Developing A Computational Approach To Investigate The Impacts of Disease Causing Mutations On Protein Function

Camilla Sih Mai Pang

A thesis submitted for the degree of

Doctor of Philosophy

January 2018

Institute of Structural and Molecular Biology

University College London

I, Camilla Sih Mai Pang confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Camilla Sih Mai Pang

January 2018

I dedicate this PhD thesis to my parents:

Sonia Louise Pang

&

Peter Tak Chung Pang.

Abstract

This project uses bioinformatics protocols to explore the impacts of non-synonymous mutations (nsSNPs) in proteins associated with diseases, including germline, rare diseases and somatic diseases such as cancer. New approaches were explored for determining the impacts of disease-associated mutations on protein structure and function. Whilst this work has mainly concentrated on the analysis of cancer mutations, the methods developed are generic and could be applied to analysing other types of disease mutations. Different types of disease causing mutations have been studied including germline diseases, somatic cancer mutations in oncogenes and tumour-suppressors, along with known activating and inactivating mutations in kinases. The proximity of disease-associated mutations has been analysed with respect to known functional sites reported by CSA, IBIS, along with predicted functional sites derived from the CATH classification of domain structure superfamilies. The latter are called FunSites, and are highly conserved residues within a CATH functional family (FunFam) – which is a functionally coherent subset of a CATH superfamily.

Such sites include key catalytic residues as well as specificity determining residues and interface residues. Clear differences were found between oncogenes, tumour suppressor and germ-line mutations with oncogene mutations more likely to locate close to FunSites. Functional families that are highly enriched in disease mutations were identified and exploited structural data to identify clusters within proteins in these families that are enriched in mutations (using our MutClust program). We examined the tendencies of these clusters to lie close to the functional sites discussed above.

For selected genes, the stability effects of disease mutations in cancer have also been investigated with particular focus on activating mutations in FGFR3. These studies, which were supported by experimental validation, showed that activating mutations implicated in cancer tend to cause stabilisation of the active FGFR3 form, leading to its abnormal activity and oncogenesis. Mutationally enriched CATH FunFams were also used in the identification of cancer driver genes, which were then subjected to pathway and GO biological process analysis.

Impact Statement

The work presented in this thesis describes a novel approach for detecting and analysing disease variants on a protein and cellular level. The approach was mainly applied to cancer as there are now significant amounts of data on disease variants linked to cancer. It used a multi-level strategy which first aggregated mutations across protein domains in families, to find mutationally enriched families. It then considered how cancer mutations cluster in 3D protein structures as a mechanism for finding putative driver mutations. By mapping putative driver genes to cellular pathways it was possible to analyse the affected cellular pathways. Structural analyses of the location of the 3D clusters highlighted how protein functions are affected and which protein regions are most targeted in cancer. Novel driver mutations were identified for Fibroblast Growth Factor receptor, a protein implicated in bladder cancer.

This work serves as valuable step in making sense of the vast amounts of next generation sequencing (NGS) data, by bridging genetic research to protein science. Since 2017, the methodology developed has formed the basis of ongoing collaborations with the Swanton research group in the Francis Crick Institute to analyse disease variants in lung cancer, linking protein bioinformatics to a clinical context. The approach is helping to pinpoint driver events in cancer - to find molecular convergence within heterogeneous genetic data, and provide a coherent explanation for disease pathogenesis. In turn this will hopefully guide drug treatment and provide insight into the nature of cancer evolution. In addition to cancer, this methodology can be applied to other disease types eg rare genetic diseases. It could also be applied to study the mutations in bacterial proteins that result in antibiotic resistance.

In terms of commercial activity within front line healthcare, the methods presented in this work could form the basis of a diagnostic characterisation tool for genetic variants, providing clinicians insight into the analysis of NGS data, and in turn guiding patient treatment. Because these methods characterise genetic variants, when linked with drug response data, they can help advance personalised medicine by increasing drug specificity, and reducing side effects - thereby increasing the quality of therapy for the patient.

The tools developed will also be valuable to pharmaceutical companies, to aid drug trials by providing complementary insights that could help rationalise structure-activity relationships for drug lead selection, thereby increasing efficiency and decreasing costs within drug lead optimisation.

As well as being reported in scientific journals, the outputs of the published research will be discussed on an online podcast within the International Creative Disturbance media outlet. This posting will highlight the analogy between protein and disease evolution and evolution in the artificial intelligence and creative arts industries - thereby bridging the gap between the intricate details of fundamental scientific research and everyday phenomena.

Acknowledgements

I would firstly like to thank the Institute of Structural and Molecular Biology for accepting me on the Wellcome Trust Interdisciplinary PhD research programme. My sincere gratitude goes towards my primary supervisor, Professor Christine Orengo. Christine has been my academic equivalent of a mother, who has been indispensable to me completing this project and the writing of this thesis. Even though the PhD ends this year, I will be forever thankful for the lessons I have learnt from her and how much she has taught me, along with her endless support, patience, and belief in me. I would also like to thank and simultaneously apologise to her partner for stealing their breakfast times for the marking of this thesis. I would also like to thank my secondary supervisor, Dr Andrew Martin, who has been supportive throughout my project, in always being there to answer any queries I have with my work, and for his great advice.

I am also grateful to the CATH lab, for their constant support and encouragement throughout this project, along with them tolerating my dark sense humour and perpetual swearing at the computer. In particular, I would like to thank postdoc Dr Paul Ashford, who has been a great "mutation" buddy to work with in this project, and for his great energy and awful puns. Additionally, thank you to Ian Sillitoe and Tony Lewis, for their support in helping me code, and always being there to explain to me how I have somehow managed to freeze my computer. I would also like to thank my lab husband, SuDatt Lam, who has literally been at my side (right hand side to be specific), throughout the 3 years. My thanks also go to Sayoni Das, Natalie Dawson, and Tolulope Adeyelu for their help with LaTeX, R, and discussions on CATH FunFams, and also for being great friends. I don't know what I would have done without the support of my london friends, and in particular I would like to thank Abbie, Amandene, Pip, Maisa, and Greg for their support during and their tolerance of me completing this thesis.

My eternal gratitude goes towards my family, who have been a constant positive force and encouragement for me throughout my project and all of my life. Thank you to aunty Sue and uncle Rob for letting me stay with them when visiting London for meetings, uncle Mike for fixing my laptop more times than desirable, and uncle John for his maths brain. Thank you to all of my little siblings, Tiger, Lilly, Aggie, and Roo for always keeping me on my toes, and grounding me to be my best self for them. Talking of siblings, I would like to say a massive thank you to my big sister and best friend Lydia, who has been my inspiration to be myself and to embrace my differences, but most of all for always understanding me and being there for me.

And finally, mum and dad, thank you thank you thank you for everything. Words can't express how loved and supported I have felt from my very first day on this planet to today. You have always believed in me, and now I do too.

Contents

List of Figures

List of Tables

Chapter 1

Introduction To Thesis

The efforts of various genome sequencing studies has shed light on the profuse amount of sequencing data, and on the importance of mutations within genetic and somatic diseases. In this thesis, the analysis of missense mutations within different cancer types were mainly considered, but the methods developed can be applied to all kinds of disease causing mutation. Since cancer harbours an abundance of genetic mutations due to its high replicative turnover, here methods were developed to identify cancer driver genes, and to distinguish cancer mutations that drive carcinogenesis from passenger mutations.

This was done using CATH protein domain family data and pathway information to identify cancer driver genes and mutations, and to understand their mechanisms. This chapter provides a general overview to disease mutations and their impacts of protein structure and function, whilst also providing a review of the methods which use such impacts to predict pathogenicity.This thesis is structured so that at the beginning of each chapter there is a literature review for the work specific to the subject of the chapter in question, followed by the methods and results performed in this work.

It was found that using CATH protein domain data enhanced driver genes detection which were implicated in cancer related processes, providing a complementary list to other domain based driver gene methods. In addition to this, CATH protein domain data also enabled the identification of driver mutations which showed greater functional relevance than other disease mutations in both cancer and non-cancer disease cases.

Mutations

Germ line mutations are inherited from the parents, and can be associated with diseases, namely inherited diseases. Mutations within the embryo, post-fertilization are referred to as *de novo* mutations [255], and can be implicated in rare diseases [225], or can contribute to the progression of adult cancers [109]. After birth, mutations which occur

throughout life are called somatic mutations, and can lead to various diseases including cancer. These types of genetic variations can include; single nucleotide polymorphisms (SNPs), insertion and deletion events and copy number variants [228].

Common SNPs are defined as occurring in at least 1% of the 'normal' population and at an approximate frequency of 1 in every 1000 bases [235]. These SNPs can be both non-synonymous (ns) and synonymous(s), where the single base change affects and does not affect the encoded amino acid respectively. This project is concerned with nsSNPs which lead to an amino acid change. Both types of SNP can exert either a neutral or negative effect on the phenotype, and would thus be described as neutral or damaging mutations respectively [263], where the latter has been implicated in diseases [27]. The emergence of nsSNPs enables the widening of the protein functional repertoire in evolution and is an important contributor to allelic variation amongst individuals - affecting the efficiency and activity of cellular events [132][236]. Despite the benefits of mutations in creating variation in evolution, the presence of nsSNPs can manifest in different disease susceptibilities and different responses to drugs between individuals [52] [116] [144].

Mendelian disease and cancer

A study of mutations involved in Mendelian, rare diseases and cancer is presented in this thesis, where the former two diseases are caused by mutations in one gene, whilst cancer is a complex disease, generally caused by mutations in many genes [165] [243] [279]. Mutations in Mendelian diseases occur at low frequencies in the population, but the proportion of mutations that cause disease are high. This is defined as them having high penetrance and is hypothesised to be a consequence of their disruptive effects on the protein and phenotype, which means that they tend to be under negative selection [273] [232].

Cancer mutations have variable selection pressures, dependent on the nature of the mutation itself. There are mainly two types of mutation, drivers and passengers. Driver mutations are primarily responsible for the oncogenic phenotype and are advantageous to cell growth [238]. These are therefore under positive selection pressure in tumour

evolution, promoting oncogenesis and drug resistance [83] [18] [198] [107]. Passenger mutations are under neutral selection pressure, as they confer no survival advantage to the tumour. Although recently, there has been a suggestion of a different type of mutation that has characteristics of a passenger mutation which eventually becomes a driver mutation due to the presence of other mutations [170]. Because of this altered structural context, the passenger mutation contributes more directly to the oncogenic phenotype. This type of mutation is referred to as a latent driver mutation.

Recently, it has been shown that certain inherited mutations can predispose an individual to cancer by altering the structural context, leading to the tolerance of otherwise harmful cancer mutations, or acting alongside them to engender carcinogenesis or drug resistance, which is referred to as mutation co-morbidity [150].

Impacts of Mutations

The impacts of mutations can be divided into structural and functional effects on the protein. The former primarily affects attributes, such as the stability and/or fold of the protein product, and the latter affects functional sites. Both effects can be consistent with damaging nsSNPs occurring at conserved regions involved in protein stability, folding and function [70] [155] [219] [246] [253] [229] [231] [66]. In terms of the amino acid substitution itself, studies showed that cancer mutations and disease nsSNPs involve substitutions which are less conserved in their amino acid physiochemical properties than non-disease associated mutations - accounting for their abnormal effects with respect to the wild type [66].

Structural Impacts

Structural destabilisation can be caused by the formation of voids and clashes in the protein as a result of substitution to a smaller and bigger amino acid respectively [229]. Both clashes and voids cause disturbance within buried regions which can in turn affect residue packing [256] [39]. This is supported by studies of Stitziel *et al* [229] which found that 88% of damaging nsSNPs compared to 68% of non-disease associated nsSNPs

form voids within the core of the protein. Martin *et al* [144] investigated the effects of mutations in the tumour suppressor gene, p53. They showed that mutations from glycine and to proline resulted in p53 undergoing conformational rearrangements as seen in Ramachandran plots. In total, 7.7% of mutations caused steric clashes, and 48.1% mutations showed disruption of residues involved in hydrogen bonding. Other studies show cancer mutations from COSMIC [15] to occur in pockets of proteins [258].

Studies by Vitkup *et al* and others [256] [43] show that mutations from tryptophan and cysteine residues have the highest probability of causing disease, followed by those from arginine and glycine, the latter accounting for 30% of genetic diseases. This reflected their roles in forming the hydrophobic core, disulphide bond formation, salt bridges, hydrogen bond creation and protein flexibility, all of which contribute to the structural integrity and therefore the functional state of the protein. Stability is a major feature dictating the protein evolutionary rate, as it determines the tolerance to a particular mutation [248]. Many studies have shown that damaging nsSNPs impact protein stability, and that this is a major cause of pathogenicity for monogenic missense mutations [16] [83] [277]. Wang *et al* showed that 83% of disease-causing nsSNPs affect stability by 1-3 Kcal/mol [263] [277]. Such a decrease in protein stability can be invoked by mutations interfering with hydrogen bonding [38], resulting in backbone strain.

Non-synonymous SNPs (nsSNPs) in buried regions are less likely to be tolerated, as these regions show a weaker ability to compensate by local mutation - due to packing constraints and the altering of complex intra-molecular interactions [16]. These impacts are especially seen with Mendelian mutations, as they tend to occur more in buried regions compared to non-disease and cancer mutations [83] [135]. Disruption of intra-molecular interactions are also seen if a mutation affects hydrogen bonding and disulphide bonds [30] [274] [62], both of which are fundamental to acquiring and maintaining the native protein fold. Other structural effects of a mutation include aggregation and the formation of amyloid-beta proteins due to changes in hydrogen bonding [225]. Such effects have been seen in recent studies that measured both the stability and aggregation propensities of proteins, showing that disease-associated mutations increase the protein

aggregation potential compared to neutral polymorphisms [45].

Although structural aggregation is generally regarded as a loss in protein function, reviews of p53 demonstrate that mutations within it can lead to neomorphic changes in p53 function. Mutations at aggregation prone sites in p53 can result in self-aggregation, and sequestering of the functional wild type p53 forms. [261] [161]. In addition to this, mutations can also affect the charge distribution in the protein, which can alter pH dependence and catalysis [225]. An example of the latter includes the L861Q mutation in EGFR (Epidermal Growth Factor Receptor), where a mutation to a polar residue causes a dramatic conformational change of the activation loop, favouring the active state [56].

Functional Impacts

Other damaging effects of mutations include impacts on protein function [165]. Mutations can affect specific functional sites, resulting in the increasing, decreasing or switching of protein function, all of which can lead to disease [248]. This may involve a mutation affecting catalytic residues or neighbouring regions in the active site thereby interfering with the chemistry or affecting substrate and metal binding [160] [230]. It is important to note that not all proteins have evolved to maximise stability, lower stability being a trade-off for their functioning [225] [248]. There are some functional sites that are inherently unstable, and mutations in these residues can even result in stabilising these structures, for example mutations involved in cancer progression [160][260] [9] [54] [177].

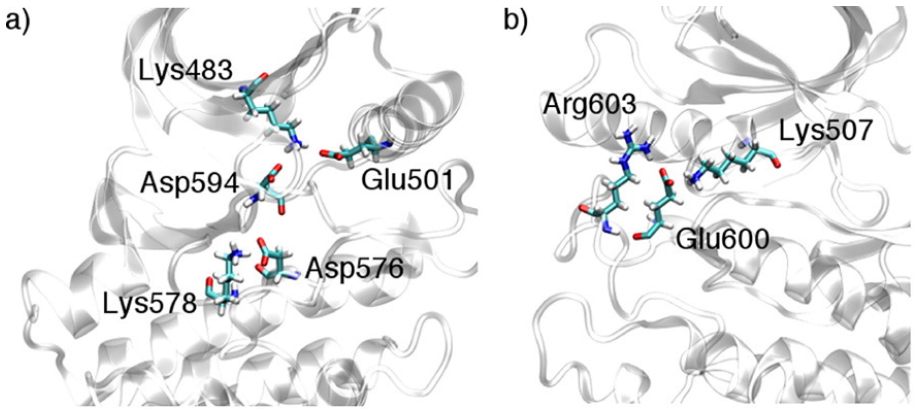
Catalytic Sites

An example of a mutation leading to a loss of function was reported for the tumour suppressor gene, p53 [144], where mutations occurring in DNA binding sites caused a change in solvent accessibility of $\geq$ 5% for the site between bound and unbound DNA. All residues within these sites were found to have mutations likely to be decreasing DNA binding and therefore the tumour suppressive role of p53.

In contrast, a mutation can also lead to an increase in function. This is demonstrated by the V600E mutation in the B-Raf kinase activation loop, implicated in malig-

nant melanoma. This mutation causes an increase of 500 fold in basal B-Raf kinase activity, causing abnormal oncogenic signalling [260], by affecting local electrostatics. Recent studies by Marino *et al* [143] show that such functional effects are due to a novel salt-bridge within the mutant structure which in turn increases the barrier for transitioning to the inactive state, as well as increasing the flexibility of the activation loop, thought to promote its phosphorylation.

Figure 1.1: Formation of a salt-bridge in the V600E mutant of BRAF. Key salt-bridge forming residues in (a) the active site of both WT and mutant B-Raf and (b) mutant B-Raf only. Taken from [143].



In addition to loss and gain of function, a mutation can also result in a switch of function, by altering the active site chemistry or substrate specificity of a protein [4] [37]. For example, this is seen in enzymes with promiscuous activity which bind many substrates [78] and enzymes involved in antibiotic resistance e.g the beta-lactamases [4]. Other types of switch of function include mutations that occur in protein interaction interfaces, where they can affect the affinity and/or specificity of binding, altering protein binding profiles and downstream networks [37] [169]. Mutations can alter post-translational-modifications (PTM), by affecting the protein motif itself or by affecting the local arrangement of the site, both important for PTMs [252]. This in turn can affect the recruitment, activity and maturation of proteins [196].

Disordered Regions

Other kinds of functional sites include conserved sites in disordered regions. These include short linear motifs (known as SLIMS) which are involved in transient and tuneable interactions such as those involved in transcription, regulation and signalling [53]. Specific roles of these motifs include binding of nucleotides and ligands, and forming sites for post translational modifications such as phosphorylation, methylation and ubiquitination. Due to the sequence variability of these regions, they are less likely to be under negative evolutionary pressures and are therefore more likely to tolerate mutations [43] [135]. Despite this scope for variability, these transient interactions are nevertheless precise and mutations can result in a loss of wanted, or gain of unwanted, interactions. This can alter binding specificity, affinity and accessibility, as well as affecting protein targeting, all of which can lead to disease [98] [283].

Examples of disease causing mutations affecting disordered regions include those in Noonan-like syndrome, Rett syndrome, Liddles syndrome [53] [252] and cancer [103]. Studies have shown that 20-25% of disease mutations obtained from UniProt, occurred in regions with predicted disorder [171] and IUPred [59]. 20% of the disease-associated mutations affected the disorder to order transitions. This in turn affects binding and activity, leading to altered signalling, gene expression and regulatory networks.

However, studies be Lu *et al* [135] show that pathogenic mutations in germ line diseases (from OMIM) [10] and cancer (from COSMIC) [15] had high propensities to occur in ordered regions, measured by DISOPRED [264] within Pfam domains, especially for germ line mutations. Furthermore, disease mutations reported in the UniProt Humsvar and HGMD databases also showed a preference to occur in protein interaction sites of low flexibility, measured by the Dynamic flexibility index (DFI) derived from molecular dynamic simulations [24].

Allosteric sites

Allostery is a form of protein regulation, whereby sites remote from the active site can have an effect on function. Effectors such as cofactors and ligands binding to these

sites can control protein function. Effector binding can be accompanied by large or small conformational changes, such as domain or residue rotations respectively. Mutations in allosteric sites, or communication paths between allosteric and functional sites, can perturb allostery and result in deregulated function [124]. Dixit *et al* showed that oncogenic mutations in kinases can occur in regions susceptible to conformational transitioning, due to their inherent instabilities. These sites tend to overlap with sites of allosteric importance [54]. In addition, oncogenic mutations have been shown to alter both the flexibility and energetics of residues distal to the mutated site, favouring the active states in ABL and EGFR kinase. This is also seen in studies which show that cancer driver mutations in BRAF and JAK2 affect structural disorder at distal sites by means of long-range coupling [135]. These studies, among others [88], suggest that disease causing mutations can be far from conserved and functional sites in the protein structure and yet still have an effect on function.

Other studies by Torkamani *et al* [250] and Dixit *et al* [56] reported germ line mutations in kinases occurring in regions not directly involved in ATP binding and catalysis [252] [66] [5]. This is seen in dystrophin and Ras, where mutations can have structural effects over long distances [68] [53]. Although the structural explanation of long distance effects of mutations remains controversial, suggestions include changes of charge distributions and disruption of packing in the protein core, along with affecting the conformational landscapes of proteins [230] [5].

In 2016, Clarke *et al* developed a method called STRESS [35], which predicts two types of allosteric site: using Monte Carlo simulations and normal mode analysis to identify surface-critical and interior critical residues respectively within a protein structure. Mutations from the 1K genomes project of varying minor allele frequencies (MAF) were mapped to protein structures. They found that both types of predicted allosteric residues had fewer rare mutations. In addition to this, they found that variants within these critical residues possessed significantly higher pathogenicity scores reported by Polyphen relative to non-critical residues, highlighting the potential pathogenicity of these predicted allosteric residues when mutated.

Cancer

Unlike mutations in Mendelian diseases, the effects of mutations are less well defined in cancer. This is partly because it depends on whether the genes affected are oncogenes or tumour suppressors. In a tumour-suppressor gene, the effects are similar to Mendelian disease mutations and tend to be detrimental structural effects, including decreasing the stability and activity of the protein [70] [161] [226]. Specifically studies by Shi *et al* showed that 50-60% of mutations in tumour suppressors decrease stability and result in impairment of protein activity [214]. Studies also showed that mutations in tumour suppressors occur more at solvent inaccessible/buried sites, similar to Mendelian diseases [214][271] [61].

In contrast, driver mutations in oncogenes tend to have less disruptive effects on the wild type protein, supported by studies which showed a lack of structurally deleterious effects for oncogenic mutations compared to mutations in tumour suppressor genes in cancer [61]. Additionally, other studies on protein stability show mutations in oncogenes to be 4 times less likely to destabilise the protein than those in tumour suppressors [226]. In cases where drivers in oncogenes do affect stability, studies show destabilisation of kinase inactive states, where oncogenic drivers shift the equilibrium towards more active kinase forms [230]. This is consistent with these mutations being often associated with an enhancement of activity, driving cancer progression. Interestingly, studies have shown driver mutations in oncogenes to be in loops and unstructured regions, on the protein surface in hydrophilic areas [83] [226] [61]. These may correspond to functional sites as it has been shown these are often in catalytic loops such as in kinases [54] and in the TIM barrel fold which is present in many enzyme structures [210]. The lack of structural constraints associated with disordered regions may explain why some cancer mutations are tolerated [267], but this subject is controversial.

Analysing Mutations

There are various public resources and tools described below, which analyse the effects on structural and functional features, to determine the impacts of a disease causing

mutation on the protein.

