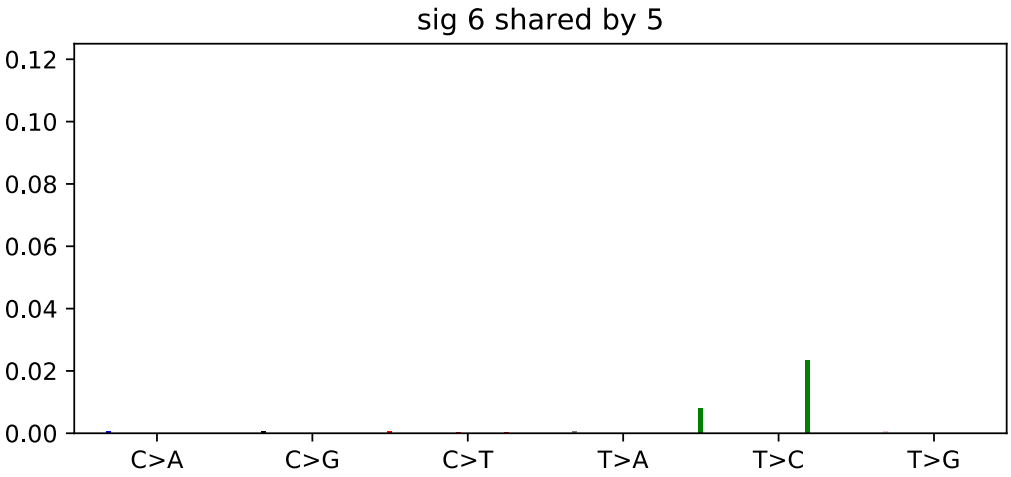
79% of r_bacillus_cereus strains

75% of c_bacillus_cereus strains

57% of r_clostridioides_difficile strains

48% of r_enterococcus_faecium strains

40% of r_acinetobacter_baumannii strains



Signature 6 was the dominant signature in:

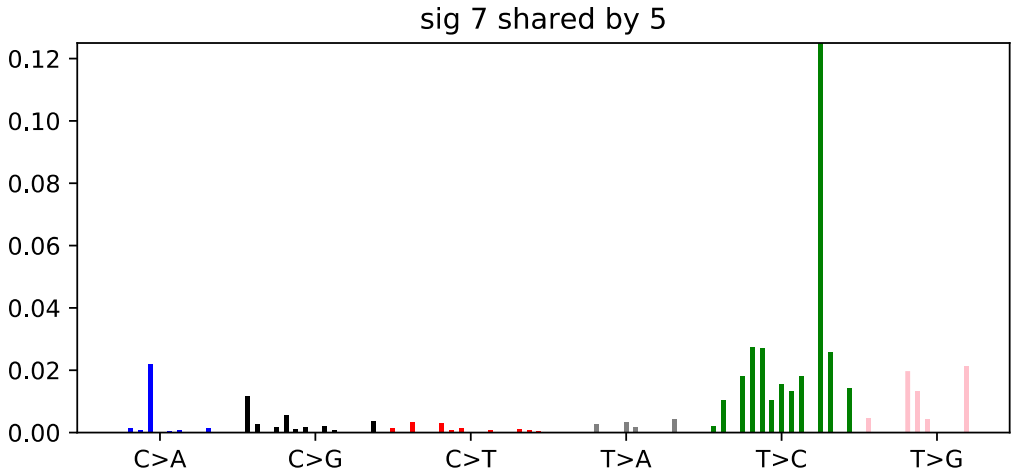
18% of r_clostridioides_difficile strains

4% of r_enterococcus_faecium strains

4% of g_pseudomonas_aeruginosa strains

2% of r_streptococcus_pneumoniae strains

2% of r_bacillus_cereus strains



Signature 7 was the dominant signature in:

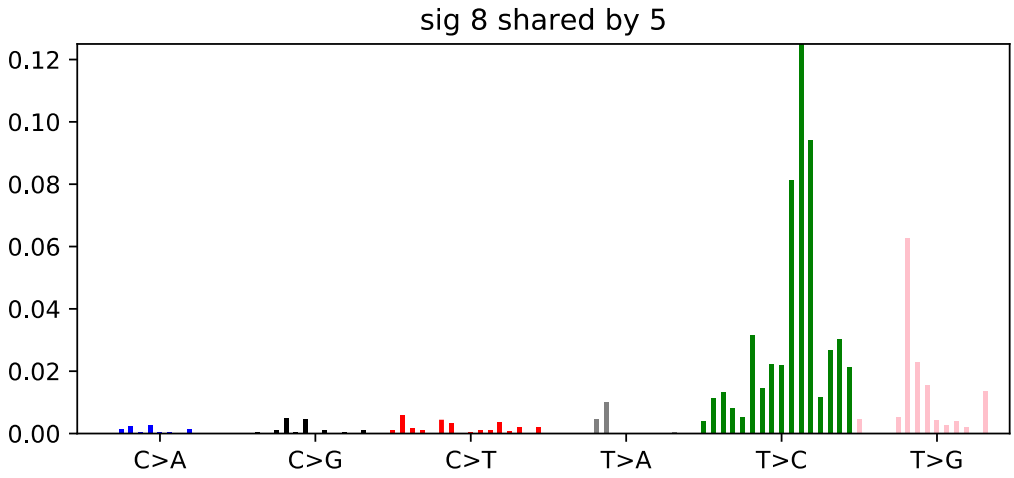
6% of r_mycobacterium_tuberculosis strains

6% of g_mycobacterium_abscessus strains

3% of c_mycobacterium_tuberculosis strains

1% of r_burkholderia_pseudomallei strains

1% of g_pseudomonas_aeruginosa strains



Signature 8 was the dominant signature in:

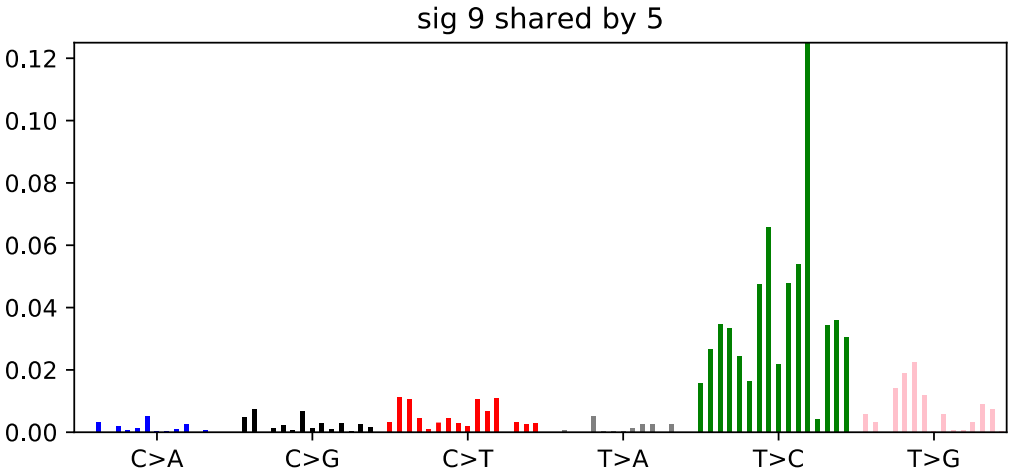
34% of r_burkholderia_pseudomallei strains

17% of g_salmonella_enterica strains

16% of c_burkholderia_pseudomallei strains

12% of c_klebsiella_pneumoniae strains

10% of r_klebsiella_pneumoniae strains



Signature 9 was the dominant signature in:

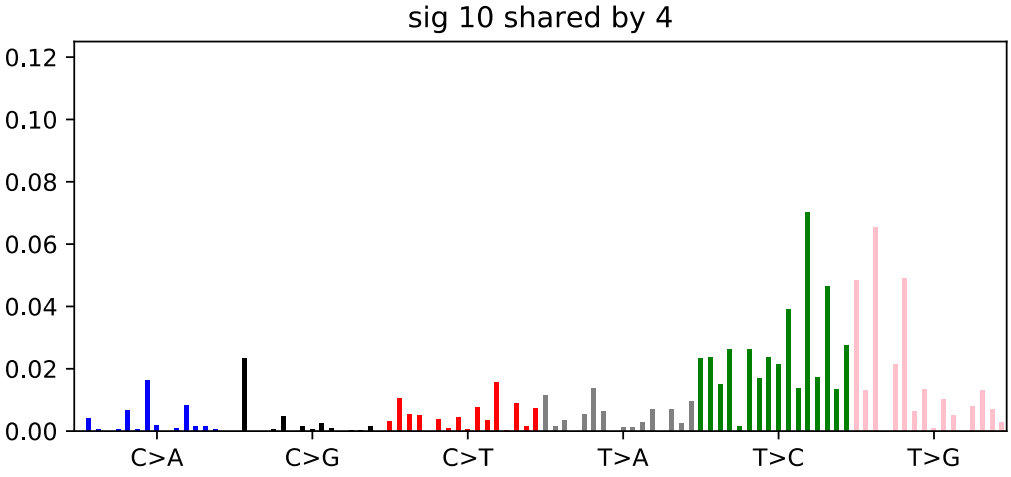
28% of c_klebsiella_pneumoniae strains

27% of g_pseudomonas_aeruginosa strains

19% of c_mycobacterium_abscessus strains

16% of g_mycobacterium_abscessus strains

12% of r_klebsiella_pneumoniae strains



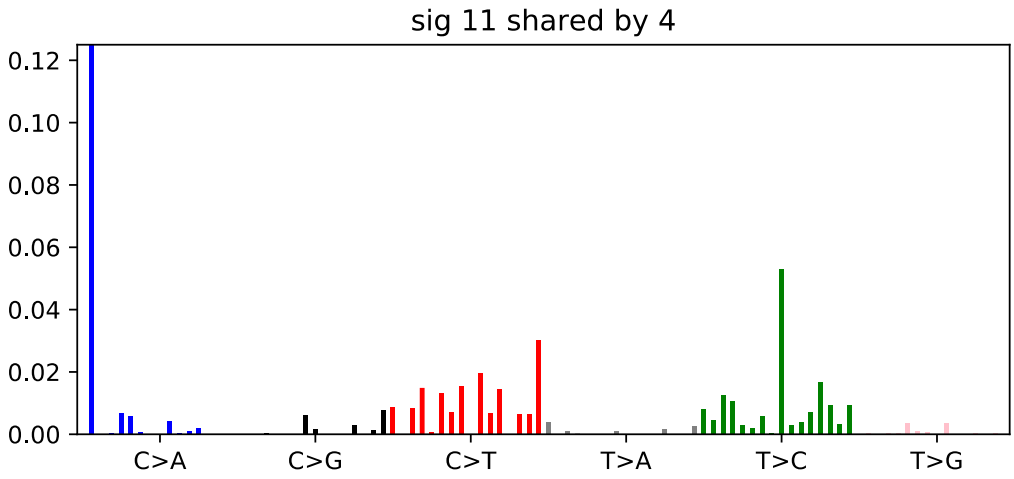
Signature 10 was the dominant signature in:

30% of c_klebsiella_pneumoniae strains

10% of g_escherichia_coli strains

6% of g_mycobacterium_abscessus strains

6% of c_salmonella_enterica strains



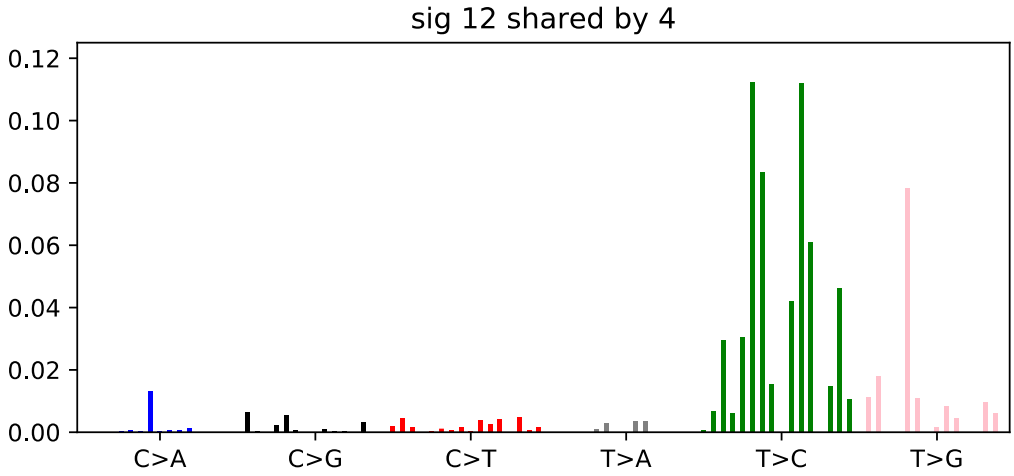
Signature 11 was the dominant signature in:

72% of r_mycobacterium_abscessus strains

8% of g_enterococcus_faecalis strains

7% of c_bacillus_cereus strains

5% of c_enterococcus_faecalis strains



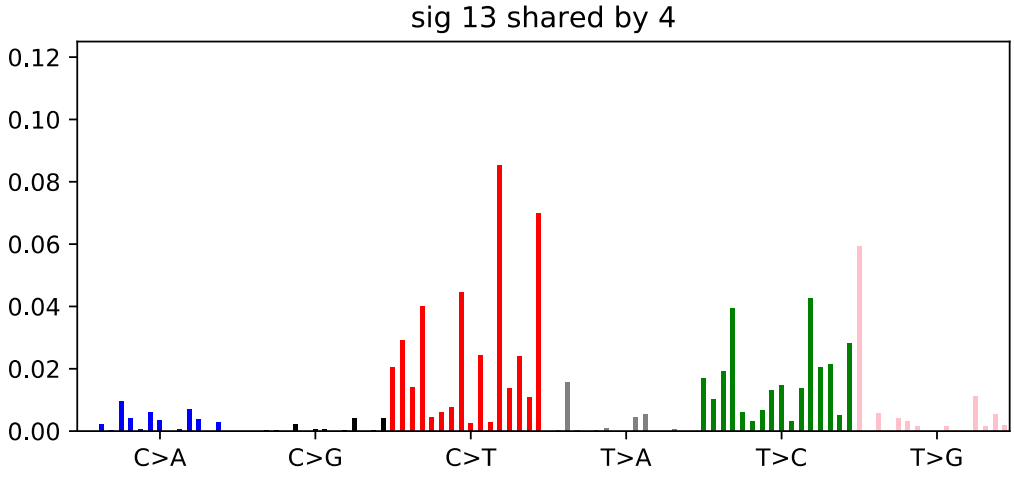
Signature 12 was the dominant signature in:

33% of r_mycobacterium_tuberculosis strains

25% of c_burkholderia_pseudomallei strains

14% of c_klebsiella_pneumoniae strains

6% of r_burkholderia_pseudomallei strains



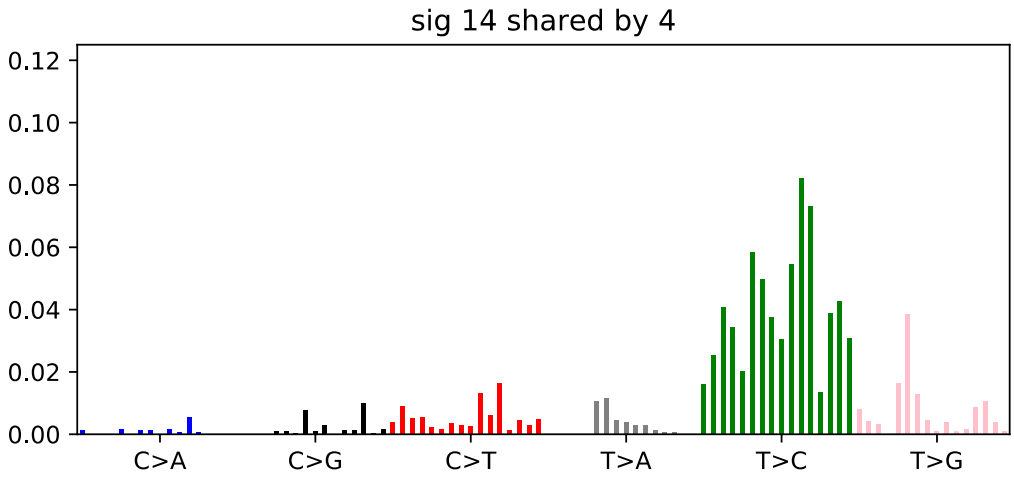
Signature 13 was the dominant signature in:

27% of c_streptococcus_pneumoniae strains

17% of r_streptococcus_pneumoniae strains

15% of c_enterococcus_faecalis strains

8% of r_enterococcus_faecalis strains



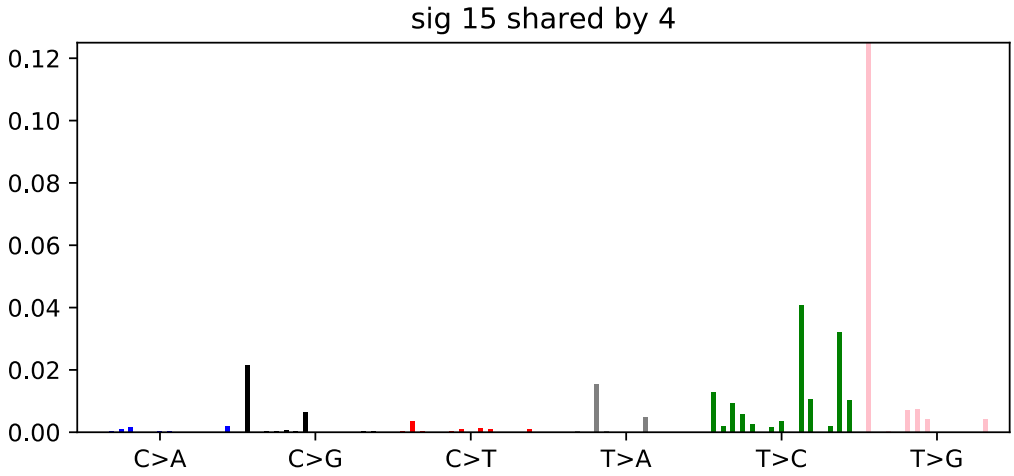
Signature 14 was the dominant signature in:

56% of g_mycobacterium_abscessus strains

46% of g_escherichia_coli strains

39% of c_salmonella_enterica strains

37% of g_salmonella_enterica strains



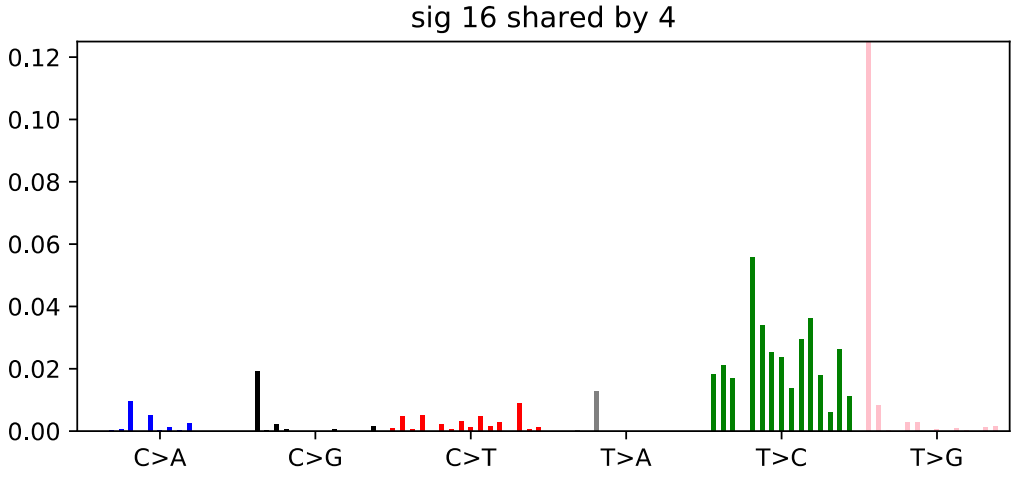
Signature 15 was the dominant signature in:

4% of r_enterococcus_faecium strains

4% of c_klebsiella_pneumoniae strains

3% of r_klebsiella_pneumoniae strains

2% of r_neisseria_gonorrhoeae strains



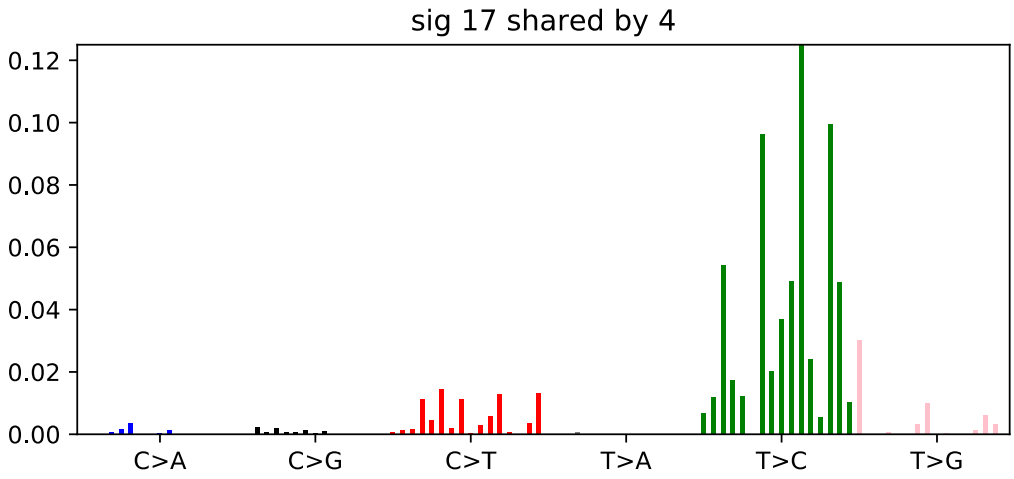
Signature 16 was the dominant signature in:

17% of g_salmonella_enterica strains

8% of g_pseudomonas_aeruginosa strains

6% of g_neisseria_gonorrhoeae strains

3% of r_neisseria_meningitidis strains



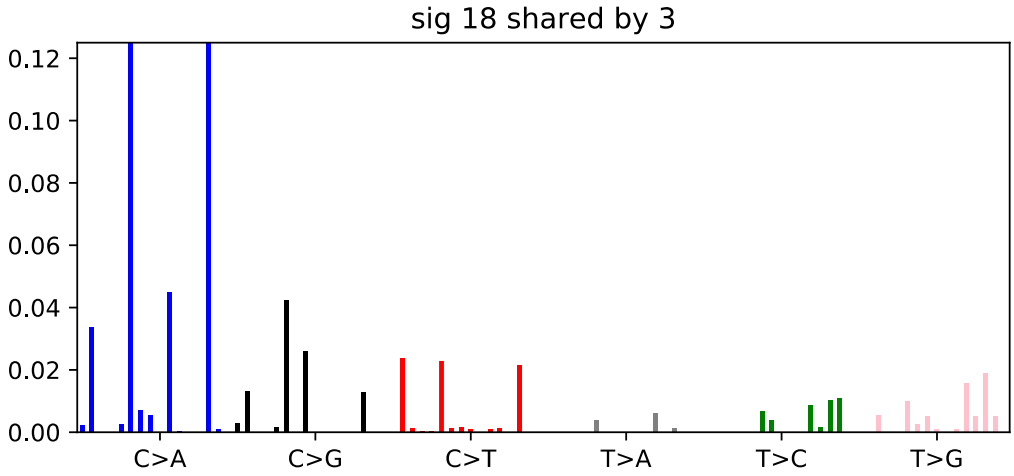
Signature 17 was the dominant signature in:

71% of r_neisseria_meningitidis strains

48% of r_neisseria_gonorrhoeae strains

32% of g_neisseria_gonorrhoeae strains

14% of r_salmonella_enterica strains

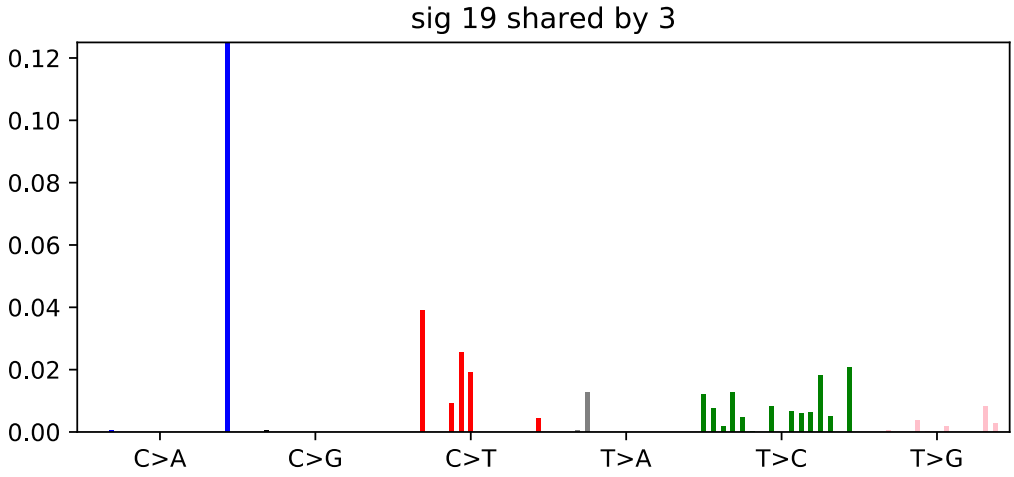


Signature 18 was the dominant signature in:

25% of c_mycobacterium_tuberculosis strains

17% of c_salmonella_enterica strains

7% of r_escherichia_coli strains

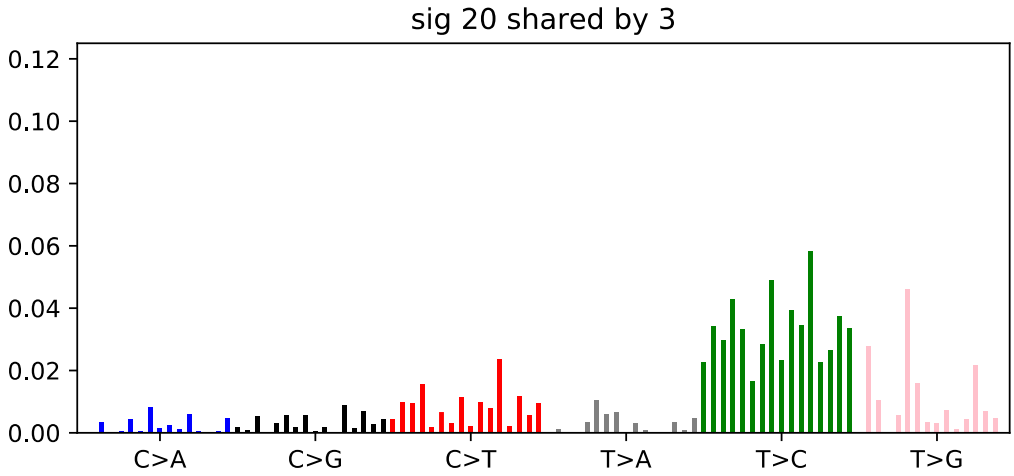


Signature 19 was the dominant signature in:

4% of c_bacillus_cereus strains

2% of r_bacillus_cereus strains

2% of r_acinetobacter_baumannii strains

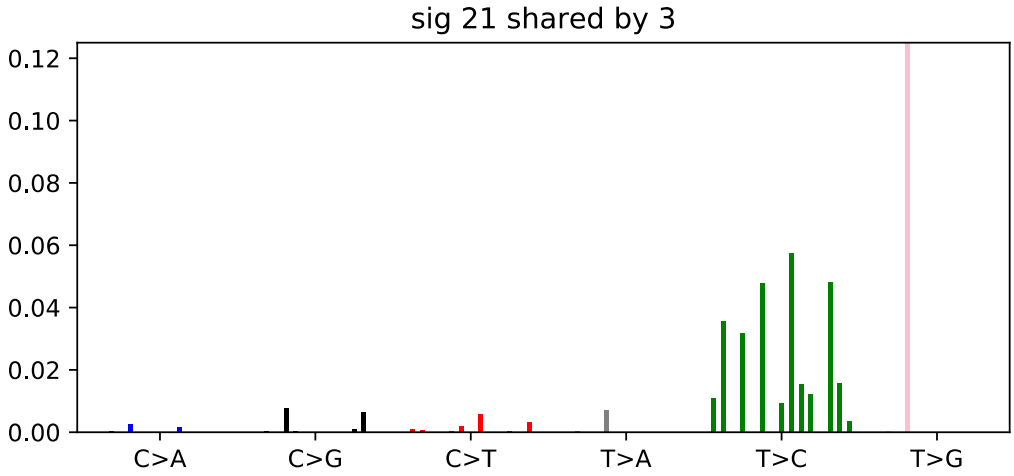


Signature 20 was the dominant signature in:

80% of r_escherichia_coli strains

18% of r_mycobacterium_abscessus strains

17% of g_escherichia_coli strains

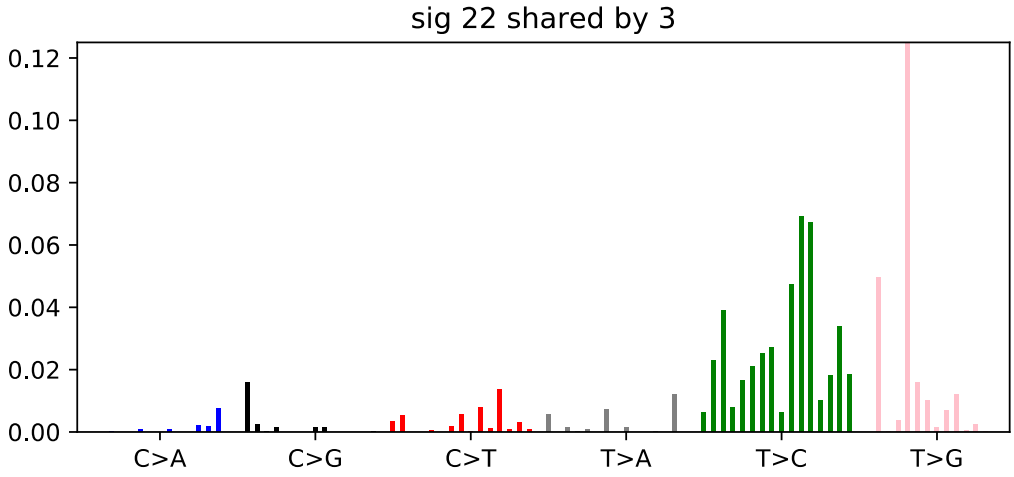


Signature 21 was the dominant signature in:

13% of r_neisseria_gonorrhoeae strains

12% of r_neisseria_meningitidis strains

6% of g_neisseria_gonorrhoeae strains

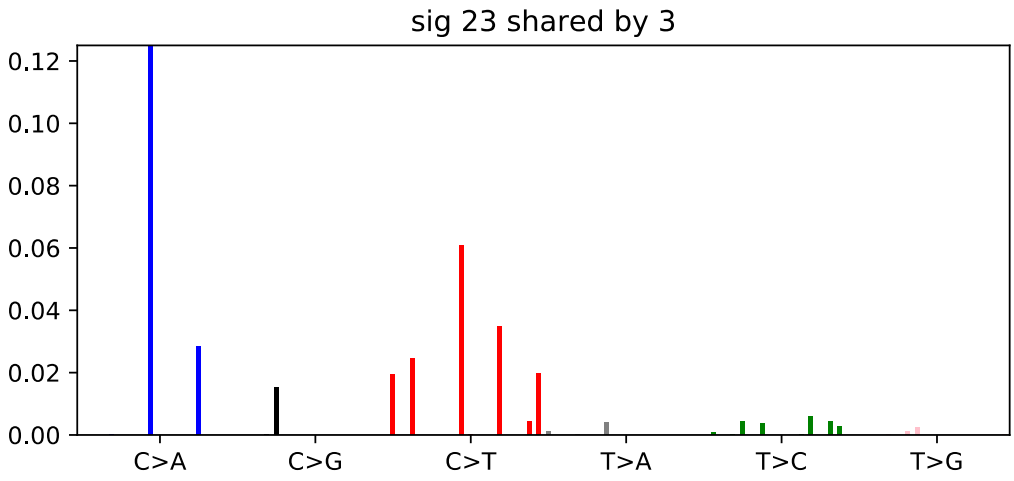


Signature 22 was the dominant signature in:

15% of c_salmonella_enterica strains

13% of g_salmonella_enterica strains

7% of r_klebsiella_pneumoniae strains

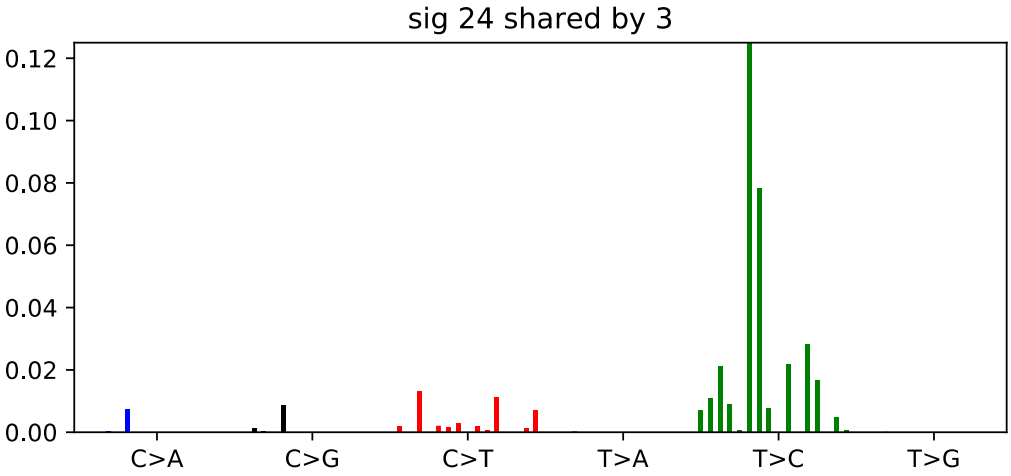


Signature 23 was the dominant signature in:

5% of g_listeria_monocytogenes strains

4% of r_bacillus_cereus strains

4% of c_bacillus_cereus strains

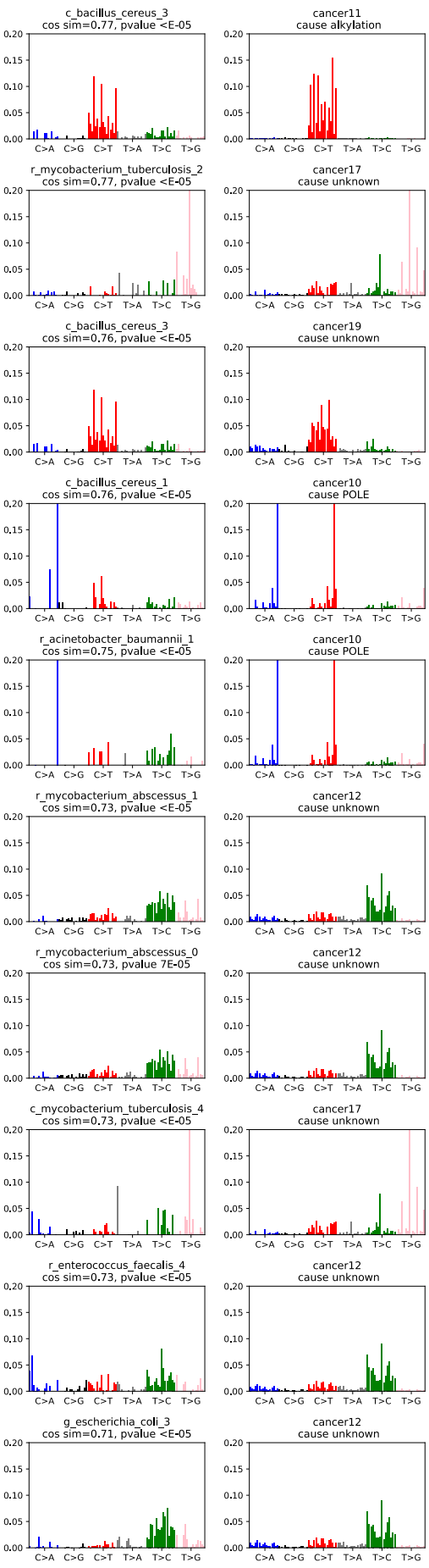


Signature 24 was the dominant signature in:

12% of g_neisseria_gonorrhoeae strains

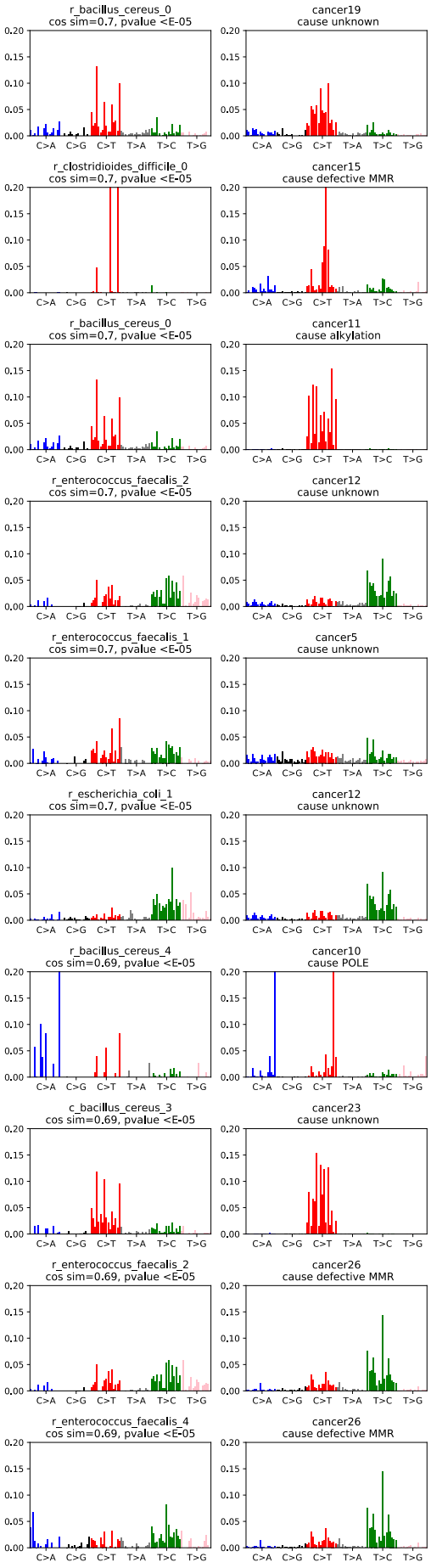
9% of r_salmonella_enterica strains

4% of r_neisseria_meningitidis strains

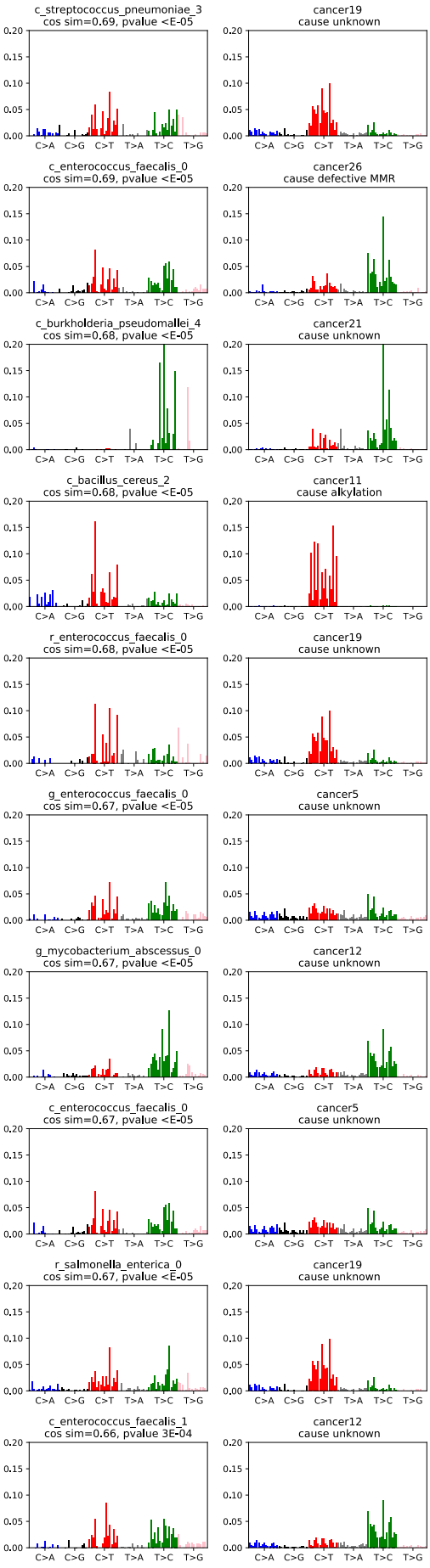


6.3.6 Supplementary figure 4.6

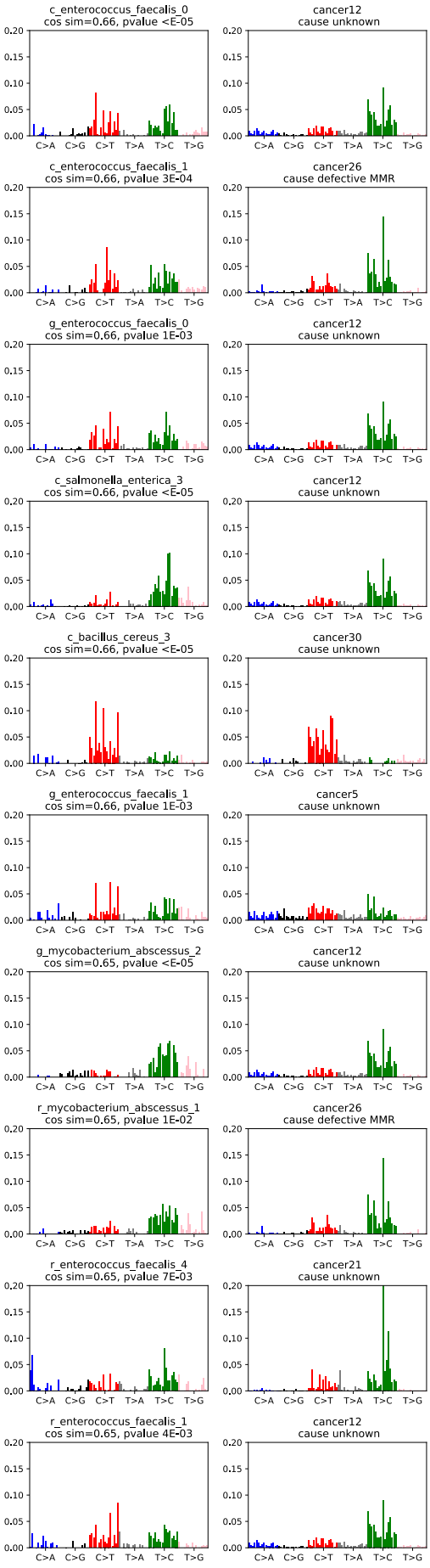
Supplementary figure 6a. Pairs of bacterial and cancer signatures showing a high cosine similarity.



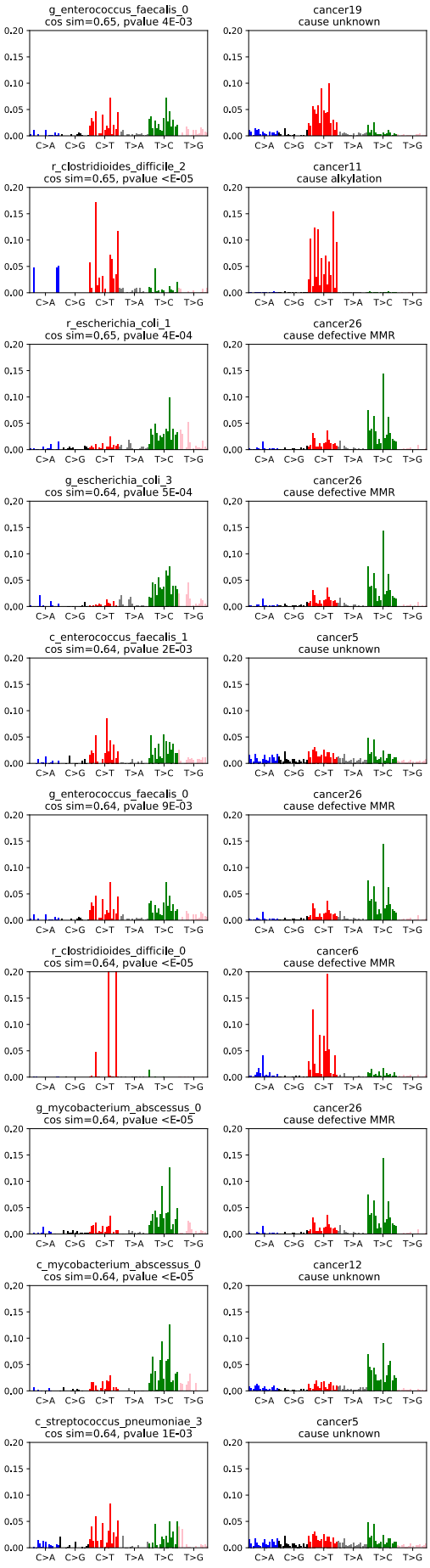
Supplementary figure 6b. Pairs of bacterial and cancer signatures showing a high cosine similarity.



Supplementary figure 6c. Pairs of bacterial and cancer signatures showing a high cosine similarity.



Supplementary figure 6d. Pairs of bacterial and cancer signatures showing a high cosine similarity.



Supplementary figure 6e. Pairs of bacterial and cancer signatures showing a high cosine similarity.

Supplementary figure 6f. Pairs of bacterial and cancer signatures showing a high cosine similarity.