SAAP Data Analysis Pipeline (SAAPdap)

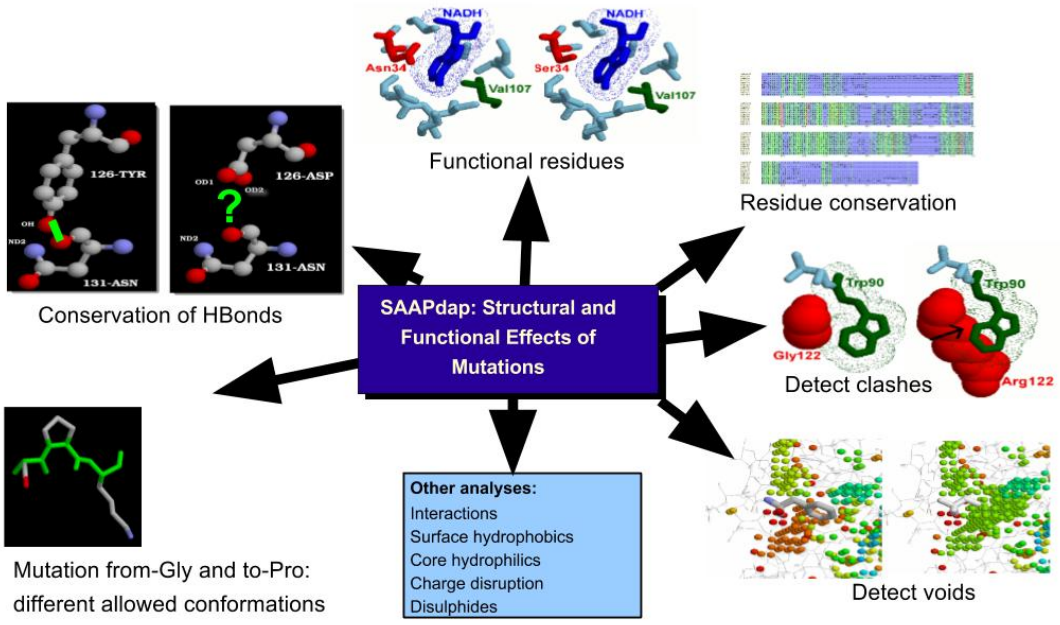
The on-line SAAP (Single Amino Acid Polymorphism) database (SAAPdap) [101], uses data from PDB, dbSNP, OMIM and other LSMDB (Locus-Specific Mutation Databases) for mutation analysis, and pre-calculates the effect of a given mutation based on a series of structural analyses in known structures. These include effects on hydrogen bonding, salt-bridges, conserved sites, UniProt annotated functional sites, clashes and voids. The conserved sites used here are measured using an in-house methodology, ImPACT, which identifies residues that have 'significant conservation' across a diverse selection of species, which suggests functional importance. The effects, if any, are reported for each mutation within a UniProt entry. The authors also developed a mutation predictor, called SAAPpred, which uses the reported effects to determine whether a given mutation is pathogenic or not [9].

A recent study performed by Baraessia and Pearl *et al* analysed the structural effects of a cancer driver mutation – V600E- on the active and inactive state of the BRAF oncogene [14]. This found that the driver mutation is 'structurally tolerated' and does not cause any structural effects within the active conformations – as reported by the SAAPdap analysis. In contrast, in the inactive BRAF state, the V600E was predicted to cause protein destabilisation by producing buried charges within the protein, thereby favouring the active conformation of BRAF.

Platinum

Other tools characterising mutations include the Platinum database developed by Pires *et al* [186]. Platinum is a manually curated and literature derived resource which uses in-house prediction tools to determine the effect of a mutation on ligand binding affinity for known ligand bound complexes in the Protein databank (PDB). Such tools include the stability predictors DUET [187], and mCSM-PPI[185]. Although Platinum does not provide an online prediction tool for query mutations, it nevertheless provides a compre-

Figure 1.2: Summary of the structural and functional effects reported by SAAPdap. Provided by Dr Andrew Martin



hensive resource for characterised ligand binding complexes with known mutations from the literature.

Cancer and kinase specific analysis tools

There are also specific tools and resources for analysing cancer mutations. Cancer3D is a database [190] which provides information on whether a mutation is a driver, obtained by using the e-driver method to analyse its occurrence in regions or domains in Pfam, associated with cancer [189]. Cancer3D also applies the e-drug method, which provides a residue level annotation of drug specificity and activity, creating a bio-marker potential. This uses data from the CCLE database which contains mutation data and corresponding drug activities in 906 human cancer cell lines [17]. Similarly the canSAR database [22] reports the propensity of both missense and copy-number variations at particular sites and their drug sensitivities, but focuses on the whole protein level and integrates protein networks.

One of the main goals in cancer bioinformatics is to identify driver genes by analysing

mutational frequencies and functional impact scores. Despite this being effective for a subset of drivers, such methods do not distinguish between driver oncogenes and tumour-suppressor genes which differ in these properties [226] [247] [91].

Kinase-specific mutation analysis tools include wKinMut [104] and MoKCa [199]. The former incorporates data from COSMIC and UniProt and performs structural analysis of the mutation using SAAPdap. The wKinMut tool also employs mutation predictors, such as SIFT [164] and MutationAssessor [198] described below. MoKCa is a similar tool which deals with mutations from the Cancer Genome Project and COSMIC which are then mapped onto Pfam kinase domains, where PROSITE annotations were used to identify functional residues.

Predicting SNP Pathogenicity

Various methods attempt to distinguish disease-causing mutations from neutral nsSNPs, as well as identifying the effects on protein structure and function that are likely to be disease-causing. Such pathogenicity prediction tools can facilitate both diagnostics and therapeutics, by prioritising mutations for study and as drug targets. The main approaches used in mutation pathogenicity prediction are based on sequence and structure based features, and are still limited in their accuracy [191].

Methods For Assessing Changes In Sequence

Sequence based methods largely exploit the effect of mutations on sequence conservation at a given residue position. Studies in 2001 by Miller *et al* [155] used a sequence and phylogeny based prediction to analyse several disease genes. Since then there have been a plethora of methods that predict the severity of a given mutation based on the pattern of amino acids observed at that position [165].

SIFT

SIFT scores the pathogenicity of a given mutation based on how well it is tolerated at a particular position within a given alignment of homologous sequences, obtained from

Swissprot/TrEMBL [164]. The score is based on the type of amino acid introduced, along with its prevalence in the alignment position, and is therefore dependent on the diversity of the aligned sequences. Changes to a completely different amino acid at positions with conserved wild type character will therefore be considered deleterious. The false positive error of SIFT is 20% and it has been used in many Meta-Predictors [84](discussed below).

FATHMM

Another sequence based method is called FATHMM, which scores the deleterious effects of mutations within HMMs derived from proteins and domains [215]. Such HMMs are created from homologous sequences derived from UniRef90 [11] and were further annotated using domain sequences within SUPERFAMILY [89] and Pfam [222]. Amino acid characteristics of both native and mutant residues are then compared to amino acid divergences at a particular HMM position, measured by the Kullback-Leiber methodology. Amino acid sequences for the family are taken from Swissprot/TrEMBL [11]. The effects of a mutation are based on whether it alters the amino acid propensities at a given position. One of the main advantages of this method is the use of species specific weightings with respect to human mutations, based on the frequency of known disease and neutral mutations from HGMD [227] and Uniprot HUMSVAR [11] respectively.

The Use Of Protein Subfamily Conservation

Some approaches detect preferentially conserved sites in a protein subfamily. If the subfamilies are functionally coherent, these partially conserved sites are likely to be associated with functional specificities. ManChon *et al* [140] showed that partial conservation i.e in subfamilies, is one of the top features for identifying SNP pathogenicity. Related studies explored signalling specificities in the kinome [37].

Mutation Assessor

Existing SNP prediction methods which use subfamily conservation usually generate HMM profiles to represent the sequence profile of a protein family and/or subfamily [198]. Mutation assessor (MA) which produces a Functional Impact Score (FIS) of a mutation, by averaging a conservation score and specificity component score based on the conservation in a protein family and subfamily respectively. The Combinatorial Energy optimization method (CEO) methodology clusters sequences within protein family and subfamily, where the latter groups were optimally separated based on a set number of 'specificity residues' which differ between these subfamilies [197]. Although MA has been shown to be an effective method for predicting SNP deleteriousness, reaching 79% in recognition accuracy, it doesn't distinguish between gain of function and loss of function mutations [183], nor does it distinguish between the different types of conservation scores derived from protein family and subfamilies. MA predictions were validated using disease mutations from OMIM [10], making it more appropriate for assessing the pathogenicity of germline disease mutations as opposed to cancer mutations.

HMMvar-func

A more recent derivative of Mutation Assessor, HMMvar-Func [134], uses the same method of identifying protein family and subfamily clusters as MA [198], but distinguishes mutations that cause gain, loss, switch or conservation of protein function. The 4 prediction classes are based on the different combinations of two logistic equations describing the probabilities of a mutation to cause a loss of function (LOF) or acquire a new function (GOF). The LOF and GOF equations themselves are derived from residue propensities in the HMMs of protein family clusters, from the wild type sequence and mutant sequence respectively.

Methods for Predicting The Functional Effects Of SNPs – Using structure and functional site data

There are limitations to sequence based approaches as they do not take into account the structural impact of the mutation likely to be important for an accurate nsSNP prediction and exploit only indirect functional information [107] [262] [57][204]. Incorporation of structure enables the mapping of the mutation in 3D and the use of structural characteristics, such as stability, intra-molecular interactions, surface accessibility, and conservation. Other features include functional site data in a 3D context. Studies by Saunders *et al* [204] highlighted the advantage of incorporating structural terms such as solvent accessibility and the C-beta density (as a proxy for degree of burial) in deleterious nsSNP prediction, especially in the case where few homologues are available.

SNPeffect

SNPeffect assesses whether a mutation affects various individual attributes related to protein homeostasis, which are then combined to produce an overall prediction score[196] [195]. Such attributes include; the co-location to catalytic sites from CSA [192], effect on protein aggregation and amyloid predictions measured by TANGO [64] and WALTZ [147] respectively, and protein stability using FoldX [208]. Variant data from the HUMSVAR database are mapped onto known protein structures and their occurrence in known and predicted functional sites is assessed.

PolyPhen-2

Studies by Sunyaev *et al* analysed the effect of mutations on various structural features [231], which serve as inputs for a probabilistic Naive Bayes classifier to build the SNP predictor PolyPhen, and the more recent PolyPhen-2 (PPH2) [6]. The structural features used in PPH2 include: solvent accessibility, B-factor, CpG context, position of mutation within a Pfam domain, change in residue volume and the difference in PSIC (Position specific independent counts) scores between wild type and mutated residue. The latter

feature reflects the propensity of an amino acid to occur in a position, considering the pattern of amino acid substitutions within an observed alignment using the PSIC algorithm [233]. Benchmarking showed that PPH2 is more effective than Polyphen, measured by true positive percentage values for damaging mutations (93%) in Mendelian diseases in UniProt HumDiv mutations in a mixture of diseases (cancer and non-cancer) in Uniprot HumsVar (73%). This further highlights the importance of considering specific types of genes and types of diseases in predictors of mutation pathogenicity.

SAAPpred

SAAPpred uses structure based features from SAAP analysis pipeline (SAAPdap) [101], as previously discussed (see section 1.4.1) to obtain a SNP prediction of pathogenicity [9].

SNPs & GO 3D

SNPs & GO and its successor SNPs & GO 3D incorporates Gene Ontology information, along with sequence and structural based features to describe the mutated residue environment respectively. These and other features are inputs to a support vector machine for SNP prediction [26] [157].

SuSPect

SuSPect uses sequence, structural, functional, and network features to predict a SNP phenotype [272]. Disease and neutral mutation data were taken from UniProt HUMSAVAR [11], dbSNP [113], and PhenCode [81], where the disease mutations were from a mixture of different disease types. A number of features were combined in a machine learning SVM including: sequence conservation measured by Jensen-Shannon divergence, difference in propensities of the wild type and mutant amino acids in a Pfam HMM position, difference in PSSM scores between wildtype and mutant in an alignment of UniRef50 sequences, protein solvent accessibility using NACCESS [127] and NetSurfP [180] and protein network centrality using DOMINE [275]. Protein-protein in-

teraction (PPI) based features were found to be beneficial for phenotype prediction.

Other studies which have used known structures in nsSNP prediction include the probalistic classifier of Chasman *et al* [30], and SNPs3D by Yue *et al* [278] which have been described in various reviews [115].

Cancer Specific Prediction Methods

The methods described above, although effective for predicting pathogenicity for germline diseases, are often less effective for cancer mutation prediction. This is because cancer mutations appear to have rather different effects on proteins [226]. CanPredict [111] is a classification method trained on known cancer variants from COSMIC [15], Mendelian disease variants from UniProt and common variants from NCBI. Features include: prediction scores from SIFT and gene type similarities classified using a Gene Ontology similarity score.

More structure-based approaches for identifying cancer mutations include the recently developed Index of Carcinogenicity (InCa) [62]. This used neutral mutations from the 1K genomes project [3], somatic cancer mutations from COSMIC and known driver mutations in oncogenes and tumour suppressors from the Vogelstein cancer gene list [257]. These mutation groups were ranked according to various features including; relative accessible and buried surface area, secondary structure, physicochemical similarity, co-location to in house PISA derived interfaces and effects on the structural environment of the mutation. The structural environment was measured using the normalised frequencies of all amino acids in the vicinity (5 Å) of the mutation. Features were incorporated in a machine learning, random forest algorithm to give an Index of Carcinogenicity (InCa) score for identifying a cancerous phenotype. [28].

Meta-Predictors of SNP pathogenicity

As described above, there are many different mutation predictors using a range of features to assess mutation pathogenicity. Meta-Predictors have been developed which combine such methods enabling feature complementarity in SNP predictions. Most of

them incorporate SIFT, mutation assessor and Poly-phen2 in their predictions. The most popular meta-predictors include CONDEL [84], and IntOgen [86] which are designed for general and cancer mutation predictions respectively, where the latter focuses more on identifying driver genes within cancer. IntOgen uses the methods in CONDEL (such as SIFT and Mutation Assessor), and also incorporates scores describing the degree of mutation clustering within genes using OncoDriveCLUST [240] and a combined functional impacts score from OncoDriveFM [85].

Limitations of current prediction methods

Although meta-predictors such as IntOgen [86] use a whole range of features to identify driver genes based on their mutations, they rarely consider the different types of disease, nor the different types of mutations seen within cancer genes, all which can diverge in their protein clustering patterns, functional impacts and stability effects within the protein.

Most methods, both general and cancer specific, classify pathogenicity based on dramatic effects on structure and stability and proximity to conserved/functional sites. Therefore they can sometimes miss disease mutations which are more tolerated i.e because they have little or subtle effects on structure or lie far from known conserved/functional sites[158]. Furthermore, there is a need for additional structurally derived information- such as dynamic effects and more comprehensive functional annotation.

Molecular Dynamic Simulations (MDS)

The use of MDS enables the modelling of the dynamic motions of proteins - providing a quantitative spatio-temporal means of studying protein behaviour, by sampling the conformational landscape which is difficult by other means [58]. Thus, MDS proves extremely beneficial for gaining insight into the stability, folding and behaviour of the protein on a dynamic level. Kumar *et al* [118] studied the effects of the G325W mutant in Aurora-A kinase implicated in hepatocellular carcinoma. The mutant caused a decreased kinase stability and binding affinity to the kinase substrate, measured using free energy calculations. The mutant also led to greater flexibility in the kinase shown by RMSF (root mean

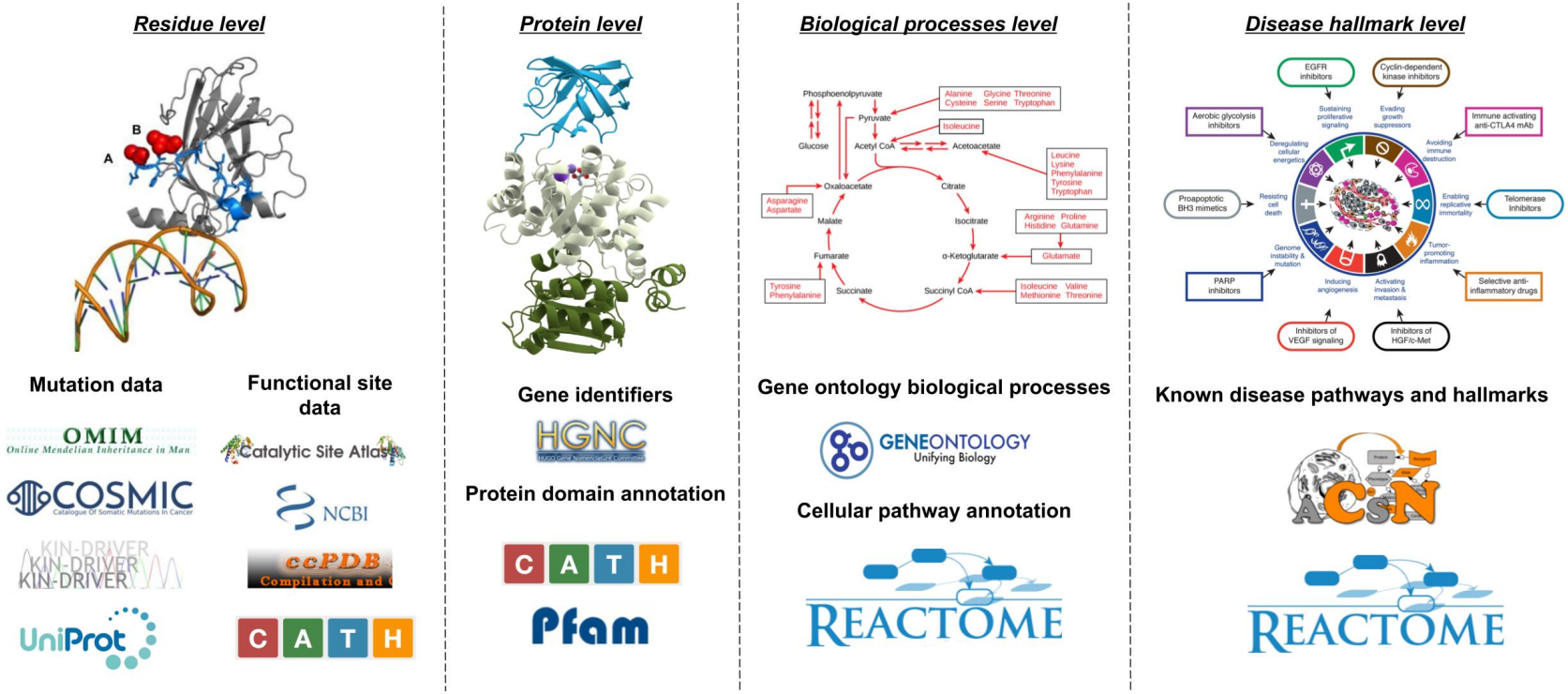
squared fluctuation) calculations.

MDS studies have also been performed to study drug efficacy profiles of the FGFR1 and FGFR3 receptors, which are often mutated in cancer [23]. This provided a structural rationale for why some cancer-associated FGFR mutants become resistant to certain drugs. MDS has also been shown to increase the accuracy of SNP prediction, by incorporating the global effects of mutations such as on the stability, flexibility and solvent accessible surface over time [194].

Other resources used in SNP impact analysis

There are many different types of functional annotations used in mutation analysis, which span the scales from the protein residue level up to biological processes and disease hallmarks, as can be seen from figure 1.3). For each level of biological abstraction used in this thesis, the data resources are described below.

Figure 1.3: Summary of the resources used in disease mutation analysis in this thesis - from the residue level to biological pathways and clinical hallmarks



Resources For Mutation Data: Germ-line Diseases And Cancer

Online database of Mendelian Inheritance in Man (OMIM)

OMIM(Online database of Mendelian Inheritance in Man) is an online resource that catalogues the germline amino acid variants and their associated diseases[10]. Mutation entries are from a range of genomic regions including those from the 22 non-sex chromosomes (autosomes), from the X and Y sex chromosomes (allosomes), and from mitochondrial DNA. The data within OMIM gives the genotype and phenotype relationships of inherited diseases, which range from monogenic rare diseases to more complex cases such as cancer, where the latest version contains 24,378 entries as from December 2017. In terms of genetic pleiotropy, the majority of genes within OMIM (68.5%) are associated with one phenotype. The phenotypes themselves are classified into four different groups which are; single gene disorders, susceptibility to complex diseases, somatic cell disease, and non disease.

UniProt HUMSAVAR

Human disease mutations are also provided by UniProt HUMSAVAR. This resource contains human variants associated with a range of germline and somatic diseases, including cancer and non cancer. The HUMSAVAR variants are all the missense mutations annotated within human UniProtKB entries, curated as being associated with diseases according to literature reports. [11]. The latest update from July 2015 contained a total of 71,795 entries, which were from 3 different associated phenotypes of ”disease associated”, neutral or ”polymorphism” variants, and ”unclassified”. Many studies use this dataset as a reference when studying the general effects of disease causing mutations on protein function [76] [42] [43].

Cancer mutations: COSMIC

For cancer mutations, the main repository is the catalogue of somatic mutations in cancer (COSMIC) database, which currently serves as the worlds largest and comprehensive resource for cancer mutations, specifically containing 4 million coding mutations across

all cancer types, along with 10 million non-coding mutations [71]. There are 2 main sections within COSMIC, the first of which encompasses large scale data from genome sequencing studies, which has been taken from other cancer mutation resources such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium ICGC. Since cancer is a heterogeneous disease with many genomic errors contributing to its pathogenicity, it is hard to infer which genes are most important in driving the cancer. Therefore there is a second section within COSMIC which is cancer mutation data that has been curated by experts covering over 600 key genes known to be implicated in cancer, called the cancer genome census (CGC) [71]. The CGC also contains additional information on the likelihood of the gene to be either a tumour suppressor gene (TSG) or an oncogene (ONC).

PinSnps: OMIM, COSMIC, dbSNP

PinSnps is an interactive web based tool which enables the analysis of mutations in different types of diseases such as different cancer types from COSMIC [15], germline diseases from OMIM [10] , and common variants from dbSNP [213]. In total there are 1,101,990 cancer mutations, 10,850 germline disease mutations, and 3,209,256 common SNPs. Proteins are annotated on their Pfam domain content [222], which are in turn linked to their available 3D structures or a homologous structure. Information on functional sites from UniProt are also available [11], along with predicted disordered inter-domain regions by DISOPRED [264]. This data was used by the same authors in analysing the enrichment of the disease mutations in such functional sites and disordered regions[135], and a more detailed discussion is provided in chapter 2. In addition to considering annotations within the protein structure, PinSnps also maps mutated proteins to human protein-protein interaction networks (PPINs) [136], enabling the analysis of mutations at both an atomic resolution and within a system level context, where the effects of mutations on binding protein partners can be investigated.

Resources for functional site annotations

A range of resources report functionally important sites in protein structures, derived from both the literature or predicted using sequence or structure based approaches.

Catalytic sites

The Catalytic Site Atlas (CSA) provides residue annotations for enzymes that have structures within the protein data bank (PDB)[192]. There are two types of catalytic sites within the CSA, the first is based on manual curation from the literature. The second set are catalytic residues identified in homologues of proteins in the first set and inferred from a PSI-BLAST multiple sequence alignment against the original entries. The latest version of CSA, CSA 2.0, contains 968 curated and 584 homologous entries with catalytic sites [72]. CSA also incorporates information from other databases such as MACiE [97] that provide information on the reaction mechanism for an enzyme, and the roles of the residues in catalysis.

Protein-protein interaction sites

NCBI-IBIS is a resource [216] which analyses and predicts protein binding partners and their interfaces based on known structural complexes from within the NCBI molecular modelling database [138]. The different types of protein interfaces within IBIS range from; protein-protein, protein-DNA, protein-RNA, protein-ion, protein-small molecule, and protein-peptide. In the latest release in January 2017, the number of chains which have these observed interactions are shown in table 1.1. In addition to providing observed entries, IBIS also predicts protein interfaces based on homologous complexes, where similar sites are clustered and inferred based on evolutionary conservation.

Ligand binding sites

Another resource for ligand binding sites is the ccPDB database [218], which uses prediction tools to identify ligand binding motifs for ligands which have binding sites within at least 30 proteins in the PDB. From the sets of structures for a given ligand, the ccPDB

Table 1.1: The number of observed chains containing the different types of protein interaction within latest release of IBIS, January 2017.

Type of interaction	No of protein domains/chains with observed interactions
Protein-DNA	6700
Protein-RNA	19833
Protein-Protein	220808
Protein-Chemical	105184
Protein-Peptide	7649
Protein-Ion	58906

analysis module deciphers the residue preferences for this ligand and incorporates this into a propensity-based predictor for ligand binding residues. The ccPDB in turn uses this predictor to annotate PDB structures.

UniProt functional features

The UniProtKB [11] resource contains a range of sequence based functional features, including residues involved in: DNA and RNA binding, active site chemistry, part of a protein domain, metal binding, and sites for various post-translational modifications such as polysaccharide addition, and methylation. These range in size from specific residues involved in functions such as post-translational modifications and catalysis, to whole protein domains.

Predicted allosteric residues

Recent efforts have used protein structure data to identify putative sites of allosteric importance. These include the works of Clarke *et al* [35], who developed the predictor STRESS which identifies both surface based and interior based allosteric sites. The STRESS methodology uses protein structure properties such as residue connectivity, and correlated movements upon ligand binding to infer allosteric function.

Resources for protein domain annotations

The use of domain annotation can enable a finer approach to assessing the effects of mutations on proteins and how this leads to disease. This is due to the increased coverage of protein sequences by domain family annotations as compared to whole protein family annotations and the fact that they tend to have conserved structures as inherited units of evolution. Pfam is a sequence based domain family resource which has been used by some mutation analysis tools [222]. However in this project the domain structure resource CATH will be used as it provides more structural information.

Sequence based domain annotations in Pfam

Pfam is a widely known resource for sequence based domain annotation[222]. There are two multiple sequence alignments used to describe a Pfam superfamily, the first of which is called the seed alignment and is composed of a smaller subet of protein sequences which are representive of the family. The second sequence alignment is more extensive since it includes sequences from all family members taken from UniProtKB. The Pfam protein families are either made by manual curation, or generated automatically and are called Pfam-A and Pfam-B families respectively. Both Pfam family types can be profiled using a Hidden Markov Model (HMM), derived from a multiple sequence alignment of relatives within the family. The latest version of Pfam version 31.0 [69] contains 16,712 protein families, and at least 73% of UniProt protein sequences match to at least one Pfam family. For a given protein sequence, domain annotations are inherited if the query falls within the boundaries of a Hidden Markov Model for a protein superfamily.

Structure and sequence based domain annotations in CATH

CATH is a hierarchical database which classifies protein structures from the Protein Data bank (PDB) according to their Class, Architecture, Topology and Homology [172]. Multi domain proteins from the PDB are split into distinct domains in CATH. Structural domains are assigned to superfamilies using structure comparison methods (SSAP, CATHEDRAL) and sequence based HMMs to confirm homology. Distant homologues

are validated by manual curation. Currently there are 300,000 domains in CATH, classified into 2,700 superfamilies [217].

To assign new sequence domains to a superfamily in CATH, their sequences are scanned against a sequence pattern or Hidden Markov Model (HMM) derived from a multiple sequence alignment of relatives in the superfamily. Superfamilies are subdivided into more specific functional families, by means of an hierarchical agglomerative clustering method, GeMMA [128] and the more recent FunFHMMer [40]. FunFHMMer is a method that separates relatives into more functionally and structurally coherent protein domain families, called FunFams, based on specificity determining positions identified by the GroupSim method [25]. CATH uses Scorecons to calculate conserved sites within a FunFam, likely to be functionally important sites [253]. Scorecons calculates conservation at a specific position based on Shannon's entropy measure of diversity. Conserved sites within a FunFam, encapsulate residues conserved within a superfamily, and in addition conserved residues specific to that FunFam, referred to as specificity-determining-positions (SDPs). Since CATH FunFams are functionally coherent, these SDPs are likely to be involved in dictating different functional specificities.

Resources for pathway and biological processes annotation

Gene ontology (GO) annotations

The Gene Ontology (GO) resource provides a unified representation of gene product attributes, using a controlled vocabulary to annotate protein function in terms of cellular compartment, molecular function, and biological process [13]. The annotations are from a range of species, from eukaryotes, prokaryotes, single and multi-cellular organisms. The terms are structured in a directed acyclic graph, where each term can have one or more defined relationships with another term within the same domain or functional grouping. For a given gene list, enrichment studies can be performed for each of the GO attributes, in order to gain an insight into its biological relevance.

Cellular pathway annotations from Reactome

There are a number of resources which enable insights into cellular pathways. These include resources such as Reactome [63], which contains pathways from a range of resources such as KEGG [241], NCBI [80], and literature derived reactions. The reactions are grouped into networks where they together represent pathways. There are resources which include GO and reactome pathway analyses to give information on enrichment of genes within biological contexts as described above - called ReactomeFVIZ, which is described more in chapter 3 [269].

The Atlas of Cancer Signalling Networks (ACSN)

The Atlas of Cancer Signalling Networks (ACSN) is a resource which provides a list of expert curated cancer hallmarks [120], which are known to be implicated in cancer progression. The ACSN is an interactive web based environment, displaying 4600 biological mechanisms covering 564 proteins involved in cancer. These are grouped in 5 major cellular processes which are effected in cancer, and are therefore considered hallmarks: cell survival, apoptosis, EMT cell motility, cell cycle, and DNA repair. The enrichment of gene lists within these hallmark processes provides an insight into their possible molecular mechanisms and contribution to the carcinogenic phenotype.

Thesis Summary

A novel method was developed to identify cancer driver mutations, and cancer driver genes within CATH FunFam domains. Cancer driver genes were identified using the MutFam protocol, which identifies significantly mutated CATH FunFam domains within 22 cancers reported in the COSMIC database [15]. The MutFam method succesfully identified domain families highly implicated in cancer, including those containing P53 and PTEN. Cancer driver mutations were then identified in 3D clusters within significantly mutated domains in the 3D protein structure. This approach filtered out mutational noise and passenger mutations. MutClusters were analysed on their proximity to both known and in-house predicted functional sites, and were compared to unfiltered cancer

mutations, mixed disease, and germline disease mutation datasets. The putative cancer driver mutations occur significantly closer to catalytic residues, protein-protein interaction interfaces, ligand binding sites, and in-house derived CATH FunSites compared to the non-clustered disease mutations.

A more detailed study was undertaken of the FGFR3 receptor implicated in bladder cancer. A number of analyses were performed including impacts on structure and proximity to functional sites, including protein-protein interfaces and catalytic sites. The effects on stability and folding rate were also examined. For a subset of mutations, the results were compared against experimental characterisation of mutation impacts on activation of the kinase. In order to characterise the functional impacts and clinical relevance of the predicted MutFam driver genes, further analyses were performed using enrichment studies based on Gene Ontology annotations, and known cancer hallmarks respectively. It was found that the MutFam driver genes were enriched in known cancer hallmarks affecting cell survival and cell motility. Gene Ontology enrichments were used to compare the MutFam driver genes within early and late stage gliomas, identifying common and distinct functional effected processes, which reflected the respective clinical phenotype of the gliomas. Whilst this work has mainly concentrated on the analysis of cancer mutations, the methods developed are generic and could be applied to analysing other types of disease mutations.

Chapter 2

Exploring The Proximity of Disease and Predicted Driver Mutations To Functional Sites

Introduction

In this chapter the proximity of putative driver mutations to functional sites was analysed. Below is a literature review of this field, followed by a summary of the work performed in this chapter.

As introduced in chapter 1, various approaches have been used to study the likely structural and functional effects of disease causing residue mutations (non synonymous single nucleotide polymorphisms, nsSNPs). These include analysing 1) the proximity of mutations to known and predicted functional sites 2) the detection of mutation hotspots in either the sequence or the structure and 3) the use of protein domain annotations. Each of these features will be discussed in more detail below.

The aims of the work described in this chapter were to establish new approaches for determining the impacts of disease associated mutations on protein structure and function. The different types of disease causing mutations considered included those in germ-line diseases (i.e cancer and non-cancer), and somatic cancer mutations in oncogenes and tumour-suppressors. The proximity of disease-associated mutations in proteins to known functional sites reported by CSA catalytic residues, IBIS protein-protein interfaces, UniProt functional features, and in house predicted functional sites were investigated. The latter are called FunSites and are highly conserved residues in CATH functional families (FunFam). FunFams are functionally coherent subsets of relatives in CATH superfamilies. Such sites have been shown to be enriched in known functional sites, e.g. catalytic sites and specificity determining residues [41].

In addition to analysing single mutations within PDB structures, the project also investigated whether disease mutations which clustered in the protein i.e. 3D hotspots showed tendencies to lie close to the functional sites discussed above. The use of

hotspots in mutation analysis has been shown to be important in deciphering pathogenic mutations, especially for cancer driver mutations. It provides a means of filtering out noise or neutral mutations and highlighting protein positions under positive selection and therefore more likely to be of functional importance in cancer. In this chapter, a proximity analysis was performed for a range of disease causing mutations, and in-house predicted driver mutations to known and predicted functional sites.

Tendency of nsSNPs to be on or near to functional sites

Some cancer mutations are passenger mutations, which do not drive the oncogenic phenotype, whilst others are likely to be drivers which engender the cancerous phenotype. Many of the cancer mutation data resources are polluted with passenger mutations and so efforts have been made to identify mutations which are likely to be drivers. In 2009, kinase specific studies performed by Izarzugaza *et al* [105] used a statistical measurement to report the significance of germ-line disease mutations from OMIM [10]) to be close to known kinase functional regions. These included conserved residues derived from the alignments of relatives in 8 kinase families in KinBase [121] , where conservation was measured using S3Det. Buried sites were those with a relative solvent accessibility $<$16%, and catalytic sites were taken from CSA [192] and FireDB [139]. Results showed that germline cancer mutations in OMIM have higher tendencies to be close to or co-located with catalytic residues and all other defined regions, compared to neutral mutations in dbSNP [213].

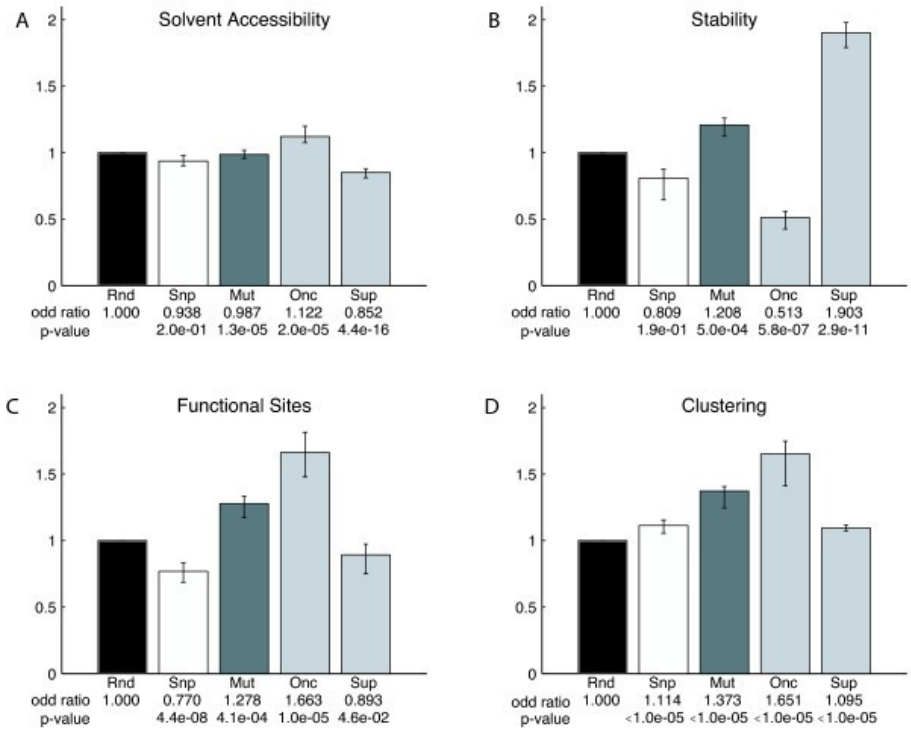
A related study by Talavera *et al* [239] analysed cancer mutations in genes having a broad range of functions, including kinases. Driver mutation sites were identified as having a dN/dS ratio $>$1, which is a proxy for positively selected residues. Functional sites were 9 residue windows containing residues involved in ligand, metal and nucleic acid binding and protein-protein interfaces, obtained from the WSsas webserver [237]. Conserved sites were identified by AL2CO [179] and buried sites were measured by ICM [2]. The authors found that cancer mutations, especially the driver mutations, were enriched at functional sites, the highest proportion being in protein-protein interaction

sites, ligand binding sites, metal and nucleic acid binding residues. They also found that genes involved in cell adhesion, multicellular development and DNA-binding functions were over-represented in the driver mutation sets.

A later study of Stehr *et al* [226] analysed the tendency of cancer mutations in annotated oncogenes (ONC) and tumour suppressor genes (TSG) from COSMIC, to lie within a distance threshold of 8 Å to a functional site. The proximity of C-β atoms of mutations was measured to catalytic, ATP/GTP and post-translational modification sites within protein domains, defined by Domain Parser [270]. Neutral mutations (snp) from dbSNP were also included [213], and 1000 random mutations were derived from the amino acid sequence (rnd) as an extra control. Significance was assessed based on the difference in the distributions of the disease mutations and random or neutral mutations. FOLDX [208] analysis was also performed to study the effects of the mutations on protein stability, and NACCESS analysis was performed to measure buried residues having a solvent accessible surface of $>$15%.

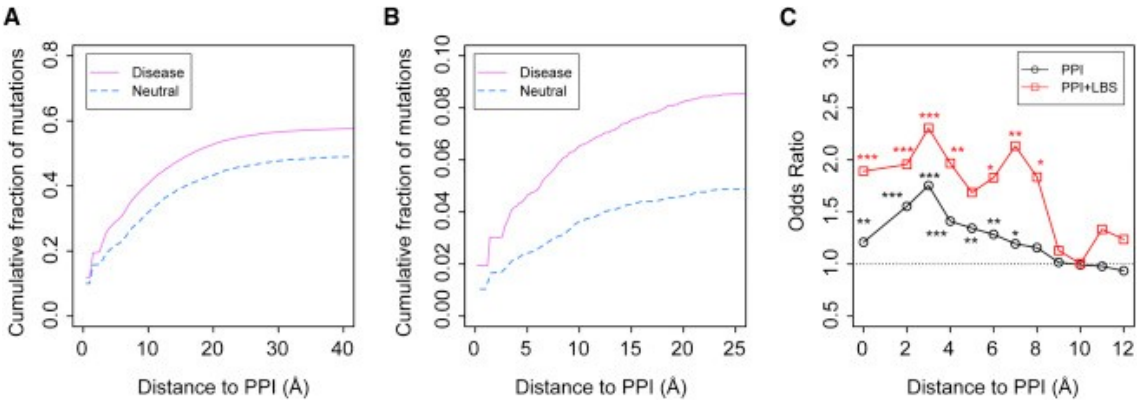
The results showed that oncogenic mutations were over-represented at $\leq$ 8 Å from functional sites, specifically GTP/ATP binding sites and also on the surface of the protein. In contrast TSG mutations were shown to be under-represented at distances $\leq$ 8 Å from functional sites and were more likely to destabilise the protein and occur in buried sites. These results are shown in figure 2.1. These results therefore also highlighted the importance of analysing regions proximal to functional sites which might have a role in altering function.

Figure 2.1: Summary of the structural and functional effects of cancer mutations, taken from [226]. The groups of mutations analysed here are; Rnd = random, Onc =mutations in oncogenes, Sup = mutations in tumour suppresser genes, Mut = all cancer mutations, Snp = neural mutations taken from dbSNP.



In contrast to Stehr *et al* [226], other studies do not constrain analysis of mutation proximity to within 8Å. These include the works by Gao and co-workers [75] who analysed the proximity of disease mutations from cancer and non cancer germline and somatic mutations (described as mixed mutations here), and neutral mutations catalogued in UniProt and PDB structures, to predicted functional protein protein interaction sites (PPI) (see figure 2.2). Residues involved in protein-protein interaction sites (PPI) were calculated using distances $<$4.5 Å between complexed partners. The closest atomic distances between the mutated residue and functional site residues were determined. This showed that UniProt mixed disease mutations have higher tendencies to occur close to PPI sites than neutral mutations. This enrichment was particularly prominent for proximity to residues 3-6 Å away from the PPI site, as opposed to the PPI sites themselves, which gave a significant odds ratio of disease to neutral mutations of 1.75 and 1.23 respectively.

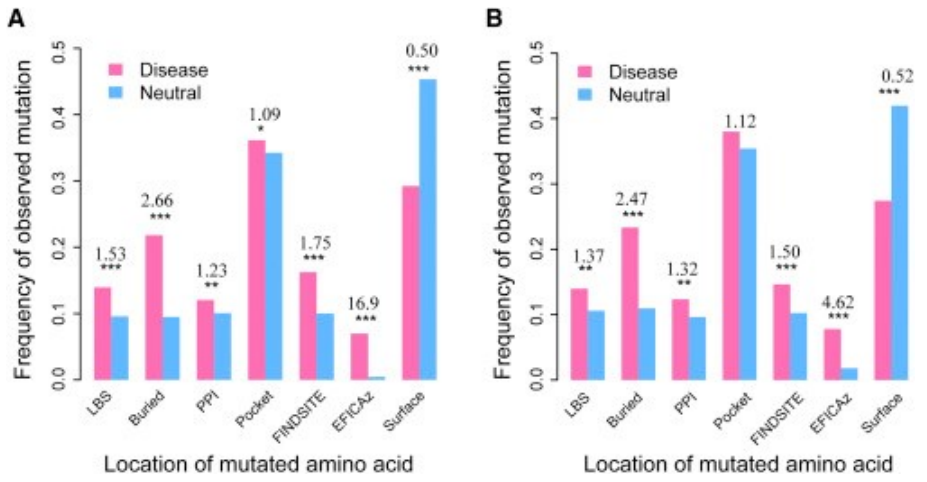
Figure 2.2: Analysing the proximity of UniProt mutations to in-house derived PPI sites, taken from [75]. The Cumulative density plots in A) are at distances from 0-40 Å B) are at distances from 0-25 Å . C) odds ratio plot of the enrichment of mutations at distances from 0-12 Å).



Mutation co-location analyses were also performed on other types of sites, including functionally discriminating residues in enzyme families, predicted using EFICAz [119] and predicted ligand binding sites using FINDSITE[21]. Known functional sites included residues with a PDB header containing a “HETATOM” type as either ligand, metal or ion binding found by the LPC (Ligand Protein Contacts) method [220]. The degree of en-

richment at these sites was determined by calculating an odds ratio of disease to neutral mutations, and the results of these are shown in figure 2.3. Disease mutations were found to be statistically enriched in or close to all predicted functional sites, particularly EFFICAz sites with an odds ratio of 16.9 compared to neutral mutations at this position.

Figure 2.3: Co-location of UniProt mixed disease-associated mutations versus neutral ones to different functional regions. (A) All mutations. (B) Subset of mutations from the same set of proteins that contain both disease-associated and neutral mutations. The y axis represents the fraction of mutations in either disease-associated or neutral data sets. The numbers above pairs of bars are the odds ratios. The different types of functional sites are; LBS = ligand binding site derived from PDB, Buried = buried site, PPI = protein-protein binding site derived from PDB, Pocket = residues within predicted protein pockets, FINDSITE = predicted ligand binding sites, EFICAz= predicted functional determinants, Surface = surface sites. Taken from [75].



Mutations within germline and somatic cancer diseases have been shown to differ in their selection pressures, as they appear to be associated with different structural and functional consequences [83] [90]. As the volume and quality of mutation data increased over time, separate studies were conducted on these different types of disease causing mutations. Such studies include the work by Yang *et al* [175] who performed an extensive analysis of somatic and germline disease mutations within the COSMIC/TCGA/ICGC/IntOGen and dbSNP resources, respectively. The authors analysed

the co-location of mutations to UniProt and CDD functional features [11] [142] including active sites, chemical binding sites and post-translational modification sites. Other PTM sites were also included from dbPTM[131] and conserved sites were derived from a BLAST alignment of proteomes from human and 5 other species. Amino acid conservation was also measured from the percentage of homologues in which the site was conserved. Significance of mutation co-location was measured using a binomial statistic. The results of this analysis showed that somatic cancer variants were generally closer to functional and conserved sites compared to the germline variants. The most significant over-representation being in UniProt binding and active sites with a P-value of 4.81E-32 and 1.83E-4 respectively. Interestingly, phosphorylation sites were shown to be under-represented for both germline and somatic variants. This study also analysed PanCancer mutations, and again revealed that different cancers exhibit discrete impacts on different functional sites.

Similar studies of germline co-location have been performed by Martinez *et al* [108], and included allosteric sites from the allosteric site database ASD2.0 [99]. The disease mutation data was taken from the Human Gene Mutation Database(HGMD) [227], and neutral mutations were taken from dbSNP [213]. The functional sites included were: N-linked glycosylation sites and different metal binding sites and residues 3 Å from these sites. Catalytic sites were taken from CSA [192]. Phosphorylation, protein-protein interaction, and DNA-binding sites were from previous studies, allosteric sites from ASD2.0 [99] and RNA and ligand sites from the [218].

Proximity of disease mutations to protein interfaces

There have been many studies analysing the frequency of mutations in interfaces of protein complexes linked to diseases, including cancer, where mutations in such functional sites can manipulate cellular signalling and perturb binding of native binding partners leading to abnormal signalling and disease [169] [37]. For example, analysis of mixed disease mutation data from UniProt – including germline non cancer disease and somatic cancer mutations – showed that protein-protein interaction sites are hotspots for

disease mutations using a UniProt neutral background as a control [42]. The interface sites were composed of residues <5 Å to residues on the binding partner in PDB structures. Later work, by the same authors [43], extended this analysis by splitting the interaction sites into core and rim segments, based on being fully or partially buried upon complex formation respectively. Mutations were analysed on their co-location to these interface regions, showing that disease mutations were 35% more likely to occur at an interface compared to the rest of the surface and are 49% more likely to occur in the core of the PPI sites compared to the rim, where they coincide with more conserved residues according to BLOSUM62 and residues important for the binding affinity of the complex, measured by FoldX [208]. Subsequent studies have further confirmed that protein-protein interaction sites are hotspots for disease causing mutations [274][24], showing particular relevance in cancer [169] [37].

Other work analysing the effects of cancer mutations on protein-protein interaction sites, includes that of Epinosa and co-workers [62], who showed that known cancer driver mutations within COSMIC putative oncogenes and tumour-suppressor genes are enriched in the partially exposed and buried residues within PISA derived protein-protein interaction (PPI) sites. Solvent accessibility was measured using PISA absolute accessible and buried surface areas as a ratio of the respective amino acids in the G-X-G peptide. The authors found that driver mutations were more likely to co-locate to PPI sites than neutral mutations, especially at partially exposed residues, where they disrupt electrostatic interactions across interfaces by replacing amino acids involved in hydrogen bonds. This information was used to develop a mutation pathogenicity score called Inca (Index of carcinogenicity).

Much like the work of Epinosa *et al* [62], more recent studies by Egnin *et al* [61] also performed independent co-location analyses of oncogenes and tumour suppressor genes from COSMIC, to PDB derived protein oligomerisation sites, and other sites of varying solvent accessibility measured by ASA [251]. In addition to this, Egnin *et al* also measured the effects of mutation on protein stability using FoldX [208], and the predicted functional impact measured by VEST [60]. An important point to note is that this

functional impact score is trained on Mendelian disease genes from the Human Gene Mutation Database (HGMD) [227], which are known to differ in their functional impacts to cancer mutations [214] [83]. Therefore predicted functional mutations from VEST may be biased towards more Mendelian like features, often seen in tumour-suppressor gene mutations, but not seen so much for mutations in oncogenes [226] [247].

In agreement with Epinosa *et al* [62], both oncogenes and tumour-suppressor genes showed an enrichment of mutations within interfaces compared to other surface residues, showing on odds ratio of 1.17 and 1.28 respectively. This analysis by Egnin *et al* [61] further confirmed the tendency of tumour-suppressor mutations to be more prevalent in the protein core than oncogene mutations, where they lead to protein destabilisation. This result is consistent with previous studies [16]. Egnin *et al* also showed that tumour suppressor genes harboured more mutations at known homo-oligomerisation sites compared to oncogenes.

Other recent studies include the work of Porta-Pardo *et al* [189] [188] who developed a method called e-driver, which detects driver genes based on them containing protein functional regions which have a cancer mutational bias compared to the rest of the protein. PPI interfaces were residues 5 Å distance from residues within a separate chain in a multi-chain PDB structure. This approach was performed on cancer mutations from TCGA [249]. e-driver identified significantly mutated protein regions by comparing the observed number of mutations in a specific region to an expected distribution in a given protein structure, normalised by the length of the given region. Statistical significance of mutation enrichment was assessed using a right sided binomial test, and genes showing significant bias were considered to be candidates for driving cancer. Again, this analysis highlighted the frequent targeting of protein interaction interfaces amongst driver gene mutations in cancer (e.g TP53,EGFR and HRAS). This study also highlighted different mutated regions within PPI sites as being implicated in different cancer types, consistent with more recent studies [61] [262] [271].

The results showed that the germ-line disease variants have a greater enrichment of mutations for all functional sites compared to neutral variants, particularly at ligand bind-

ing sites, with a relative proportion of 3.07 and 0.87 for the disease and neutral mutations respectively. Other enriched sites included metal binding sites, with relative proportions of 2.88 and 0.10 for the disease and neutral mutations respectively. Sites with the lowest proportions of both disease and neutral mutations were the post translational modifications and allosteric sites.

Proximity of disease mutations to allosteric sites

More recently, analyses of proximity to allosteric sites have been performed. The precise functions and locations of these sites are harder to infer, and in the past have solely relied on experimental validation and manual curation. Allostery plays an important role in modulating protein functions, distinct to the functional site itself.

Kinase specific studies, performed by Dixit *et al* [54] used predicted allosteric sites, and analysed their co-location to somatic cancer mutations in kinases. These authors demonstrated that residues at certain sites within the kinase have an inherent instability which makes them more likely to experience conformational transitions, which would result in slight structural changes. These sites were also described as “frustrated sites”, and were found to overlap with allosteric residues which play a role in initiating functional allosteric transitions. This study showed that cancer mutations from COSMIC [15] altered the position of frustrated sites, affecting the activation equilibrium of ABL and EFGR kinases implicated in oncogenic signalling, thereby favouring the active forms.

More recently, Kumar *et al* [234] performed an extended study of frustrated sites in many protein types, in the analysis of cancer driver mutations from COSMIC and TCGA and cancer genome census (CGC) which were separated into oncogenes (ONC) and tumour-suppressor genes (TSG) mutations. Pathogenic germline mutations were taken from HGMD [227]. For each of the mutation data sets, Kumar *et al* explored their co-location to predicted unstable regions within the protein, termed frustrated sites, as in the Dixit study. Frustrated sites within protein structures were defined as interacting amino acid pairs, which when mutated to all possible amino acids, possess a relative instability compared to the whole protein in the wild type form, and were measured using

the Frustratometer tool [176]. Solvent accessibility was also measured using NACCESS, where core residues had a relative solvent accessibility of $<$25%. Results showed that oncogenic cancer mutations are more likely to co-locate to frustrated residues on the protein surface, compared to pathogenic germline and cancer TSG mutations, which lead to greater disruptions to frustrated sites within the protein core. This result is also consistent with other studies by Stehr *et al* [226] and Egnin *et al* [61].

Another related study analysed the co-location of germline disease variants from ClinVar [209] and HGMD [227], and benign variants from ExAC [221] to predicted allosteric sites. Unlike the Dixit study, this study [35] predicted 2 types of allosteric site which are STRESS-surface and STRESS-interior. STRESS-surface sites were residues which exhibited significant conformational changes upon ligand binding, and were modelled using Monte Carlo simulations. STRESS-interior residues are residues with a high betweenness centrality within a protein residue network, which was calculated using normal mode analysis, and were hypothesised to act as communication hubs and therefore sites of allosteric action within the protein. This analysis showed that both of the STRESS-sites harbour statistically more disease mutations from HGMD, ExAC, and ClinVar, compared to non-STRESS residues.

Using enrichment studies to identify disease driver mutations and assess co-location to functional sites

Information on protein regions, residue hotspots or mutation clusters

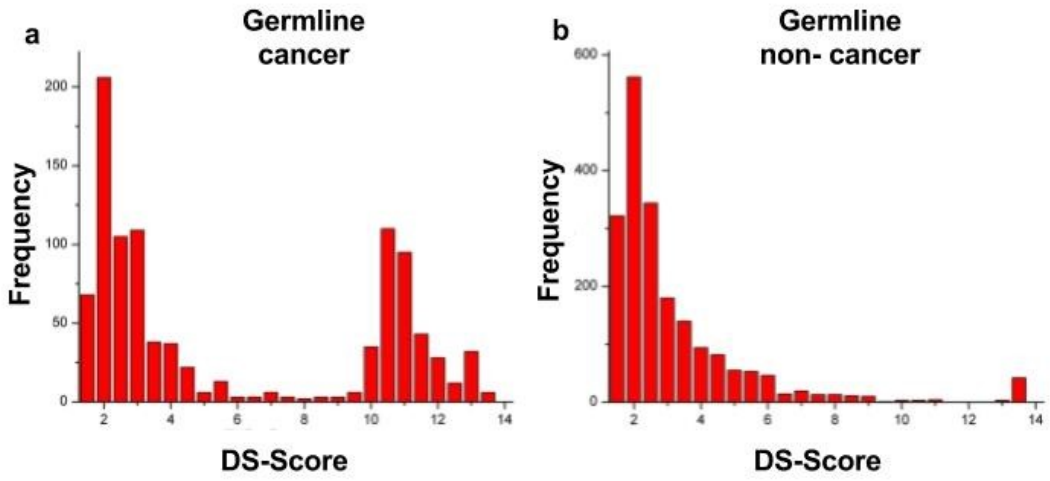
The enrichment of mutations at particular sites within linear protein sequences and individual positions (described as 1D hotspots) and structures (described as 3D hotspots) has been widely regarded as an indicator of positive selection, and is of particular relevance in diseases such as cancer. In order to aid functional validation of these predicted driver mutations, various studies identifying such 1D and 3D hotspots have analysed their co-location to various functional sites to infer effects on function. Within the last 10 years, there have been a plethora of sequence based 1D hotspot detection methods, which are based on identifying a statistically significant enrichment of mutations at protein positions [126] and regions within protein sequences [152], where such regions can be defined using domain annotations or adjacent residues surrounding the mutation [240]. The use of a domain centric approach in hotspot identification can help refine driver mutation analyses, as domains from a particular domain family can occur within many genes [273] so the accumulation of mutation data for domains within a single family can increase statistical power in predicting the functional impacts of mutations. With regards to identifying 1D hotspots, many studies have exploited the use of domain based annotations in hotspot analysis, and these will be discussed below.

Peterson *et al* explored the clustering tendencies of non-cancer and cancer mutations from OMIM and SwissProt in a variety of protein domain annotations [182]. The domain annotations used were taken from CDD (Conserved Domain Database), Pfam, COG and SMART. A domain-significance score (DS-score) was calculated for each position within the domain to measure the mutational enrichment. This score is based on a method developed in earlier work performed by Yue and Forrest *et al* [276], and was used to determine the probability of observing a mutation cluster of a particular size at a single residue, given the number of mutations in total and their positions in a domain. Positions with high DS-scores were referred to as disease hotspots and were further

analysed according to their co-location to CDD functional features and conserved sites measured by AL2CO. This work identified disease hotspots in both germline cancer and germline non-cancer diseases, shown in figure 2.4, where 58.1% of the cancer and 51.2% of the germline non cancer hotspots occurred in highly conserved positions. In addition to this, the authors found that a greater proportion of germline cancer mutation hotspots (69.8%) were co-located with functional features compared to germline non cancer disease hotspots (35.9%).

Interestingly according to figure 2.4, the germline cancer mutations showed 2 groups of mutations possessing different DS-scores, the first associated with a peak at lower scores, similar to the germline non-cancer mutations. The second peak was primarily composed of putative oncogenes, which more frequently gave higher DS-scores, suggesting high mutational enrichment. The results highlighted again the importance of considering germline non cancer and cancer mutations separately, since their clustering patterns differ.

Figure 2.4: The DS-score distribution for a) germline cancer and b) germline non cancer mutations. Taken from [182]).



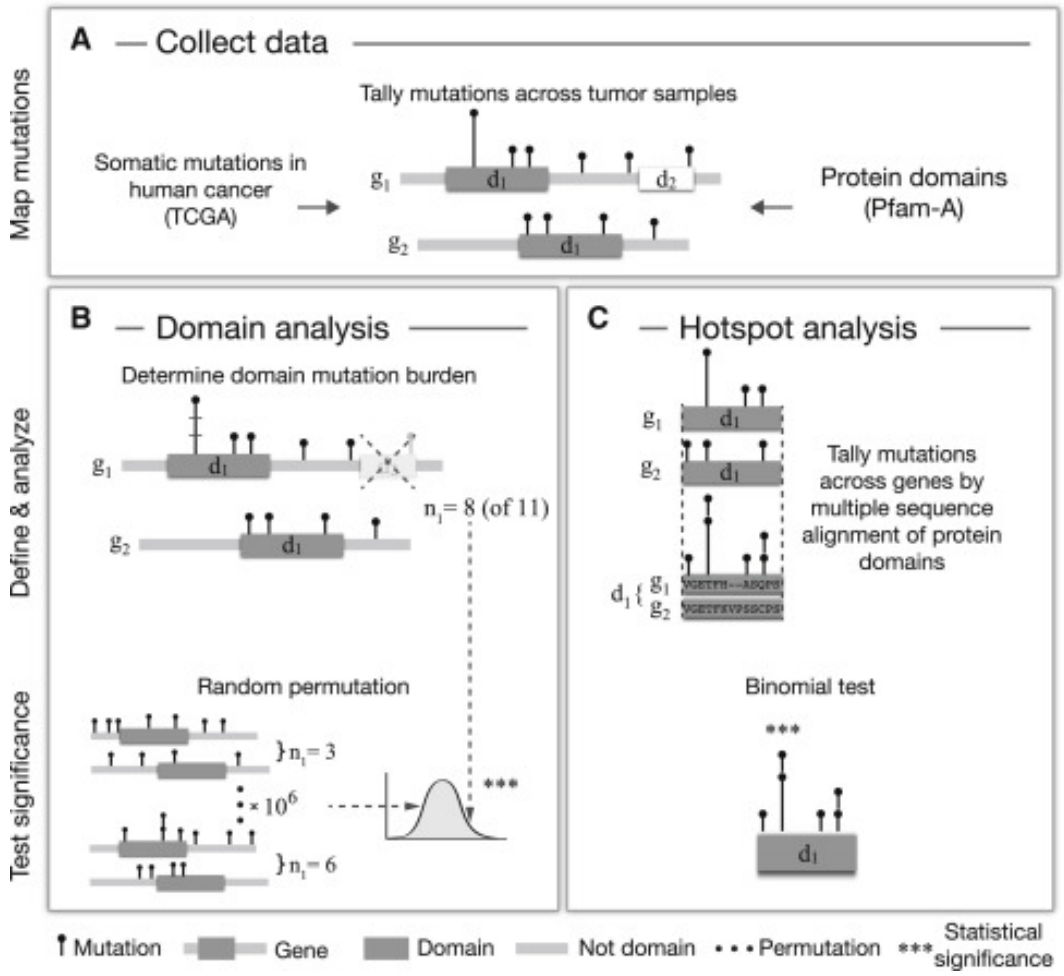
More recent work by Miller *et al* [154] explored both the domain and position based enrichment of cancer mutations in Pfam domains [222]. Miller *et al* analysed somatic cancer mutations obtained from 22 cancer types from the TCGA [249]. The domain enrichment was referred to as domain burden and was quantified by an enrichment score, shown in equation 1. The expected mutation burden in this study was based on the total number of mutations observed and the fraction of amino acids assigned as domains compared to the total number of residues of all genes in a domain family.

Equation 1:

$$ed(domainenrichmentscore) = md(observedmutationburden)/me(expectedmutationburden)$$

The enrichment score in equation 1 was made by tallying mutations across multiple domain relatives identified within specific domain boundaries. These were compared to a random mutation permutation test outside the domain, 10^6 times across each gene. The P-values were calculated as the proportion of permutations that contained a larger mutation count compared to the observed mutation counts in a specific domain. This work revealed 14 significantly enriched domains involved in canonical signalling, in particular kinase domains, which were enriched in cancer causing mutations. As well as considering whole domain enrichments, this study also calculated a within-domain 1D hotspot based on whether certain positions possessed significantly higher mutation counts in a multiple sequence alignment of domain containing genes, compared to other residues within the domain. The process of both domain and positional enrichment, applied in this study, is illustrated in Figure 2.5.

Figure 2.5: Schematic of the method for determining domain and positional enrichment of mutations within Pfam domains, taken from [154].



This work identified 82 hotspots encompassing well known mutations, along with 68 novel mutations which would not have been detected by individual gene analysis. These mutations were within known cancer-associated genes including receptor tyrosine kinases such as; EGFR, FGFR1 and EPH. The results of this analysis are available from an online webserver, MutationAligner [79]. For selected examples, the authors also analysed the co-location of specific hotspots to known conserved regions. An example of this is a hotspot within the Forkhead domain, which was located in the 3rd helix of the conserved wing structure, involved in DNA binding.

Similar to the works of Miller *et al*, Fan Yang *et al* [271] used Pfam domains in the analysis of pathogenic mutations predicted by the IntOGen platform [86], taken from 21

cancer types from COSMIC [15], ICGC [281], TCGA [249]. Mutations with high, medium and low impact predictions were used in detecting significantly enriched Pfam domains and position based 1D hotspots within each cancer type. In contrast to Miller's work, Yang *et al* didn't aggregate mutations across a Pfam domain family, but identified 1D hotspots as significantly mutated positions within genes containing a significantly enriched Pfam domain, where oncogenes (ONC) and tumour suppresser genes (TSG) were considered separately. A Fisher's exact test was performed to see if a domain instance or 1D hotspot was significant in a given cancer type. In addition to this, they performed a co-location analysis of the 1D domain hotspots to general functional sites: catalytic sites from the CSA 101, post-tanslational modification sites from phosphositeplus 109 and interface residues from ProtInDB [193] and Mechismo [20]. Significantly mutated domain instances (SMD) were calculated as the total number of somatic missense mutations in the Pfam domain region of a particular gene, normalised by the corresponding domain length.

This analysis found that mutations in tumour-suppressor genes have strong biases towards particular domains, in agreement with Miller *et al* [154]. They also found that a single gene could be mutated in different domain regions in different cancers, therefore requiring discrete therapeutic attention, also shown in [189]. 1D hotspot co-location analysis revealed that mutations in oncogenes occur more on catalytic and phosphosite functional sites compared to tumour suppressor genes, by 32%. For the latter, mutations were shown to occur more at domain interaction sites and within the protein core. Finally mutations in oncogenes showed higher clustering tendencies than mutations in tumour suppressor genes which had a more dispersed distribution, in agreement with previous studies [226].

More recently studies have used other domain families and regions to analyse cancer mutations and identified 1D hotspots using alternative statistical approaches. For example Peterson *et al* adapted their methods, developed for previous studies [181] of TCGA [249], and analysed somatic missense mutations from 20 cancer types, within CDD and Pfam protein domains. In contrast to their previous work [182], they used an alternative

approach for 1D hotspot detection. For each of the 20 cancers, in each domain model, mutated domain hotspots were identified by counting the number of mutations at each position in a domain, across all patients. Significantly mutated positions in the domain were associated with an FDR derived P-value of 0.05 or 0.01. This was modelled using a hybrid distribution of a zero-inflated Poisson which accounts for a large amount of zero counts, and adjusts this accordingly. Random mutation models were derived by a random distribution of residues of equal number to the observed mutations.

The mutation hotspots were then analysed for their co-location to various functional sites including: UniProt functional features (nucleotide binding sites, DNA binding sites, calcium binding site, active site, and metal binding sites), and predicted conserved residues using the AL2CO method. A Fisher's exact test was used to measure the degree of enrichment at these sites, which was corrected for multiple testing using Bonferroni correction. The authors identified domains which contain hotspot(s) - derived from one or more genes in the domain alignment - as "Oncodomains". Candidate cancer genes were identified if they contained variants in an identified domain hotspot, and were analysed for enrichment in particular GO terms, using Pfam2GO annotations. Gene overlaps with other driver gene predictors were performed, including genes identified by MutSigCV, and known cancer genes from the Cancer Genome Census (CGC) and the NCI cancer gene index.

This analysis identified 185 CDD and 673 Pfam mutated protein families across 20 cancer types, containing 2126 CDD and 3563 Pfam hotspots respectively. The hotspots were found in very different locations in different cancer types and between patients. Functional feature analysis showed a statistical enrichment of oncodomain hotspots on all UniProt functional features (P-value $<$3.63E-87) - the highest co-location being for nucleotide binding sites, together with a significant enrichment of hotspot mutations on conserved residues (P-value $<$1.45E-9). In total, 3041 candidate cancer genes were identified, which included 56% of genes predicted by CHASM [28], 34% of genes predicted by MutSigCV, and 34% of CGC and NCI-CGI genes. Since many cancer genes can harbour mutations on many distinct interfaces, this study also highlighted the phe-

notypic pleiotropy within cancers, seen in previous studies [188].

Various studies have also incorporated other types of genetic variations, such as insertions and deletions (INDELS). Baissa *et al* [14] recently analysed the enrichments of 3 different mutation types, taken from COSMIC, in Pfam domains; missense mutations, INDELS, and truncations. This work used Pfam domains, but considered the annotation of the cancer gene types: oncogenes (ONC) genes and tumour suppressor genes (TSG) genes. Enrichment of mutations in Pfam domains were performed by counting the frequencies of the mutation type in that domain, and comparing this to the frequency in all other Pfam domains. This was further normalised using domain frequency, domain length, and number of samples. Statistical significance was measured using a Chi-squared association test.

This analysis was done for genes annotated as either ONC or TSG using the PANTHER functional classification tool [153], for dominant or recessive genes respectively. In order to measure the clinical significance of these cancer gene enrichments, the mutation frequencies of each Pfam domain containing cancer gene were compared to the mutation frequencies within 450 'random' domains considered "not cancer associated", where significant domains were identified using a Bonferroni correction. In addition to identifying enriched Pfam domains for the different types of genetic variation, 1D hotspots were identified for each variation type, as variations across gene members within a Pfam family were accumulated for common positions. Functional annotations of the Pfam domain, specifically GO annotations, were taken from the MOCKa database [199], where residues implicated in PTM (phosphorylation, glycosylation, ubiquitinylation), and PROSITE patterns are annotated. Pathway enrichments were taken from the KEGG database.

This study reported that the most frequent type of genetic variation in ONC and TSG genes were somatic missense mutations, showing a mutation percentage of 85% and 62% respectively. The ONC and TSG cancer genes types were associated with 310 and 197 Pfam domain families respectively, with 44 common Pfam domains between them. The Pfam domain enrichment results were further used in a machine learning

classification tool, in order to predict if a query gene is either an ONC or TSG based on their domain content (e.g types of Pfam domains). The classifier was trained on CGC genes, and assessment of performance showed an AUC of 0.72. For the hotspot analysis, the method identified 341 hotspots in total for the 3 mutation types, within 66 domains. The ONC genes had fewer mutation hotspots per domain, overall, compared to TSG genes for all 3 mutation types considered. In agreement with other studies, the 1D hotspots were in different locations within the overlapping ONC/TSG Pfam domains (SET, PKinase, RhoGAP). The 481 ONC and 133 TSG genes, derived from the enriched Pfam domains, were analysed for pathway enrichments, revealing 306 Pathways for ONC genes including those involved in biosynthetic processes, transcription, and protein amino acid phosphorylation. There were 76 pathways reported for the TSG genes, implicated in the cell cycle, cellular stress response, and DNA Damage response. GO analysis revealed that the ONC genes contain more transcription factors, and the TSG genes contain more enzymes. For genes containing both ONC and TSG Pfam domains, 14 enriched pathways were reported involved in the immune system, cell proliferation, and apoptosis.

So far, the studies discussed have considered the analysis of all mutations in protein structures. However, in more complex diseases such as cancer, the presence of both passive passenger and active driver mutations is likely to hinder mutation analysis. Therefore, in order to prioritise mutations which are more likely to be causing the disease phenotype, various approaches have used mutation clustering methods in order to identify protein positions enriched in mutations, which are therefore likely to be sites of driver mutations [189]. Furthermore, many studies have assessed the enrichment of mutations within defined regions within the protein sequence. These regions can include 1) residues surrounding a frequently mutated residue, 2) within protein domains, and 3) within disordered regions. For example, Tamberero *et al* [240] developed a method called OncodriveCLUST that analyses the enrichment of cancer mutations within a residue window in a protein sequence, and calculates a clustering score. A binomial model was used to identify significantly mutated residues and positions containing mu-

tations within 5 amino acids. Genes identified by this method were enriched for known cancer driver genes in the Cancer Genome Census, mostly having a dominant phenotype suggestive of oncogenes. Clusters were also found for recessive cancer genes but mutations within these clusters were more dispersed in their nature, and more likely to be associated with loss of function genes, consistent with studies described above [226] [271] [247].

Other studies used enriched domain annotations in driver gene detection. For example, the work of Lu *et al* [135] who compared the propensities of common (control), germ line and cancer variants - from OMIM and COSMIC respectively - in protein regions. Such protein regions were defined using Pfam domain annotations and disordered residues measured by DISOPRED [264]. The regions studied were intra-domain ordered, intra-domain disordered and inter-domain disordered regions. The use of propensities provides another means of quantifying regional enrichment of mutations in the protein sequence, and here this was normalised using region length and the relative frequencies of mutations in the rest of the protein. Both germ-line and cancer mutations were significantly depleted in both types of disordered regions, while possessing high propensities for ordered regions, especially in the case of germ-line variants.

More recent studies have also incorporated structure-based domains from the CATH database in driver gene detection. Hashemi *et al* [94] analysed the enrichment of cancer missense mutations in 29 cancers from the TCGA, within sequence based Pfam and structural CATH superfamily domains in identifying candidate driver genes. Significant mutation enrichment of mutations within a domain region was based on a binomial model of observing k mutations within a specific gene or domain of length l, compared to their occurrence within all coding regions of the genome of length L. Significantly mutated domains/genes were identified for each cancer type, using a p-value cut-off of 0.05, which underwent a Bonferoni correction. For the candidate genes within domains, these were also analysed on their degree of connectivity within the STRING database, compared to a random set of genes of equivalent size. In addition, domain candidate genes were compared with known cancer genes from COSMIC to test for cancer gene indication.

Lastly the SNPeffect predictor was used to predict impact scores for mutations within the Pfam and CATH domain genes.

Mapping domain annotations on mutated genes identified 759 CATH domain families within 2993 proteins, for 19 cancers. For the Pfam domain genes, there were 6009 Pfam domains within 17,722 genes, for 29 cancers. Although the CATH domains had a lower coverage of mutated genes, compared to Pfam domains, the CATH domain genes showed a higher overlap with cancer causing genes from COSMIC (that have specific domain type), at 65%, compared to the Pfam dataset, which gave a 52% overlap. In addition to this, the SNPeffect analysis showed that 14.3% and 17.7% of the mutations were reported as likely to have a high functional impact within the Pfam and CATH domains types, respectively. For the STRING connectivity analysis, the CATH domain candidate genes showed a significantly higher connectivity compared to random. This was not the case for the Pfam gene sets which showed a lower degree of connectivity between cancer types. This study therefore highlighted that the use of CATH domains in mutation analysis may provide a more informative and more functionally coherent insight into cancer genes and the mutations within them.

Structure-based Enrichment: 3D hotspots

Despite the success in detecting 1D hotspots, they are limited to detecting recurring and frequent mutations within the same nucleotide position, and depend on the number of samples available. A more sensitive method for detecting rare mutations is to see how mutations cluster within structural space. Therefore, other studies identify 3D hotspots, which also include the region surrounding highly mutated residues.

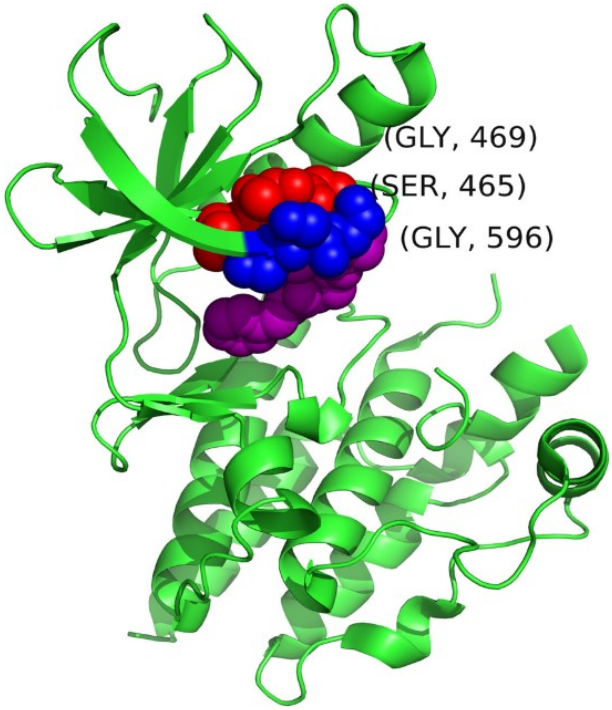
By considering the clustering of mutations in 3D, this not only detects positive selection at a given position, it also detects hotspots formed by mutations far from each other in the sequence but brought together in 3D upon folding, which may be overlooked by 1D hotspot methods. Other studies have shown that oncogenes are more likely to have mutation clusters at individual sites, compared to tumour suppressor genes which are more likely to contain mutations dispersed throughout the protein sequence [271] [247]

To date, many of the analyses of the 3D clustering of mutations have been performed with cancer missense mutations in whole protein structures. One of the earliest studies was performed by Stehr *et al* in 2011 [226]. They developed a clustering score for each gene which was proportional to the sum of distances between the centroids of mutated pairs across the protein, and normalised by the number of mutated residue pairs. Clustering was performed in protein domains, defined by DomainParser [270], to avoid bias of domain architecture and gene size. Significance of clustering was based upon a random control population, which was derived from a 1000 random mutations within the amino acid sequence. Much like the work on 1D hotspots, this approach revealed that mutations in oncogenes have higher tendencies to cluster within protein domains compared to tumour-suppressor gene mutations.

Another early study [200] also measured distances between mutation pairs. Ryslik *et al* analysed mutations from COSMIC comprising both oncogenes and tumour suppressor genes. The SpacePAC method was applied, which identifies spheres of different radii (between 1-10Å) around each mutation in the protein structure [200] and reports significantly mutated clusters based on the observed mutational counts falling in the tail of a Poisson distribution. Spheres of various radii from a central clustering position are analysed, where the cluster centre has the maximum normalised mutation count compared to randomly simulated mutations. SpacePAC identifies between 1 to 3 non-overlapping spheres of various radii, where different sphere combinations are assessed by their P-value to identify which combination encapsulates the most mutations. In doing this, one of the main assets of SpacePAC is that sphere combinations are assessed using the most enriched amino acids first, and so once an optimal sphere combination is found for a given protein, the programme terminates. This removes the necessity of having to sample every possible combination of overlapping spheres to capture mutations, unlike earlier approaches [201].

SpacePAC successfully identified 3 clusters in the ALK tyrosine kinase, at positions known to harbour activating mutations, thus further supporting the use of clustering in studying activating mutations in cancer. Other results from this study included the obser-

Figure 2.6: SpacePAC detects 3 mutation clusters within the known cancer driver BRAF kinase. The cluster centre positions are labelled and coloured in red, blue, and purple. Taken from [201].

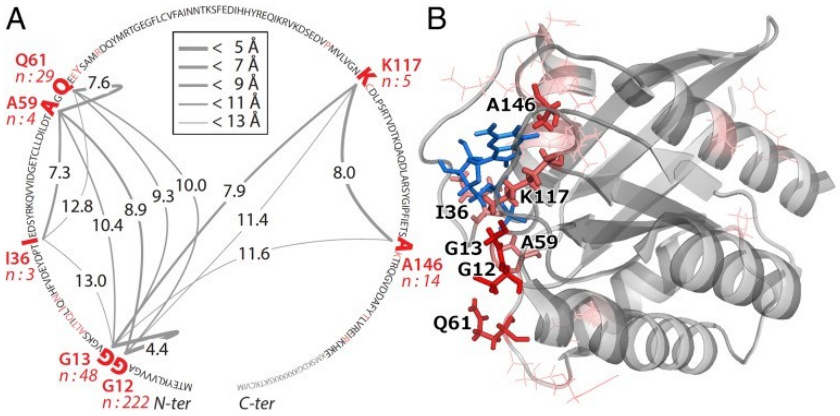


vation of co-location of these spheres to larger functional regions derived from UniProtKB. These include domain types, and other residue specific features; ATP/GTP binding sites, and other binding sites. Analysis showed that 77% of the hotspots within 18 proteins overlapped with a binding site or a domain. These included 3 within the BRAF kinase domain, which contains the known driver mutation position V600 on the activation loop (shown in figure 2.6).

Whilst 3D clustering methods have helped identify regions of cancer driver mutations, so far the methods treated all distances between cluster residues equally, and mutation frequencies were not explicitly accounted for in the clustering scores. In 2015, a similar method measuring centroid distances between mutated residue pairs was developed by Kamburov *et al* [110], called CLUMPS. A weighted average proximity (WAP) score was used, where distances between mutation pair centroids were accumulated for each residue, and then weighted by how many samples the residue was found mutated in (Figure 2.7). CLUMPS was used to analyse cancer mutations from COSMIC [15]

and the PanCancer compendium. Both oncogenes and tumour suppresser genes were considered.

Figure 2.7: CLUMPS methodology applied to the KRAS domain from the PanCancer dataset. A) Distance arcs of less than 13 Å between mutated centroids (red) are shown, where thicker arcs are shown for closer distances, within the protein sequence. B) Structure of KRAS where the mutated residues in hotspots are shown in red, and the GDP substrate is shown in blue. Taken from [110].



Unlike 1D hotspot enrichment studies, CLUMPS analysis showed a comparable enrichment of clusters in both oncogenes and tumour suppressor genes. As with earlier studies, the identified clusters were analysed according their co-location to functional annotations taken from PDBsum [123], which identified an enrichment of tumour suppressor gene clusters on protein-protein interaction sites.

Another recent analysis identifying 3D mutation hotspots in oncogenes and tumour suppressor genes includes the method of Tokheim *et al* [247]. This work also measures the structural properties of the hotspots, which were assessed according to region size, mutational diversity, amino acid physico-chemistry, and evolutionary conservation. Evolutionary conservation was measured using Shannon's entropy. The cancer mutations were taken from TCGA for 23 tumour types, and the hotspots were significantly mutated regions within the whole protein structure. A local mutation density for each residue (r) in a given structure was measured as the sum of the missense mutation count at the residue r and those residues occurring proximal to it, within 1nm or 1 amino acid side chain away. The observed value of this local density was compared to 1000 ran-

dom simulated permutations, where a significant local density has a P-value of 0.01 with Bonferroni correction.

This analysis found that hotspots in oncogenes were smaller in their size and encompass a smaller range of amino acid variant types compared to those within tumour suppressor genes, which were more diverse in their physicochemical properties and more dispersed throughout the structure, in agreement with related analyses [226] [271]. The authors also found an increased tendency for hotspot mutations to contain more evolutionary conserved mutations and occur within protein-protein interaction sites compared to non-hotspot mutations. These features were then used to identify 3D hotspots in different cancer genes, using a Naive Bayes classifier.

The results of this analysis showed that cancer associated proteins from the Cancer Genome Census (CGC) [257] contained 3D clusters of mutations exhibiting a significantly higher degree of cluster closeness than those found in non-cancer proteins. Regions of high cluster closeness correlated with Pfam functional domains, and there was significant preference of clusters to occur in conserved regions as measured by PhastCons, assessed using a T-test. However, this method only prioritises clusters with high closeness scores and therefore overlooks mutated regions of the protein that are more disordered, less globular, and that contain more dispersed clusters - such as in tumour suppressor genes.

More recent studies have also included an additional filtering step to select genes that are expressed above a certain threshold, thereby reducing genetic noise. These include the works of Gao *et al* [74] who studied 3D cancer mutation using TCGA and ICGC somatic missense mutation data from 41 cancers. Genes with low RNA expression levels ($<$0.1 TPM) in 90% of tumours were excluded. Clusters were identified using a threshold of 5Å , and mutations were aggregated across all samples using protein structure alignment. A robust statistical method was applied involving 10^5 randomisations of mutated residues. For specific examples, mutations coinciding with known functional sites reported in the literature were identified, and western blot analysis was used to measure the abundance of activated downstream signalling effectors (Phosphorylated

ERK1) and activated mutated genes (GTP-RAC1). The authors found 943 significant 3D clusters, in 503 genes and identified 3D hotspots in known tumour suppressor genes - PTEN, CDH1, and KEAP1. Further analysis showed the 3D clusters to occur near the catalytic site in PTEN, within the ca^{2+} binding region of CDH1, and in the NRF-2 binding region of KEAP1. Western blot analysis revealed that the 3D cluster mutations in MAP2K - occurring in the regulatory helixA region - disrupted MAP2K activity, leading to its abnormal activation, and elevated downstream phosphorylated-ERK1 levels.

In this chapter, domain annotations from the CATH database were used to detect cancer missense mutation enrichment in domains and enriched 3D clusters, and thereby predict putative driver genes. Specific examples of these 3D clusters (MutClusters) are analysed in further detail to determine possible functional effects based on functional site proximity. In summary, a large-scale analysis of proximity to functional sites was performed for germline non-cancer, somatic cancer, and predicted driver mutations. This showed that putative driver mutations were enriched at various functional sites from known sources, and in house predicted sites based on conserved residues seen in CATH FunFams.

Materials and Methods

Mutation Data

Germline mutations:

Germ-line disease mutations from OMIM: OMIM germ line mutations were taken from the Gene3D Database containing the OMIM version in 2012 [10], and labelled according to whether they were associated with cancer or not. This was done using a perl script to identify different terminologies describing cancer; “Cancer”,”Carcinoma” and “Blastoma”. This resulted in a non cancer dataset of 212,544 mutations. The OMIM cancer dataset was further divided into oncogenes and tumour suppressor genes using a cancer gene list containing oncogene and tumour suppressor gene annotations created by Vogelstein et al [257] and the TSGene database [282] respectively.

Disease mutations from UniProt: UniProt disease associated mutations and neutral mutations were taken from UniProt-HUMSVAR release March 2014 [11], annotated as “disease” and “polymorphisms” respectively. To avoid bias of heavily mutated genes, both UniProt mutation datasets were filtered to exclude UniProt accessions with more than 50 mutations. Furthermore, if a position harboured more than 1 mutation, one variant was selected at random. Both of these criterion have been used in previous studies [75]. Filtering the UniProt datasets reduced the disease mutation dataset from 770 genes to 688 genes for the disease mutations set, and the neutral dataset from 1947 genes to 1926 genes. This gave a total of 6300 mutations in 688 UniProt accessions for the disease mutations, and 8838 mutations in 1926 UniProt accessions for the neutral group. See table 2.1.

Somatic cancer mutations:

Somatic mutations implicated in cancer were taken from the COSMIC Cancer Gene Census curated by Wellcome Sanger for genes [15]. These 600 genes included somatic missense mutations in annotated oncogenes and tumour-suppressor genes.The cancer

mutations within the COSMIC datasets were within 29 oncogenes and 40 tumour suppressor genes.

Mapping of mutations to 3D structures

For each of the mutation groups, the UniProt to PDB mapping was extracted from the Gene3D tables in the CATH database, which are based on the SIFTS mapping algorithm [254]. The best structure for a UniProt sequence was selected based on 1) the maximum mapped UniProt sequence length and the best structural resolution, as in studies by Stehr et al [226]. The number of mapped entries for each mutation type is shown below in table 2.1.

Table 2.1: The different disease mutations analysed and the number of entries mapped to a PDB structure.

Mutation dataset	**Total Mapped Entries**
OMIM Non cancer	7,523
COSMIC oncogenes	1,893
COSMIC Tumour suppressors	3,184
UniProt Disease	6,300
UniProt Neutral	8,838

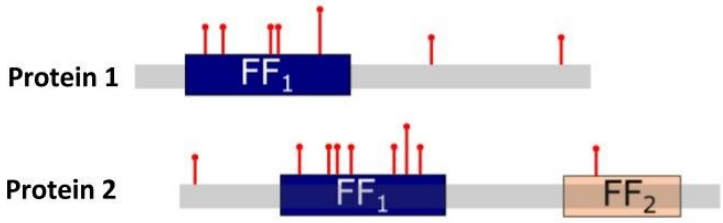
Identifying domains enriched in mutations (MutFams) and 3D clusters enriched in mutations (MutClusters)

We identified CATH domain functional families (FunFams) that are enriched in mutations (MutFams). A number of cancer disease mutation datasets were analysed to identify MutFams, which included mutations implicated in bladder cancer (BLCA), glioblastoma multiforme (GBM), low grade glioma (LGG), gliomas (GLI), breast cancer (BRCA), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), kidney and renal cell carcinoma (KIRC), prostate adenocarcinoma (PRAD), skin cutaneous melanoma

(SKCM), acute myeloid leukemia (LAML), stomach adenocarcinoma (STAD), renal adenocarcinoma (READ) and colorectal adenocarcinoma (COAD). Cancer mutations were taken from COSMIC [15], TCGA [249] and the ICGC [281] resources. Mutations were associated with their respective CATH FunFam domains to detect enriched FunFams (MutFams) for each cancer type. The enrichment analysis was performed using a program written by Dr Paul Ashford in the Orengo group, and a schematic of the methodology to detect enrichment is shown in figure 2.8.

Akin to previous methods by [154], we used a mutation enrichment score for a CATH FunFam, described by equation 6. The FunFam enrichment score is based on the observed mutations in a FunFam (mf) by summing all mutations in all domains within a FunFam domain boundary. The expected mutation count (me) is calculated based on the total number of mutations observed in all genes containing the FunFam, and the fraction of amino acids within FunFams compared to the total length of the genes.

Figure 2.8: Enrichment of mutations in CATH FunFams(FF). The red bars reflect the number of mutations. The equation calculates the enrichment factor for the FunFam in question



To assess the significance of the observed enrichment (ef) of mutations in each FunFam, a permutation test was used based on previous work by Miller *et al* [154]). For each FunFam, the set of (human) genes was collected. To create a random mutation model, these genes were then subjected to a residue based permutation test. For each of the 1000 permutation iterations, the total number of mutations within the FunFam genes were counted, as mi. The P-value was defined as the proportion of iterations where mi $>$mf, where mf is the observed mutations count for a FunFam, see equation 6.

Equation 6:

$$ef(enrichment factor score) = mf(observed mutation burden)/mi(expected mutation burden)$$

Correction for multiple testing

To reduce noise, for each cancer type, positions where the total mutations count was $<$than 10 were excluded. This was applied within FunFam boundaries across all human genes, and for multiple-spanning discontinuous sequence ranges, where applicable. Additionaly, MutFams with enrichment factor ef $<$1 , were removed as these are, by definition, not enriched. P-values obtained for MutFams were corrected for multiple testing, using the Benjami-Hochberg (BH) correction. A false Discovery Rate (FDR) of 5% was applied using the R function “p.adjust” on all filtered MutFams, for each cancer type.

Identification of mutationally enriched 3D clusters (MutClusters)

MutClusters are distinct from 1D mutation hotspots as they involve multiple residues that cluster in 3D, as opposed to an enrichment of mutations at a single residue. Firstly, mutationally enriched regions within CATH FunFam domains were identified by the MutCluster program written by Dr Paul Ashford in the Orengo group. For each MutFam with a significant enrichment factor of $>$1.5 and corrected P-value $<$0.01, the observed mutations at a specific MutFam position are tallied from all FunFam gene members and mapped onto the FunFam representative structure. This is the structure with the highest average structural similarity, measured by SSAP [173] to all other relatives in the

FunFam. The density of mutations within a spherical volume of 5Å is calculated and a permutation test is performed for each sphere, to identify regions which harbour more mutations than expected by chance, with a P-value of <0.01. The permutation test is done by generating 5000 random mutations within the protein structure.

Proximity analysis of single and clustered mutations to functional sites

Known functional sites

To analyse the proximity of single mutations and MutCluster mutations to functional sites, we used known functional site data including; catalytic residues from CSA [192], and protein-protein interaction sites from NCBI-IBIS [216]. We also used UniProt functional features extracted from the UniProt functional features table on May 2009 [11]. These were mapped to the PDB structure, as described in (section 2.2.1.1). The functional feature types include; MOD_RES, METAL, NP_BIND, ACT_SITE, SITE, CARBOHYD, see table 2.2. Other UniProt features which were considered but excluded due to lack of data were; CA_BIND, ZN_FING, DNA_BINF, MOTIF, ACT_SITE. CARBOHYD and LIPID.

Table 2.2: UniProt functional features included in the proximity analysis and their descriptions.

UniProt feature	Description
MOD_RES	Post translational modification of a residue. Includes, Acetylation, amidation, hydroxylation, methylation, phosphorylation, formylation, blocked group, pyrrolidone and sulfation
METAL	Binding site for a metal ion, such as iron or copper
ACT_SITE	Amino acids involved in the catalytic activity of an enzyme.
BINDING	Amino acids involved in the binding of a chemical group, such as a prosthetic group or a co-enzyme.
CARBOHYD	Describes the occurrence of the attachment of a glycan (mono- or polysaccharide) to a protein residue. This include C-,N-, or O- linked glycans.
SITE	Any interesting amino acid on the sequence that is not defined by any other feature key. It can also apply to an amino acid bond which is represented by the positions of the 2 flanking amino acids, such as a cleavage site for a protease

Other known functional residues included ligand binding residues, were taken from the IBIS resource [216], and from ccPDB [218]. The ccPDB data were from compiled datasets updated in 2012 available online. The choice of ccPDB ligand types was based on a recent article by Martinez *et al* [108] as shown in table 2.3.

Table 2.3: Ligands considered in the MutDist proximity analysis

Ligand abbreviation	Molecule name
ATP	Adenosine triphosphate
ADP	Adenosine diphosphate
FAD	Flavin adenine dinucleotide
FMN	Flavin mononucleotide
GDP	Guanosine triphosphate
HEM	Protoporphyrin IX containing Fe
NAD	Nicotinamide adenine dinucleotide
PLP	Pyridoxal – 5'- phosphate
UDP	Uridine diphosphate

Predicted functional sites

In addition, proximity to predicted functional sites was analysed using predicted sites from a range of resources.

Sequence based: Conserved sites within a FunFam alignment - FunSites

For a given FunFam, predicted functional sites (named FunSites) were identified using an in-house programme called Scorecons [253], which detects sites with highly conserved residues. Scorecons scores range from 0 for unconserved to 1 for completely conserved sites, where significantly conserved residues possess a scorecons value of greater than or equal to 0.7. Proximity to scorecons sites was analysed only if the mutated protein was in a FunFam with a Diversity Of Positions (DOPS) score of at least 70. DOPS score measures the information content of a FunFam multiple sequence alignment. Here a DOPS threshold of 70 has been shown to be optimal in obtaining conserved positions enriched in known functional sites [41].

Structure based - Predicted allosteric sites

We predicted allosteric residues in a protein structure by using a novel in-house method, BC-Site (Aurelio Goya Marcia, personal communication). The protein domain structure is modelled as a residue network, where each residue is a node connected to other residue nodes. The protein domain contact network is derived from the cross correlation matrix obtained by normal mode analysis (NMA). This matrix is based on whether 2 residues in a pair exhibit correlated fluctuations, and network edges are weighted by the degree of strength of these motions. These weights can be transformed into “effective distances” between contacting nodes, where a high correlation suggests a strong information flow between the 2 residues, associated with shorter effective distances [35].

The effective distance between 2 residues nodes are referred to as vertices. The betweenness centrality of a residue is the number of shortest paths from all vertices that pass though that residue, compared to all other vertices. This method measures the residue’s position in linking between residue communities, and captures sites involved in

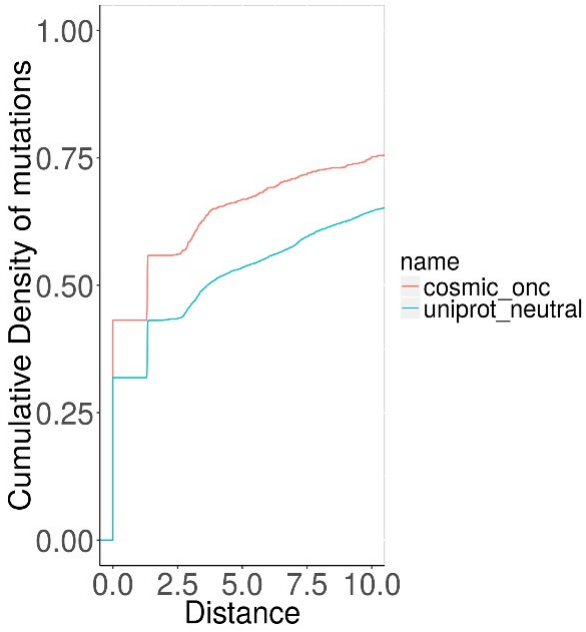
transferring signals between distal regions within the protein [35]. Since these residues have been shown to be important in controlling the flow of information within a residue network, they are thought to be important for allosteric communication [50] [112]. The BC- site residues do not distinguish between surface/non surface or sites that are directly affected upon ligand binding on the surface of the protein, but rather residues which act as linkers between residue hubs within the protein to facilitate cross-protein communication. Therefore in addition to considering BC sites, for specific analyses of MutCluster residues, we used surface allosteric sites predicted by the STRESS program [35]. These predicted allosteric residues were identified based on them undergoing significant conformational changes upon ligand binding, measured using Monte Carlo simulations (see Chapter 1, page 61).

MutDist method

The MutDist program written in perl consists of a number of modules which parse the mutation data, structural data and functional site data. Subsequently, the closest atomic distance between a mutated residue and a functional site residue is determined. This was performed for each of the mutated residues in the OMIM and COSMIC datasets, which could be mapped to a structure. The proximity of mutated residues within MutClusters to functional site residues was also analysed, again taking the closest atomic distance between the mutated residue to the nearest functional site residue. Furthermore, a program was written in R to generate plots of the cumulative density functions (CDF) of these distances for each mutation dataset, where the cumulative probability of mutations was plotted for each distance to the functional site being considered. An example of this is shown in figure 2.10.

Statistical Tests: Whilst a number of statistical tests could be applied to test for significance (e.g Fisher's exact text, KS-test), the Fisher's exact test was used as this has been widely reported in the literature for these types of studies and allows us to directly compare our results with those of Gao *et al* [75]. In order to assess the significance of the different distributions of the disease causing mutations compared to neutral mutations,

Figure 2.9: An example of a cumulative density plot (CDF) for proximity of disease mutations from COSMIC-ONC to IBIS-PPI sites



odds ratios (OR) were calculated and the Fisher's Exact test was performed in R for the contingency table 2.4 shown below. This is a similar analysis to that employed by Gao *et al* [75]. In this work, the significance of a mutation type to occur close to a functional site (between 0-8Å) is described by a P-value, which was considered significant if the P-value is less than or equal to 0.01, representing the 1% significance threshold.

Table 2.4: Contingency table for disease versus neutral mutations at a given distance, for odds ratio calculations

Distance (X or Y) boundary	Disease mutations	Neutral mutations
(X-Y) or (0-X)	A	C
Not in boundary or threshold	B	D

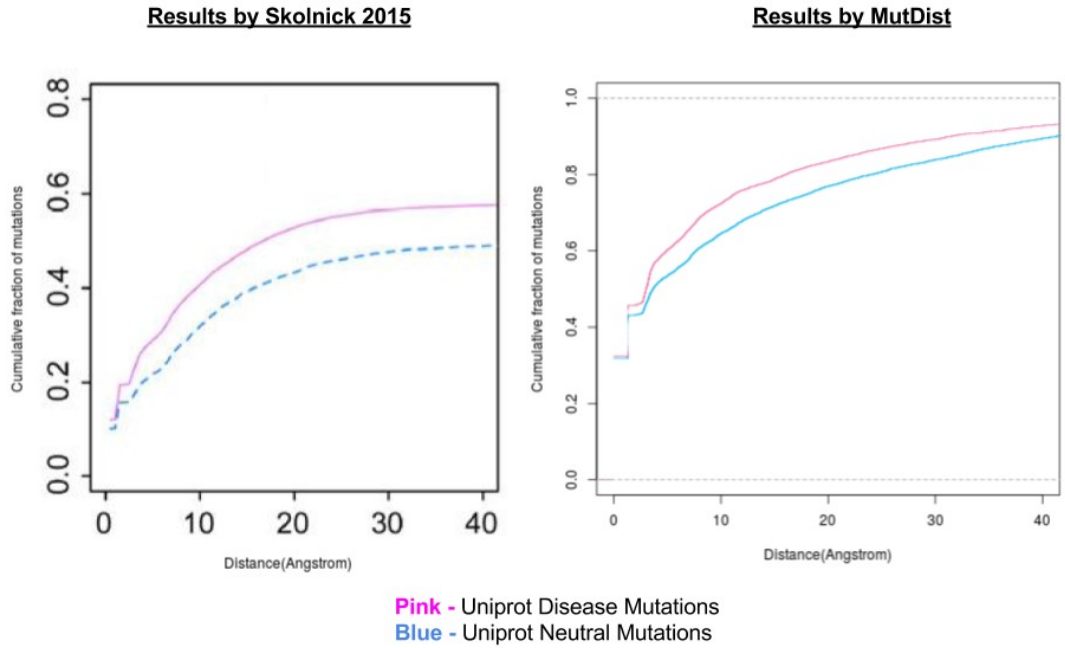
Comparing the results obtained by MutDist with those for similar analyses of proximity of mutations to functional sites

In order to test the reliability of MutDist and the developed computational framework, analyses of UniProt disease mutations and neutral mutations were compared to the re-

sults obtained from similar studies performed by Gao *et al* [75]. The study by Gao *et al* explored the proximity of mutations to in-house defined protein protein interaction sites, where the complexes used were identified as entries in the PDB which had a "protein complex" label within their biological unit. The interface sites were then defined as residues having heavy atom distances of at most 4.5Å between protein partners. MutDist was run on protein protein interaction sites collected from IBIS [216] which are obtained from both experimental data reported in the literature and homology analyses. The cumulative density functions for each of the studies are shown below in figure 2.10.

In both studies, disease mutations show higher probabilities to be close to interface sites than neutral mutations in UniProt. This relationship is similar in the 2 studies. In terms of differences, Mutdist showed overall higher probabilities of both neutral and disease mutations to be close to a PPI site, at each distance. This is likely to to be due to the fact that the interaction sites used in the MutDist analysis were more comprehensive as they were derived from a wider set of sources. The similarity in the trends observed justified the use of MutDist in larger scale studies.

Figure 2.10: Comparing the proximity of UniProt disease and neutral mutations to PPI sites, analysed using the Skolick method and the MutDist method



Results

Analysis of the proximity of known disease associated mutations and predicted cancer driver mutations to functional sites

MutDist analysis was performed on germline mutations using datasets taken from OMIM, for non-cancer diseases. We also considered mutations in different types of cancer genes identified in COSMIC – oncogenes and tumour suppressor genes, which included both germline and somatic cancer mutations. To analyse the functional effects of the MutClusters on a large scale, a proximity analysis using MutDist, was also performed for all MutClusters. As discussed already in methods, the MutCluster protocol is designed to detect mutationally enriched regions and thereby filter out noise from passenger mutations. For each dataset, the distances of mutations to functional sites were shown as an empirical cumulative density functions for each functional site type.

Analysis of proximity of mutations to catalytic sites

According to figure 2.11, UniProt disease mutations show a modest tendency to occur close to CSA sites, similarly germline non cancer mutations. The UniProt disease mutation dataset were included for reference and contain both germline non cancer and cancer mutations. This result is consistent with studies by [75], which reported that 17% of UniProt disease mutations co-locate to predicted enzyme functional determinants identified by EFFICAz [75].

Somatic cancer mutations in both oncogenes and tumour-suppressor genes showed the lowest tendency to occur close to CSA sites. However MutClusters showed a high tendency to be close to CSA sites, producing the highest enrichment with an odds ratio of 2.98 at distances 0-8 Å, compared to UniProt neutral mutations (see table 2.5) In summary, all disease mutations show a higher tendency to be close to catalytic residues than neutral mutations. This is enhanced for cancer mutations by filtering out passenger mutations using the MutCluster enrichment protocol.

Figure 2.11: Comparing the proximity of UniProt disease and neutral mutations to CSA sites.

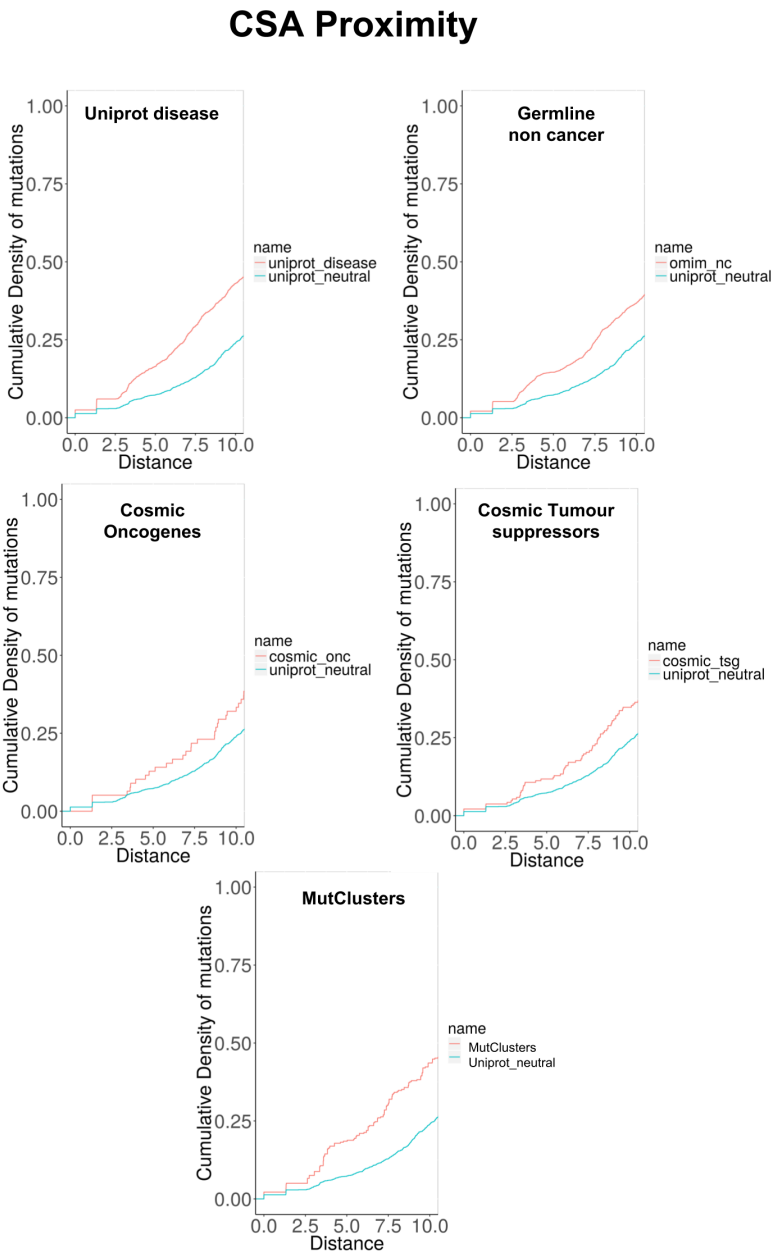


Table 2.5: Overall tendencies of disease mutations to occur close to CSA sites compared to the UniProt neutral mutations. Fisher's exact test was performed against all mutation types to the UniProt neutral.

Mutation type	Number of distances	Distance boundaries	Odds ratio	P-value	significance
Germline non cancer	2598	(0-8)	2.270614922	2.20E-16	**
Cosmic oncogenes	78	(0-8)	1.723322684	0.05304	
Cosmic tumour supp	187	(0-8)	1.715344338	0.004351	**
UniProt disease	1562	(0-8)	2.809229542	2.20E-16	**
MutClusters	319	(0-8)	2.981621786	3.34E-15	**
UniProt neutral	2111	-	-	-	-

Analysis of proximity of mutations to protein-protein interaction sites

According to figure 2.12, the UniProt disease mutations, consisting of both non-cancer and cancer mutations, show a slight but significant tendency to occur close to IBIS-PPI sites, consistent with results from previous studies [42] [43] [75] showing PPI sites to be hotspots for UniProt disease mutations. But since this dataset contains a mixture of diseases in both non-cancer and cancer, we also analysed germline non-cancer mutations.

The OMIM germline non cancer mutations showed no significant tendency to occur close to IBIS-PPI sites. This may be due their propensities to occur within the protein core, where they cause protein destabilisation [277], or their occurrence in ligand binding sites [108]. Schuster *et al* has also showed that only 4% of germline disease mutations from OMIM and UniProt are involved in protein interaction sites [206]. A recent large scale study was performed by Gress *at al* in 2017 [90], who compared cancer and non-cancer disease missense mutations. They revealed that protein-protein interaction interfaces are not enriched for either disease causing mutations. However, somatic cancer mutations in oncogenes and tumour-suppressor genes both show significant tendency to occur close to IBIS-PPI sites. This tendency is further increased for the MutCluster mutations, which show the highest odds ratio of 2.51 compared to neutral mutations.

Figure 2.12: Cumulative density functions of germline and somatic cancer single mutations and MutCluster mutations to protein-protein interaction sites.

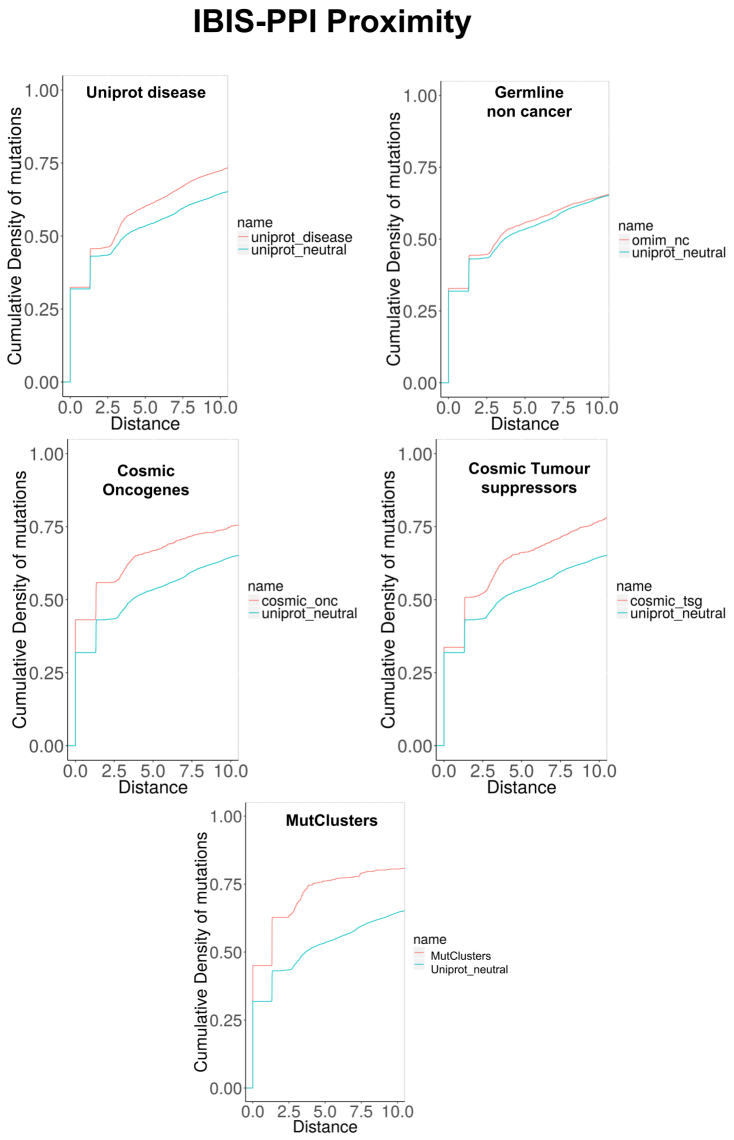


Table 2.6: Overall tendencies of disease mutations to occur close to IBIS-PPI sites compared to UniProt neutral mutations. Fisher's exact test was performed against all mutation types to the UniProt neutral

Mutation type	**Number of distances**	**Distance boundaries**	**Odds ratio**	**P-value**	**significance**
Germline non cancer	5479	(0-8)	1.049274235	0.2018	
Cosmic oncogenes	990	(0-8)	1.71150847	3.53E-13	**
Cosmic tumour supp	1179	(0-8)	1.709584988	4.65E-15	**
UniProt disease	4478	(0-8)	1.412395121	2.20E-16	**
MutClusters	1897	(0-8)	2.517012866	2.20E-16	**
UniProt neutral	6582	-	-	-	-

Analysis of proximity of mutations to ligand binding sites

According to figure 2.13, the UniProt disease mutations show a significant enrichment at ligand binding sites, with an odds ratio of 1.6. A significant signal is also seen for the germline non cancer mutations from OMIM, consistent with other studies showing that germline disease mutations from HGMD, exhibit a high tendency to co-locate to ligand binding sites [175] (see table 2.7).For the somatic cancer variants, the mutations in oncogenes produced a significant enrichment at ligand binding sites, with an odds ratio of 1.46. In contrast, somatic mutations in tumour suppressor genes were not significantly enriched. Mutations within MutClusters showed the highest enrichment at ligand binding sites, with a significant odds ratio of 1.78.

Figure 2.13: Cumulative density functions of germline and somatic cancer single mutations and MutCluster mutations to ligand binding sites.

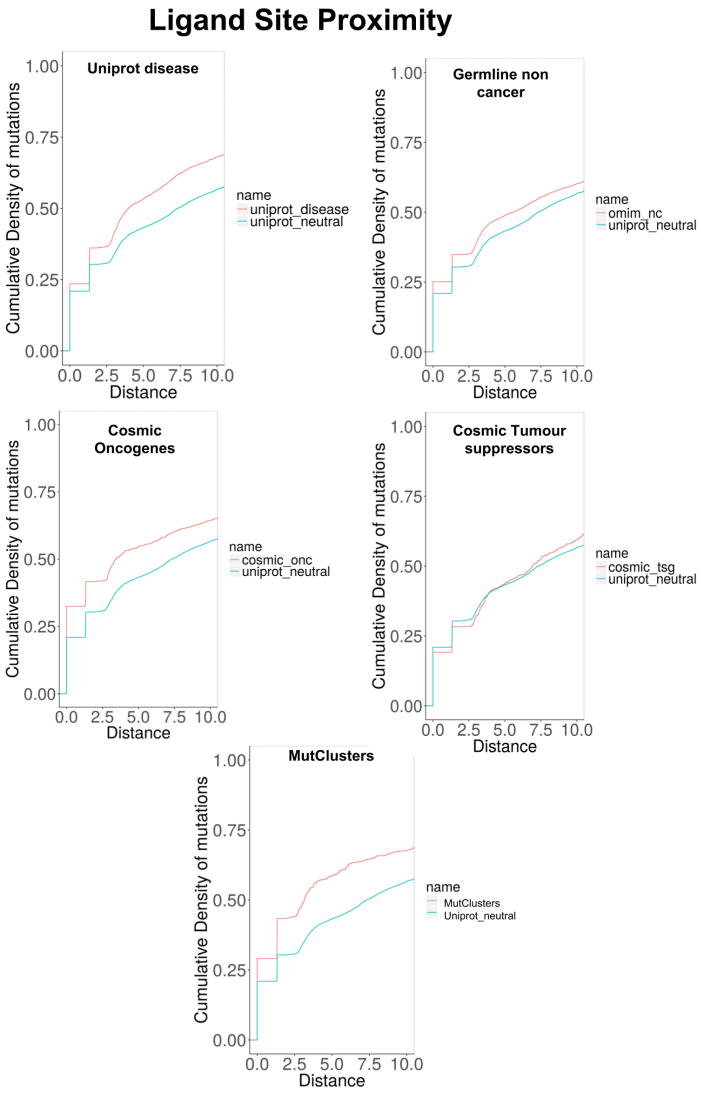


Table 2.7: Overall tendencies of disease mutations to occur close to ligand binding sites compared to the UniProt neutral mutations. Fisher's exact test was performed against all mutation types to the UniProt neutral.

Mutation Type	Number of Distances	Distance Boundaries	Odds Ratio	P-value	significance
Germline non cancer	5575	(0-8)	1.202882585	5.28E-07	**
Cosmic oncogenes	936	(0-8)	1.462814882	9.39E-08	**
Cosmic tumour supp	788	(0-8)	1.096129955	0.2269	
UniProt disease	4205	(0-8)	1.6235527	2.20E-16	**
MutClusters	1584	(0-8)	1.776610699	2.20E-16	**
UniProt neutral	6414	-	-	-	

Analysis of proximity of mutations to predicted FunSites

According to figure 2.14, the UniProt disease mutations show a modest but significant tendency to occur close to FunSites. This dataset contains germline non cancer, and cancer mutations. Germline non cancer mutations showed a higher and very significant tendency to lie close to these conserved sites, suggesting a possible functional explanation of the impacts of germline non cancer mutations. However, a low proportion of these mutations are close to catalytic and interface residues (see figure 2.11 and figure 2.12), suggesting that these mutations may be impacting other types of sites identified by the FunSite protocol, i.e ligand binding sites or allosteric sites.

Somatic cancer mutations in oncogenes also produced a significant tendency to lie close to FunSites, which was further enhanced for the MutCluster mutations which produced the highest odds ratio of 84.7 compared to neutral mutations shown in table 2.8. There was not enough distance data for the somatic mutations in tumour-suppressor genes, and so they were not analysed.

In summary, all disease mutations showed a significant tendency to occur close to FunSites, and the highest tendency was seen for MutCluster mutations and the germline non cancer mutations. This is consistent with studies that show both germline non cancer and somatic cancer disease mutations to occur close to conserved sites [140].

Figure 2.14: Cumulative density functions of germline and somatic cancer single mutations and MutCluster mutations to predicted FunSites.

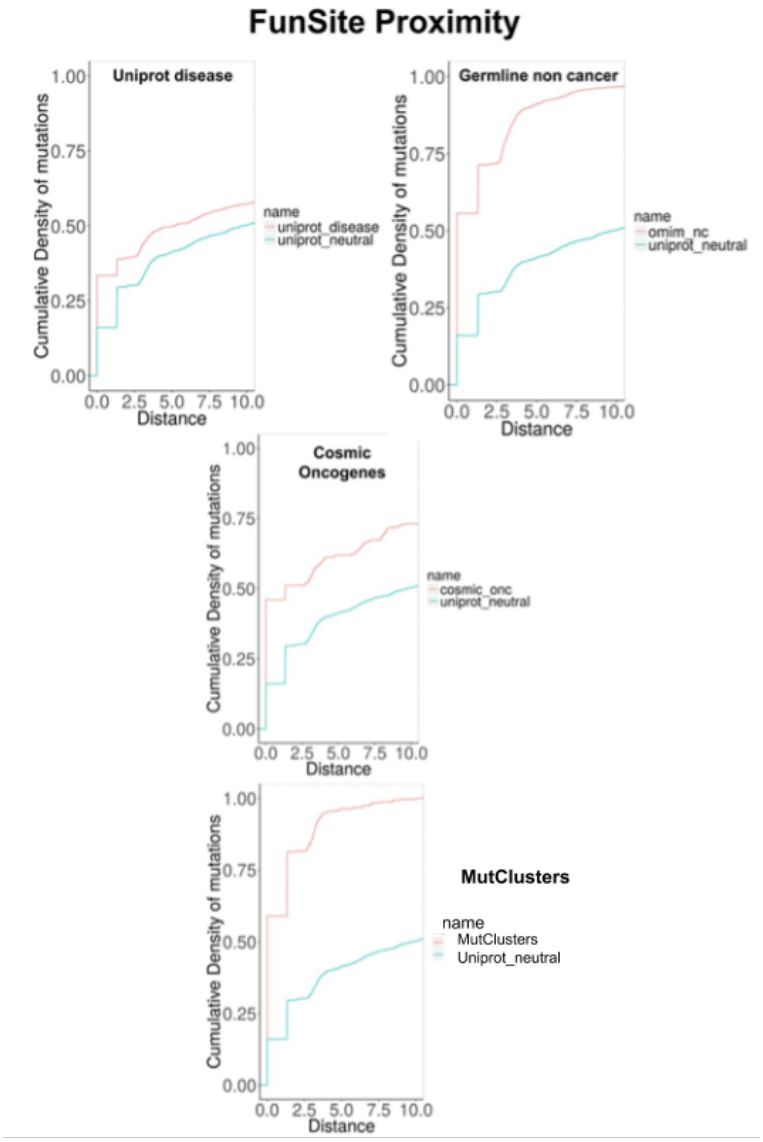


Table 2.8: Overall tendencies of disease mutations to occur close to predicted FunSites compared to the UniProt neutral mutations. Fisher's exact test was performed against all mutation types to the UniProt neutral.

Mutation type	**Number of distances**	**Distance boundaries**	**Odds ratio**	**P-value**	**significance**
Germline non cancer	2360	(0-8)	25.50585912	2.20E-16	**
Cosmic oncogenes	278	(0-8)	2.451820678	3.54E-11	**
UniProt disease	1289	(0-8)	1.365189157	3.36E-05	**
MutClusters	845	(0-8)	84.67358626	2.20E-16	**
UniProt neutral	1613	-	-	-	-

Analysis of proximity of mutations to UniProt functional features

In addition to CSA, IBIS-PPI and predicted Funsites, UniProt functional features were also considered. The first is the "MOD-RES" functional feature, which describes residues that play a functional role in post translational modifications, such as acetylation, amidation, hydroxylation, methylation, sulfation, formylation, pyrrolidone addition, and phosphorylation.

Modified residues (MOD_RES)

All disease mutation types show significant tendencies to occur near MOD_RES sites, shown in figure 2.15 and table 2.9. The somatic cancer oncogenes shows the highest tendency to be proximal compared to neutral mutations, producing an odds ratio of 2.45. In contrast, the somatic cancer mutations in tumour-suppressor genes produced the lowest tendency to be proximal to the MOD-RES sites. The tendency of MutCluster mutations to be proximal to MOD_RES sites is in-between the somatic oncogenes and tumour-suppressor genes, producing the second highest significant odds ratio of 2.15.

A difference between oncogenic and tumour-suppressor gene mutations was also seen in studies by Stehr *et al* [226] , who also examined proximity to post translational modifications. Studies by Fan Yang *et al* [271] also showed that oncogene hotspots co-locate more with post-translational modifications, than hotspots within tumour-suppressor genes.

Figure 2.15: Cumulative density functions of germline and somatic cancer single mutations and MutCluster mutations to UniProt MOD_RES sites.

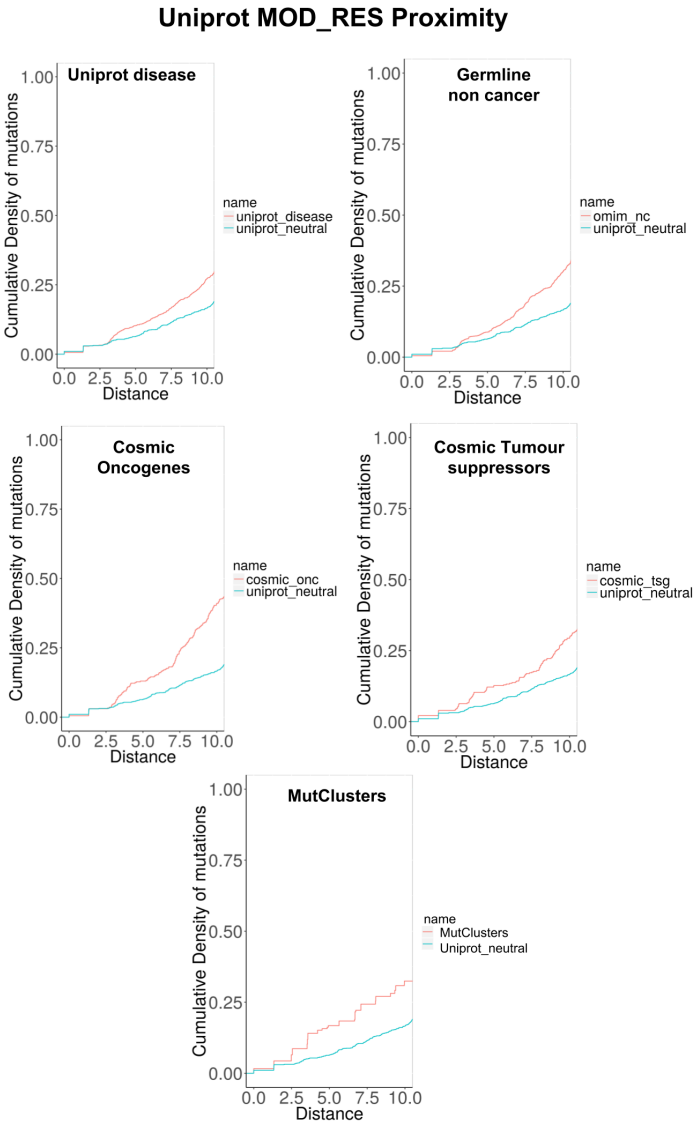


Table 2.9: Overall tendencies of disease mutations to occur close to UniProt MOD_RES sites compared to the UniProt neutral mutations. Fisher's exact test was performed against all mutation types to the UniProt neutral.

Mutation type	Number of distances	Distance boundaries	Odds ratio	P-value	significance
Germline non cancer	1570	(0-8)	1.842581789	2.00E-09	**
Cosmic oncogenes	392	(0-8)	2.44789357	6.42E-10	**
Cosmic Tumour supp	378	(0-8)	1.520661157	0.009245	**
UniProt disease	1385	(0-8)	1.568368174	2.90E-05	**
MutClusters	185	(0-8)	2.150649351	0.00012	**

Metal binding residues (METAL)

According to figure 2.16, all disease mutations show significant enrichment to UniProt metal binding residues. The MutCluster mutations show the highest proximity to UniProt metal binding residues (see table 2.10). This high enrichment for cancer driver mutations agrees with the study by Talavera *et al* [239], which showed that cancer mutations under positive selection are enriched at metal binding sites compared to cancer passenger and neutral mutations.

Figure 2.16: Cumulative density functions of germline and somatic cancer single mutations and MutCluster mutations to UniProt metal binding sites.

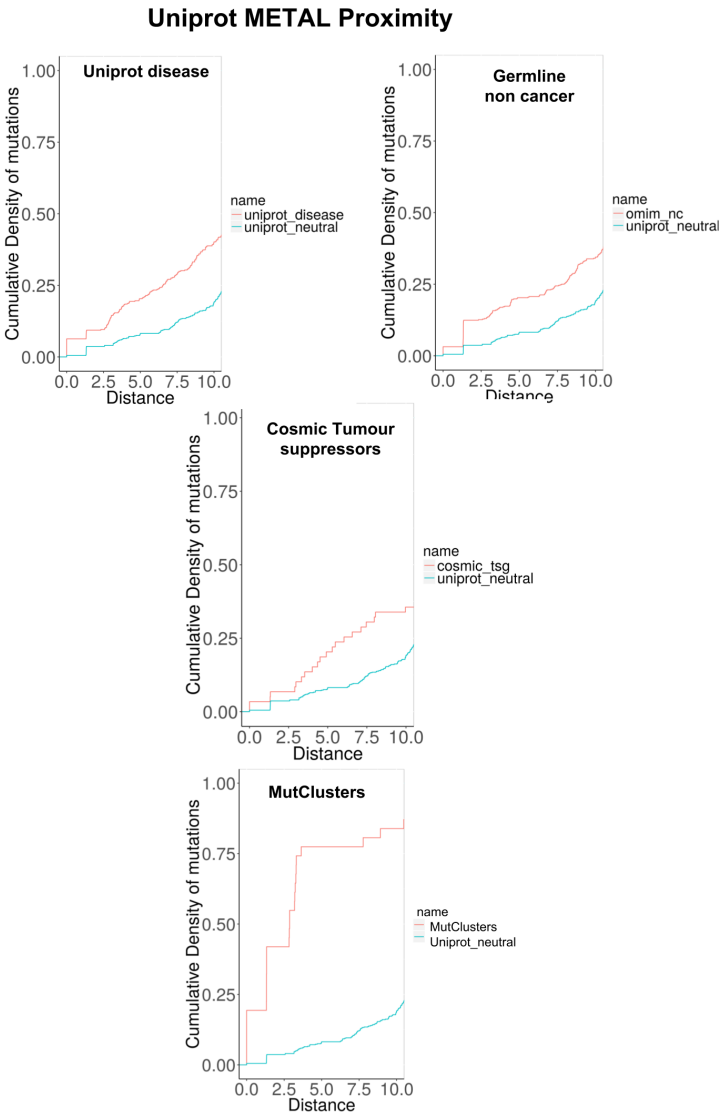


Table 2.10: Overall tendencies of disease mutations to occur close to UniProt metal binding sites compared to the UniProt neutral mutations. Fisher's exact test was performed against all mutation types to the UniProt neutral.

Mutation type	**Number of distances**	**Distance boundaries**	**Odds ratio**	**P-value**	**significance**
Germline non cancer	507	(0-8)	2.169788107	8.88E-07	**
Cosmic Tumour supp	59	(0-8)	3.05974026	0.0004567	**
UniProt disease	428	(0-8)	2.77913391	1.66E-10	**
MutClusters	31	(0-8)	26.83982684	1.05E-15	**

Analysis of proximity of mutations to predicted allosteric sites

According to figure 2.17, the germline non cancer mutations show a statistically significant enrichment at betweenness centrality (BC) sites, with an odds ratio of 1.92 and a P-value of 2.2E-16 between 0 and 8Å from BC sites (table 2.11). This tendency of germline disease variants to occur proximal to sites of high betweenness centrality is consistent with studies by Clarke *et al* [35].

In contrast, the unfiltered cancer mutations from COSMIC show no enrichment compared to the UniProt neutral model. Furthermore, the MutCluster mutations show a statistically significant depletion with an odds ratio of 0.25. This trend conflicts with studies by Shen *et al* [211] which showed cancer mutations to be enriched at allosteric residues. However, the predicted allosteric sites used in such studies were more frequently found on the surface compared to BC sites.

Other studies have also showed that cancer mutations are more likely to be observed on the surface of the protein than germline disease mutations [135] [83], which may explain the depletion of cancer mutations at BC sites. In contrast, the enrichment of germline disease mutations at BC sites is likely to be due to the fact that these mutations tend to occur in the protein interior.

Figure 2.17: Cumulative density functions of germline and somatic cancer single mutations and MutCluster mutations to predicted betweenness centraility (BC) sites.

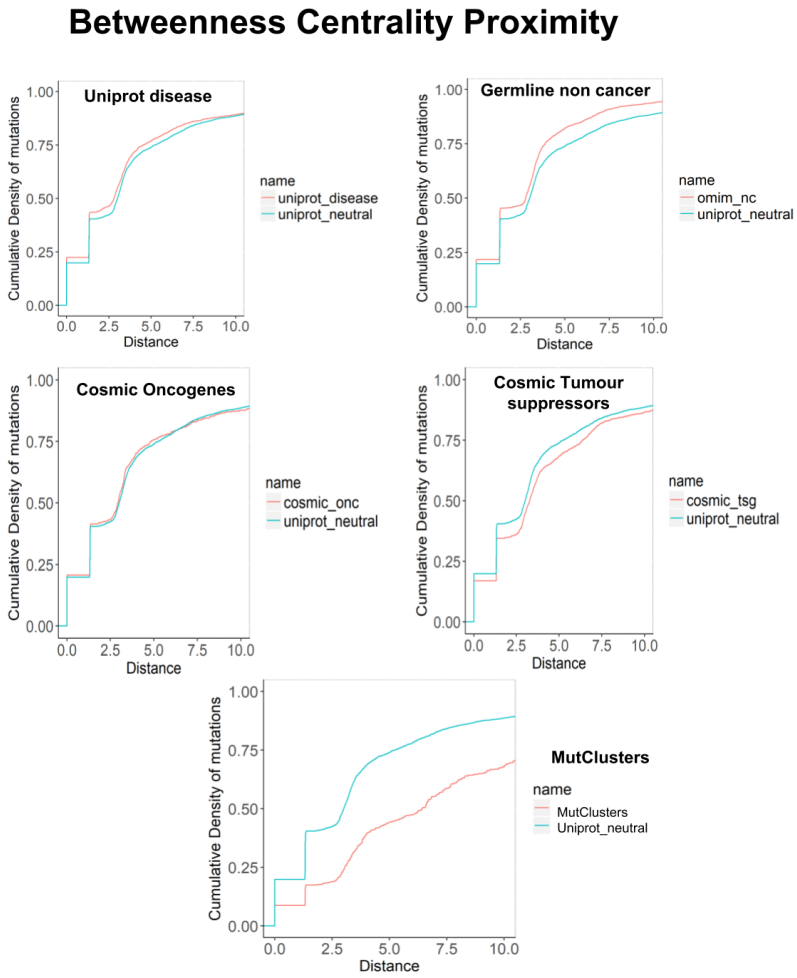


Table 2.11: Overall tendencies of disease mutations to occur close to predicted allosteric sites compared to the UniProt neutral mutations. Fisher's exact test was performed against all mutation types to the UniProt neutral.

Mutation type	**Number of distances**	**Distance boundaries**	**Odds ratio**	**P-value**	**significance**
Germline non cancer	3649	(0-8)	1.920792423	2.20E-16	**
Cosmic oncogenes	625	(0-8)	0.9297891166	0.545	
Cosmic tumour supp	1337	(0-8)	0.8579148061	0.07905	
UniProt disease	3038	(0-8)	1.123280002	0.09929	
MutClusters	1574	(0-8)	0.2495323392	2.20E-16	**
UniProt neutral	4188	-	-	-	-

Specific examples of mutationally enriched clusters close to functional sites

We identified 42 MutFams using the enrichment method described in materials and methods (section 2.2), which uses mutation data from the PanCancer datasets obtained from COSMIC/TCGA/ICGC [15] [249] [281]. The clustering algorithm detected 175 clusters within the 42 MutFams, comprising 970 mutations. Table A.1 in the appendix summarises the proximity of the clusters to the different types of functional sites. Mutationally enriched clusters in MutFams are described as MutClusters.

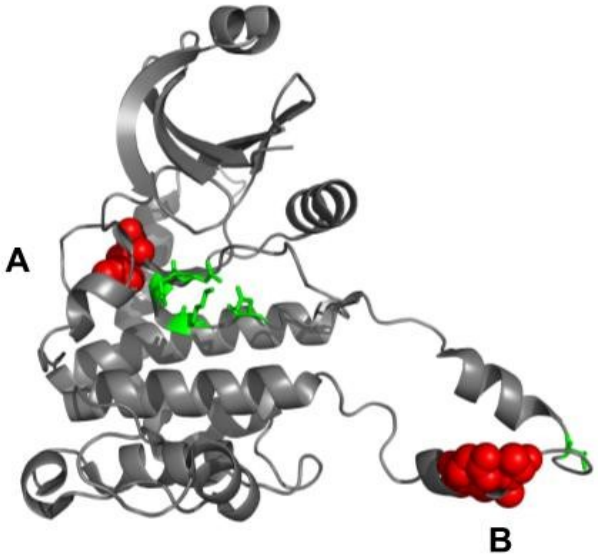
MutCluster mutations in Chk2 kinase

It has been shown that MutCluster mutations show a tendency to locate close to catalytic sites. This is particularly prominent in regions between 3-5Å from the active site residues (figure 2.11). Such positions are likely to be affecting substrate binding or primary shell electrostatics. An example of MutCluster mutations affecting a catalytic residue in this way is shown in figure 2.18. Chk2(checkpoint kinase2) is a driver gene in both low-grade glioma (LGG) and glioblastoma multiforme (GBM) cancers. This is a kinase which helps to maintain genomic integrity during the cell cycle checkpoint and there is increasing evidence that Chk2 plays a tumour suppressive role in cancer [7] where it evokes downstream signalling processes that activate proteins that regulate cell cycle progression, genomic integrity and the activation of proteins resulting in cell death – including P53. It is activated in response to DNA damage. The 4 clusters of mutations within Chk2 kinase are located in 2 major regions in CHK2, known to be important for kinase function. These regions are shown as, a) the ATP binding pocket and b) the activation loop/APE motif in figure 2.18 [117][242]. The location of MutClusters here are consistent with studies showing that both of these regions preferentially harbour cancer causing mutations in protein kinases [55] [105] [156].

Region A in figure 2.18 shows a single cluster centred on residue 355 - spanning the ATP binding pocket and the hinge region connecting the two lobes. This region is close

to a ligand binding site and very close to the catalytic site and has also been implicated in co-ordinating global kinase motions [54] [56], when the protein undergoes conformational changes initiated by substrate binding. The second region, B, encapsulates the clusters centred on positions 392,394,396 – and spans the activation loop of the kinase and the APE motif, both of which are heavily involved in kinase function and activation [242] [117].

Figure 2.18: CHK2 kinase MutCluster Mutations. Transferase (Phosphotransferase) domain 1 is shown from a MutFam domain commonly found in LGG, GBM gliomas, which occurs in the FunFam Calcium/calmodulin-dependent protein kinase type II FunFam (CATH id 1.10.510.10, 79008). The central cluster residues are coloured red and are located in the functional regions labelled A and B. The CSA residues are coloured in green. Coloured boxes in the table below the structure indicate the functional site that the MutClust is close to (less than or equal to 5Å).

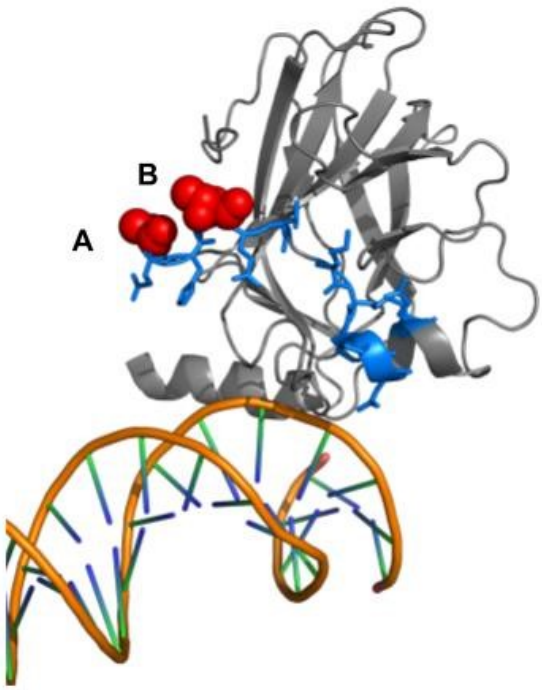


Cluster	Residue	CSA	IBIS_PPI	LIGAND	FunSite
A	A355				
B	A392				
B	A394				
B	A396				

MutCluster mutations in p53

One of the most mutationally enriched MutFams in the COSMIC/TCGA PanCancer dataset contains the cellular tumour antigen p53 domain, within the TP53 gene. Figure 2.19 shows that this domain contains two MutClusters (see A and B) which are near to each other, and within 5Å of an IBIS-PPI site (blue residues), that binds to a homologous transcriptional factor, P63. Therefore, any mutations proximal to the p63 binding site would disrupt its binding, and the tumour suppressive role of p53 in preventing DNA damage and maintaining genomic quality control within the cell cycle [144] [65]. Affecting DNA repair is one of the main hallmarks of cancer and it can lead to further genetic abnormalities within the tumour [92].

Figure 2.19: MutClust mutations in the TP53 domain which is in the FunFam cellular tumour antigen p53 (CATH id 2.60.40.720, 232) found in all cancers. The central cluster residues 195 and 199 are coloured red, and the IBIS-PPI site is coloured blue.



Cluster	Residue	IBIS_PPI	FunSite	Betweenness centrality
A	C195	1	1	0
B	C199	1	1	1

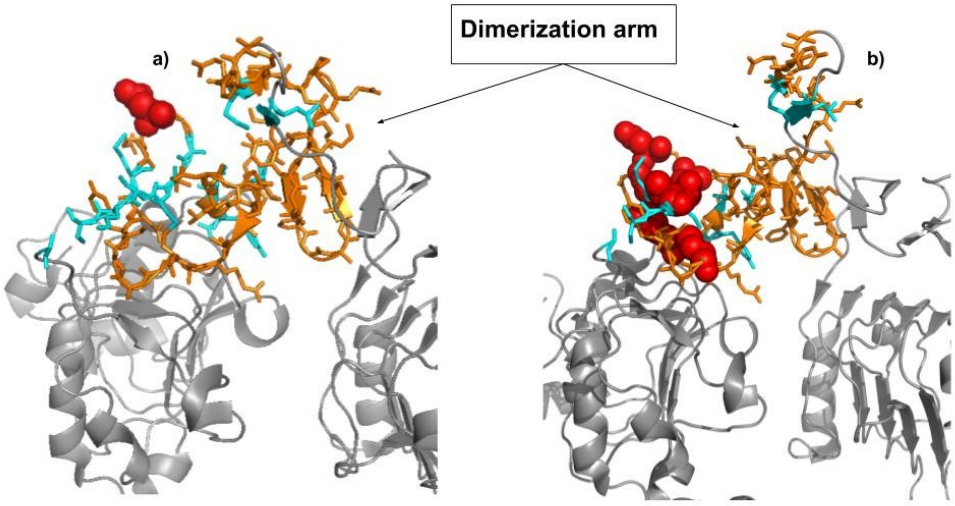
The two clusters are centred on residues 195 and 199. The cluster at position 195 is 5 residues away from a known DNA binding site of P53. Therefore the mutations within the cluster, and possibly those within the second cluster at 199 may have additional functional effects on the P53 interactions to other transcriptional regulators i.e in addition to affecting binding to p63 [34]. The proximity of TP53 mutations to functional sites has also been shown by Sahni *et al* [202]. Other functional sites close to the mutations in the MutClusters include conserved sites (FunSites) and sites possessing a high betweenness centrality i.e containing residues predicted to be involved in allosteric communication within the protein, which lie close to the cluster residue 199.

MutCluster mutations within predicted allosteric residues in EGFR

Epidermal growth factor receptor (EGFR) is a growth factor signalling receptor which governs cellular proliferation and is one of the most well-known cancer causing genes. Several MutClusters were identified within the extracellular region of EGFR, in the L-domain, responsible for receptor dimerization, imperative to its activation.

The binding of the cognate ligand to an EGFR monomer, epidermal growth factor (EGF), induces dimerization of the receptor monomers mediated by the extracellular domains. This dimerization event then evokes a conformation change in the receptor which is transmitted via the transmembrane regions to the intracellular kinase domains. This causes their auto-phosphorylation, leading to subsequent kinase domain activation and the recruitment of signalling adaptor proteins and growth factor signalling [44]. In total, there were 6 MutClusters within this domain, spanning the residues between positions 276-304. According to Figure 2.20, it can be seen that this mutated region lies on top of the L domain within the dimerization arm, which is involved in making disulphide contacts between the EGFR monomers, imperative for receptor dimerization and subsequent receptor activation [44].

Figure 2.20: Mutations in the FunFam Epidermal Growth Factor Receptor domain (1mox structure). The central cluster residues are coloured red. a) the 297 cluster b) and all MutCluster mutations, shown in orange. The surface allosteric sites are shown in cyan. The IBIS-PPI and FunSites were not highlighted in the diagram for clarity. Coloured boxes in the table below the structure indicate the functional sites that the MutCluster is close to (<5 Å).



Residue	IBIS_PPI_	FunSite	Surface Allosteric sites
A287	1	0	1
A293	1	0	1
A295	1	0	1
A297 (a)	0	0	1
A299	1	0	1
A300	1	0	1
A302	1	0	1

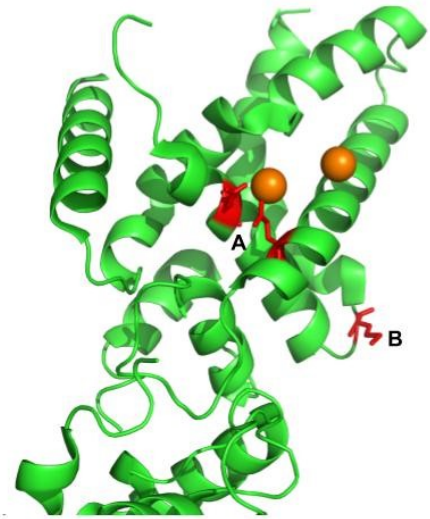
This critical position provides a focal point for altering the orientation of the L domain. The MutCluster locations shown in figure 2.20, suggest a functional effect. Specifically the clusters occur proximal (0-5Å) to surface allosteric sites, which are involved in enabling allosteric communication upon ligand binding, and also to IBIS-PPI sites involved in protein binding. These observations agree with studies by Porta-pardo *et al* [189] which used e-driver to identify significantly mutated regions in whole protein structures. The authors also showed that mutations lay in the dimerization interface, and caused increased EGFR phosporylation by affecting dimerization in GBM [188]. Furthermore, mutations at residue 289, have been shown to increase mean intracellular phosphoprotein levels, compared to wild type EGFR [129]. The activating nature of these mutations

are consistent with studies showing that glioblastoma tumour cells, harbouring these mutations, display a carcinogenic phenotype, of anchorage independent growth [129].

MutCluster mutations in metal binding and predicted allosteric sites in DICER RNAase

Figure 2.21 shows an example of a cluster within the RNAase domain in the DICER gene, located close to a magnesium binding residue. DICER uses magnesium ions to bind RNA and catalyses RNA cleavage to produce small interfering RNA and MicroRNA molecules. According to figure 2.21, it can be seen that all MutCluster residues in clusters A and B (highlighted in red in figure 2.21) are within 5 Å to conserved FunSites, and predicted allosteric residues of high betweenness centrality. Both MutClusters contain at least 1 residue within 5 Å to a UniProt metal binding site, which binds magnesium. This observation is consistent with studies showing that many somatic mutations in DICER affect its metal binding activities within ovarian tumours [96]. It has been suggested that this effect contributes to the oncogenic phenotype by altering the production of micro RNAs (miRNAs), which are important for cell differentiation and fate determination [96]). Germline non-cancer mutations and mutations from a range of disease including cancers, have shown enrichment at metal binding sites reported by studies performed by Martinez *et al* [108]

Figure 2.21: Mutations within the RNAase DICER domain (PDB code 2eb1) in the endoribonuclease Dicer homolog 1 domain FunFam (CATH id 1.10.1520.10, 4026) found in all cancers. The central cluster residues are coloured red, and the magnesium ions are coloured as orange spheres. Coloured boxes in the table below the structure indicate the functional site that the MutCluster is close to ($<$5 Å).



Cluster	CLUSTER_RES	FunSites	UniProt Feature - Metal binding	Betweeness- centrality
A	B152	1	1	1
A	B47	1	0	1
B	B167	1	1	1

MutCluster mutations in DYRK kinase with apparently no functional effect

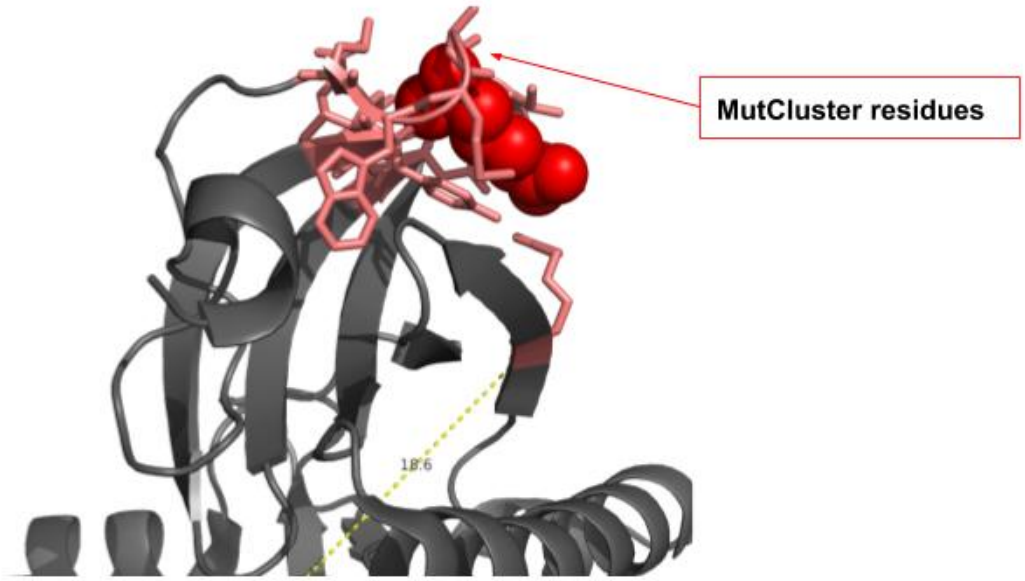
We also examined a MutCluster within the DYRK-1A kinase, involved in regulating DNA splicing and apoptosis [223]. Both processes have been implicated in cancer, where the former has recently been linked to protein network rewiring. DYRK-1A showed no MutCluster residues within 5Å to any of the functional sites considered in the previous studies, i.e CSA, IBIS-PPI, ligand binding, allosteric, or predicted functional sites (FunSites).

One possible explanation for this would be that no FunSites were predicted for the DYRK-1A structure. One of the main determinants for producing the FunSites requires the structure to be a member of a FunFam with an appropriate DOPS score ($\geq$ 70). Lower DOPS scores imply less informative multiple sequence alignments involving less diverse sequences, and therefore less reliable conserved site identification. The DOPS score for the the MutFam containing DYRK-1A was lower than 70 (at 68.45) and so the conservation sites for this MutFam could not be safely identified.

To explore the consequences of this mutation further, FunSites were inherited from the closest FunFam with a DOPS score $>$70, specifically a predecesor of the DYRK-1A FunFam in the FunFam tree hierachy. These FunSites may encapsulate some of the functional residues within DYRK-1A. The MutCluster residues and inherited FunSites were mapped to the DYRK-1A structure, and show a high degree of co-location with the MutCluster residues, illustrated in the figure 2.22.

Furthermore, literature searches revealed that the inherited FunSite residues are conserved between different isoforms of DYRK kinases in different species. In a related isoform - DYRK2 kinase - these inherited FunSites facilitate autophosphorylation of the activation loop, and are referred to as NAPA-1 and DH-box regions. Although the DYRK-1A structure lacks such regions, this suggests that mutations in this region of DYRK-1A mimic the enhancing effects on autophosphorylation, seen in relatives, to increase kinase activity [223].

Figure 2.22: Mutations in the DYRK-1A kinase (structure 2vx3) in the Dual-specificity tyrosine-phosphorylation-regulated kinase 1B domain FunFam (CATH id 3.30.200.20, 64610). The central cluster residues are coloured red.



Conclusion

The functional impacts of different disease mutations have been explored, by analysing the proximities of germline non-cancer, cancer, and putative cancer drivers within MutClusters to known and predicted functional sites. In terms of the specific trends of proximity reported from the MutDist analysis, this revealed that the different disease mutations show some enrichment at specific functional sites, highlighting possible mechanisms of pathogenicity, which are summarised in table 2.12. These patterns are consistent with the literature, suggesting that cancer mutations are more likely to affect protein-protein interaction sites than germline non cancer mutations [169][271], and that germline non cancer mutations show more enrichment at betweenness centrality sites, which are buried within the protein core [83].

Table 2.12: Proximity trends of different disease mutations to the various functional sites. A significant enrichment within 0-8Å of the functional site is where P-value is less than or equal to 0.01. All P-values are based on the enrichment of mutations at the functional sites highlighted in bold, apart from the BC-site P-value for a mutation depletion indicated by an asterisk (*). N/A indicates sparse data.

Mutation type	**CSA**	**IBIS-PPI**	**Ligand**	**FunSite**	**MOD_RES**	**Metal**	**BC-sites**
Germline non cancer	**2.2E-16**	0.201	**5.28E-07**	**2.2E-16**	**2.0E-09**	**8.88E-07**	**2.2E-16***
Cosmic oncogenes	0.053	**3.53E-13**	**9.39E-08**	**3.54E-11**	**6.42E-10**	N/A	0.545
Cosmic tumour supp	**0.0044**	**4.65E -15**	0.23	N/A	**0.0092**	**0.00046**	0.08
UniProt disease	**2.2E-16**	**2.2E-16**	**2.2E-16**	**3.36E-05**	**2.9E-05**	**1.66E-10**	0.099
MutClusters	**3.34E-15**	**2.2E-16**	**2.2E-16**	**2.2E-16**	**0.00012**	**1.05E-15**	2.2E-16*

According to table 2.12, it can be seen that cosmic oncogene and tumour suppressor gene mutations differ in their proximities to some sites, where oncogenes are more enriched at ligand binding sites, and the tumour suppressor gene mutations are more slightly enriched at CSA sites. The predicted driver mutations within the MutClusters, show even greater enrichment at such sites. This study also demonstrates the utility of the CATH FunFams protocols and structural mapping to identify putative cancer driver

mutations in selecting more functionally relevant mutations compared to analysing a mixture of passenger and driver cancer mutations. The predicted driver mutations show the highest enrichment at various functional sites compared to all other mutation types considered, as summarised in table 2.12.

An additional dataset was included in this analysis, comprising of the disease and neutral mutations from UniProt. These underwent filtering to avoid the bias of heavily mutated genes, according to the approach used in Gao *et al* [75]. Further work would include analysing the excluded genes on their specific functions. Since the study of proximity to post-translational modifications (PTMs) treated all PTMs as one type of site, as more data becomes available, more detailed analyses of each site type would also provide a valuable focus for future study.

3D mutation clustering can detect mutations distributed throughout the protein sequence, and is therefore more likely to identify rare driver mutations than 1D based hotspot methods. In addition to identifying MutClusters, the use of CATH FunFams has extended functional site information by predicting functional sites - FunSites, shown to be highly enriched for all types of disease mutations in germline non cancer, and cancer variants. For those genes mapping to a FunFam with low information content, it is possible to infer functional sites by mapping to a closesly related FunFam, by using the FunFam hierarchy structure, referred to as FunTree [73].

Clearly proteins can have different numbers of functional sites and different numbers of residues within them, and this can have an effect on the proximity analysis. For proteins that have multiple sites, there will be a greater likelihood of mutations lying near functional sites, by chance. The main purpose of this analysis is to show that the cancer mutation residues, which cluster within 3D, have a greater tendency to lie near functional sites. Here the same genes are being used for comparison. That is, the bias has been addressed by performing our analysis of proximity to functional sites, also for neutral mutations observed in the same proteins. Our results demonstrate that disease mutations, especially cancer mutations filtered by clustering (MutClusters), are statistically significantly more likely to be close to functional sites than neutral mutations.It is possible that

proteins in the other datasets eg proteins with OMIM germline mutations, do not have as many functional site residues in them as the proteins with cancer mutations. This possible bias could be explored in future analyses. Other future work would be to include the study of 3D clusters within other disease mutation types and also within neutral mutation datasets. It would also be of interest to analyse the structural and chemical properties of the clusters themselves such as, such as cluster size, polarity or hydrophobicity.

In our study, we developed a new method for cancer cluster detection, to see whether cancer mutations in the MutFam genes are clustered in 3D, and whether these 3D clusters are located on or close to known and predicted functional sites. In the future, it would be interesting to compare the cancer mutation 3D clusters (MutClusters) identified by our approach with those identified by other algorithms which use structure for driver gene detection, for example CLUMPS [110], e-driver [189], and HotMAPS [247].

Chapter 3

The Use Of CATH Functional Families In Cancer Mutation Analysis

Introduction

In this chapter, putative driver genes in cancer were identified and subjected to functional analyses. Below is a literature review of this field, followed by a summary of the work performed in this chapter.

Chapter 2 described the identification of CATH domain families enriched in cancer associated mutations (MutFams). Therefore, human genes within these MutFams can be considered as putative driver genes containing a common mutated functional domain. Our putative driver genes were compared with driver genes predicted by another domain-based method, based on Pfam [154] [271]

A gene ontology (GO) biological process analysis was performed on the common genes and for the each of the non-overlapping driver genes within the MutFam and Pfam based methods, to see whether the different predicted driver genes converged on their enriched biological processes, thereby suggesting shared mechanisms of pathogenicity. In order to assess the clinical relevance of these driver genes within the cell itself, the GO biological processes for each driver gene set were compared to the literature to infer their relevance in cancer. Further validation of the MutFam driver genes was done by comparing their ACSN hallmarks to those associated with known cancer genes from the Cancer Genome Census (CGC) [257]. In addition, functional analyses were performed on MutFams within the specific cancers; LGG and GBM. These included pathway and GO biological process enrichment studies.

MutFams

As described in chapter 2, MutFams are significantly mutated CATH FunFam domains, where a FunFam is a functionally coherent set of protein domain sequences. Relatives

within a FunFam are grouped based on structural and sequence similarity and are associated with at least one experimentally characterised GO functional term [41]. Since FunFams are functionally coherent domains families, this approach offers greater accuracy in analysing the functional impacts of mutations compared to previous domain based approaches which used less functionally specific domain families e.g Pfam [154] [182].

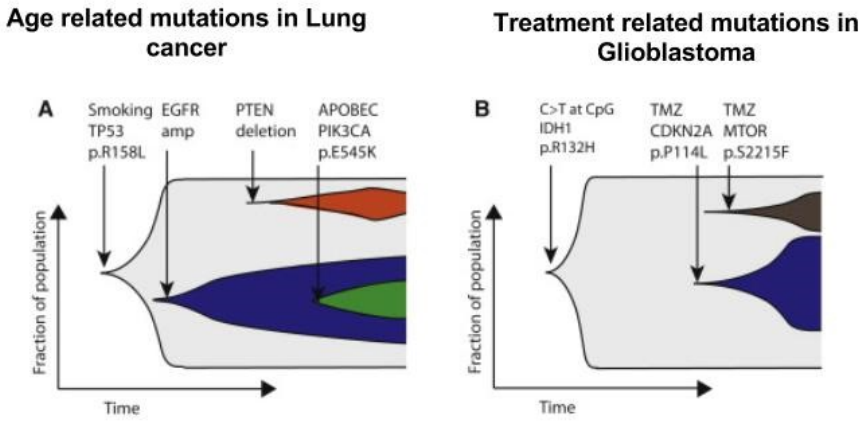
There are two advantages of this increased functional resolution, the first is more accurate identification of driver mutations within protein domains, as mutated positions are aggregated from each member of the FunFam multiple sequence alignment on to a FunFam representative structure and clustered in 3D. This thereby increases statistical power and filters out noise due to passenger mutations. In addition to this, MutFams can aid driver gene detection, since membership of a MutFam implies a common, significantly mutated functional domain. We hypothesised that identifying driver genes in this manner may produce a more refined set of driver genes, specific to a set of cellular pathways, and cancer hallmarks from the ACSN online tool [120] than using functionally coarser Pfam domain families.

Cancer – a heterogeneous disease

Cancer is an evolutionary disease, where the various genetic abnormalities act in concert to promote net growth, while also providing a genetic buffer for adapting to selection pressures. Within a given cancer, there are regions of the tumour (subclones) that have distinct mutational profiles in specific pathways. These sub-clones can be geographically distinct within the tumour, and can also emerge at different times. This is referred to a spatio-temporal heterogeneity, which makes therapeutic intervention a great challenge. It is known that these subclones can act together to reach functional complementarity, where mutated pathways support the growth of new clones, which in turn shape the evolutionary trajectory of the tumour [148]. For example, this sub-clonality can be seen for a lung tumour where an initial mutation occurring within P53 causes the outgrowth of a major clone containing this mutation (grey zone in figure 3.1 1a). Within this major

clone, later mutational events including; EGFR amplification (blue zone), PTEN deletion (red zone), and APOBEC and PIK3CA mutations (green zone) cause the emergence of other sub-clones, further complicating the mutational profile of the tumour over time.

Figure 3.1: Modes of cancer evolution. A) Age related lung tumour, containing different mutational events that lead to emergence of subclones. B) Glioblastoma tumour evolution before and after treatment with Temozolomide (TMZ). Different colours represent the different clones and their effected pathways. Taken from [148].



In addition to age related mutational processes, tumour subclonality can also be invoked by anti-cancer treatments, such as Temozolomide (TMZ), shown in figure 3.1. TMZ has been used in chemotherapy for treating glioblastoma multiforme, which is one of the most common and aggressive adult brain tumours [280]. It can be seen that even though this drug acts to prevent tumour recurrence, it also causes mutations within the genes; MTOR and CDKN2A. This in turn initiates 2 separate tumour subclones, further shaping the evolutionary trajectory of the tumour. Therefore, analysing pathways within cancer cells can help make sense of this mutational diversity, whereby different mutations in the same gene or differences in mutated genes, may be affecting the same pathway. By mapping genes to pathways, we can highlight the main pathways effected within the tumour as a whole, or that are under positive selection in cancer, thereby providing a therapeutic opportunity.

Cancer hallmarks

Cancer is one of the most complex diseases to date, where cancers differ in their affected components and pathways. Elucidating and identifying the different processes affected can be a challenging endeavour, due to genomic instability and the heterogeneous nature of cancer. However, it has been suggested that the various effects converge into 11 main cellular hallmarks [92]. These have been used as a means of classifying the diverse effects of the neoplastic disease, and a summary of these is shown in figure 3.2. The hallmarks range from interfering with genome stability, growth factor signalling, invasion and metastasis, cell cycle and apoptosis regulation, and altering cellular energetics such as metabolism.

It has been proposed that there are 2 general categories of hallmarks within cancers, the core and the emerging hallmarks [92]. The core hallmarks are ones which are affected in nearly all cancers, such as impacting genome stability and the endowment of genomic errors which alter growth factor signalling and resistance to cell death. Another core hallmark is the promotion of inflammation, leading to an infiltration of immune cells. These core hallmarks alone can result in a carcinogenic phenotype of abnormal cell proliferation and are present in both early and late cancers. In order to sustain this level of cell growth, the emerging hallmarks play an important role. It is within these hallmarks that different cancers have adopted distinct mechanisms to enable sustained proliferation. These emerging hallmarks include those affecting energetics and carbon-based metabolism to increase ATP production for the increased energy demand.

Other emerging hallmarks include those which exploit the immune response, whereby immune cells are incorporated into the tumour environment and provide support for tumour growth. Here it is thought that immune cells provide a defence mechanism for the tumour against other immune factors and also against external drugs [92]. These emerging hallmarks, particularly the latter, are mainly characteristic of later stage cancers, which have already established a sustained level of abnormal cellular growth.

Other emerging hallmarks, associated with later stage cancers include promoting angiogenesis to increase oxygen uptake, especially in response to hypoxic conditions [178].

Figure 3.2: The 11 cancer hallmarks (blue writing), Taken from [92].

In order to metastasise, tumour cells leave the primary site and colonise elsewhere. This is initiated by invading the extra-cellular matrix (ECM) and surrounding tissues. This can involve many different processes, such as affecting cellular adhesion to other cells and the underlying ECM components, degrading of the ECM, and changing cell morphology to be more conducive to migration. Processes encapsulating such invasive events are within the “activating invasion” and “metastasis” hallmarks. In order to figure out which of the hallmarks and corresponding pathways are being affected by a set of genes implicated in a given cancer, various studies have exploited online resources to determine and characterise pathway enrichments [125].

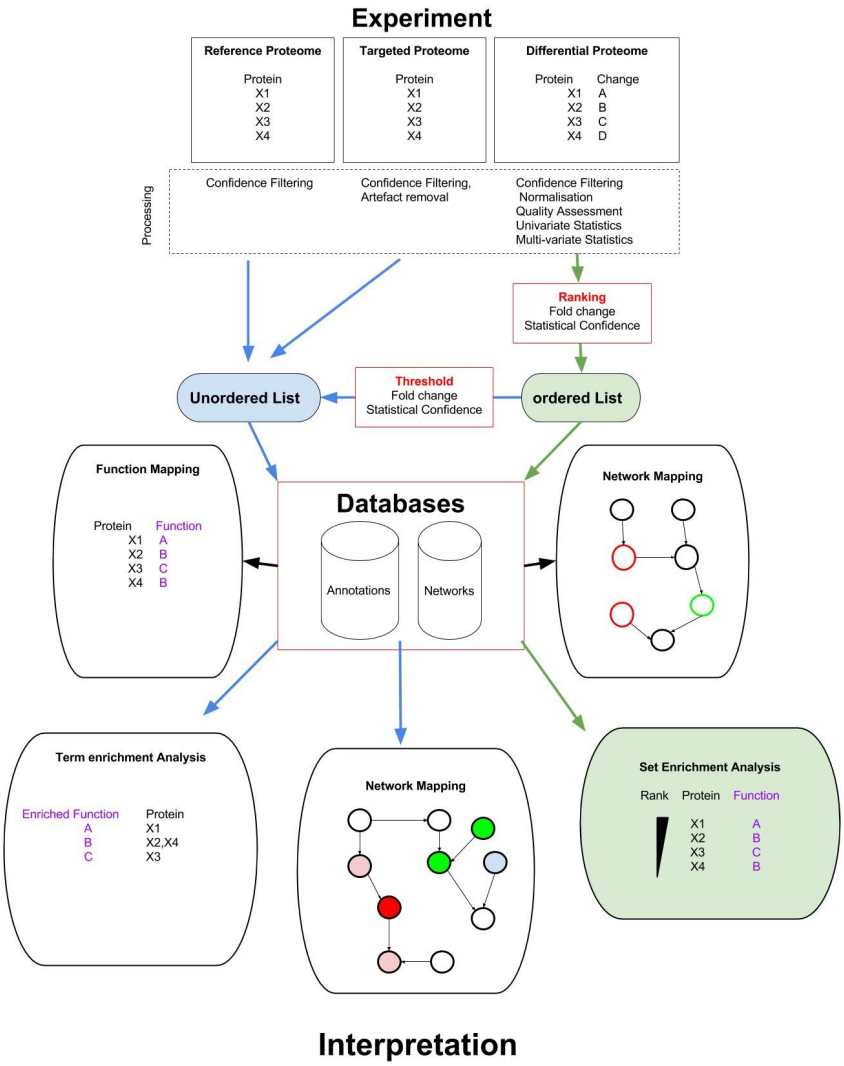
The complex nature of cancer, caused by its genomic instability, leads to a plethora of mutated components in multiple pathways. In addition to mutations which confer a growth advantage to the cancer, referred to as driver mutations, there are also mutations that happen as a result of genomic instability and do not contribute to a carcinogenic phenotype. The latter are referred to as passenger mutations. When considering the

mutation data in its entirety, there is a lot of noise due to passenger mutations, and it can be hard to prioritise which mutational events contribute most to carcinogenesis. Therefore, in order to prioritise proteins for study, various efforts have used mutation enrichment methods, to detect genes that are significantly mutated and likely to be drivers. The various methods for detecting driver cancer genes have been discussed in chapter 2.

Assessing the functionality of candidate genes

To make sense of driver gene lists on a cellular level, functional analyses are needed. There are many ways of doing this, some of which are summarised in ([125]) and illustrated in figure 3.3.

Figure 3.3: Overview of gene list analysis and enrichments of functions. Edited from [125]



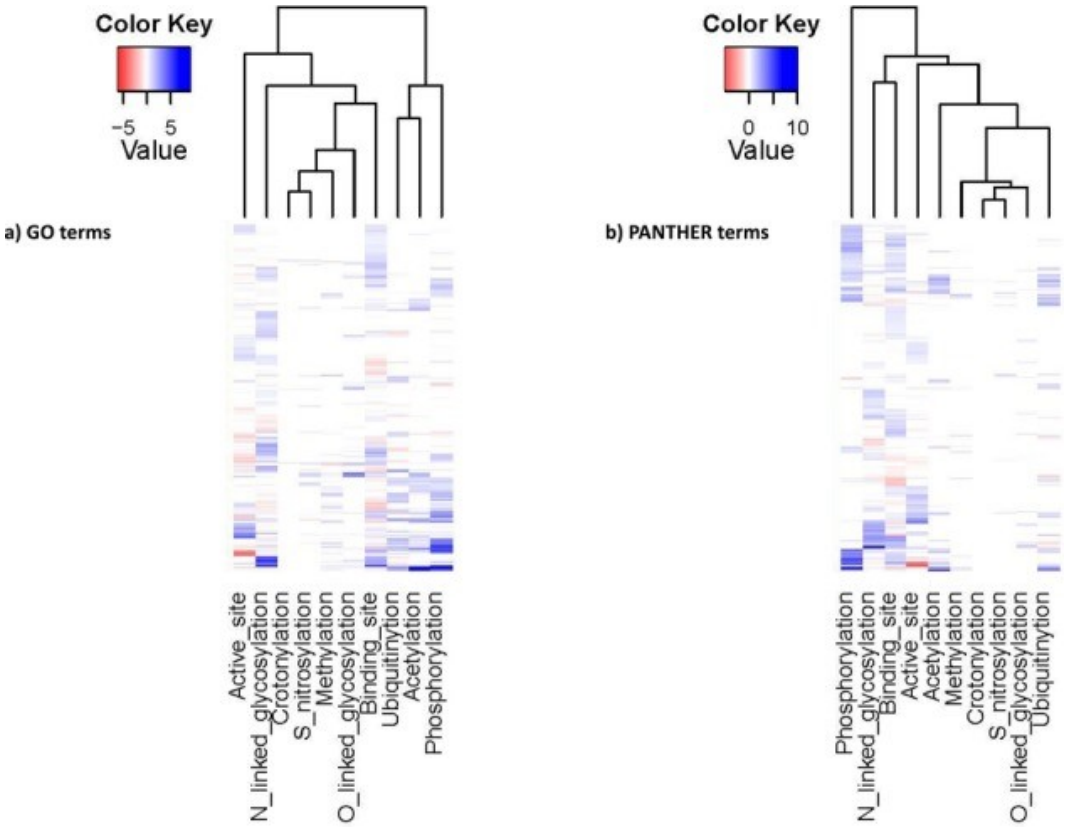
Analysing functional enrichment in putative gene driver lists

Pan *et al* [175] analysed somatic pan-cancer mutations from TCGA, in 127 significantly mutated genes, annotated with known UniProt functional sites [11]. The aim of this study was to detect whether functional sites were affected in specific pathways from GO [13] and PANTHER [153]. This study identified 10 main types of functional sites affected.

In order to see whether the mutated functional sites were associated with distinct cellular events, the gene lists were subjected to GO term and PANTHER pathway enrichment tests, and the mutation events affecting functional sites were then clustered based on the similarity of enrichments in GO and PANTHER pathways. A P-value was determined which reflected the enrichment of the pathway in genes with mutations on functional sites compared to genes with mutations not located on functional sites. P-values were calculated for each type of functional site. This resulted in a heat map of enriched GO biological processes and enriched pathways, shown in figure 3.4.

The genes identified as having the most mutations were TP53, HIST1H4A, HIST1H3A, RELN, SMAD4, CTNN81, DICER1, KRAS, NRAS, BRCA2 and PTEN. This gene list was enriched in pathways mainly involved in cell-cell adhesion, the nervous system, and embryonic development. Genes with mutations in ubiquitylation and acetylation sites were associated with similar PANTHER pathway terms. The data also showed mutations in phosphorylation sites to be enriched in many PANTHER pathways, consistent with the impacts on growth factor kinase signalling processes, often affected in cancer [169].

Figure 3.4: Clustering of mutated UniProt functional sites using a) GO term and b) PANTHER pathway enrichments. Unique terms are along the vertical, under and over representation are coloured red and blue respectively. The columns show the association of functional site mutations with a particular GO term. Taken from [175].



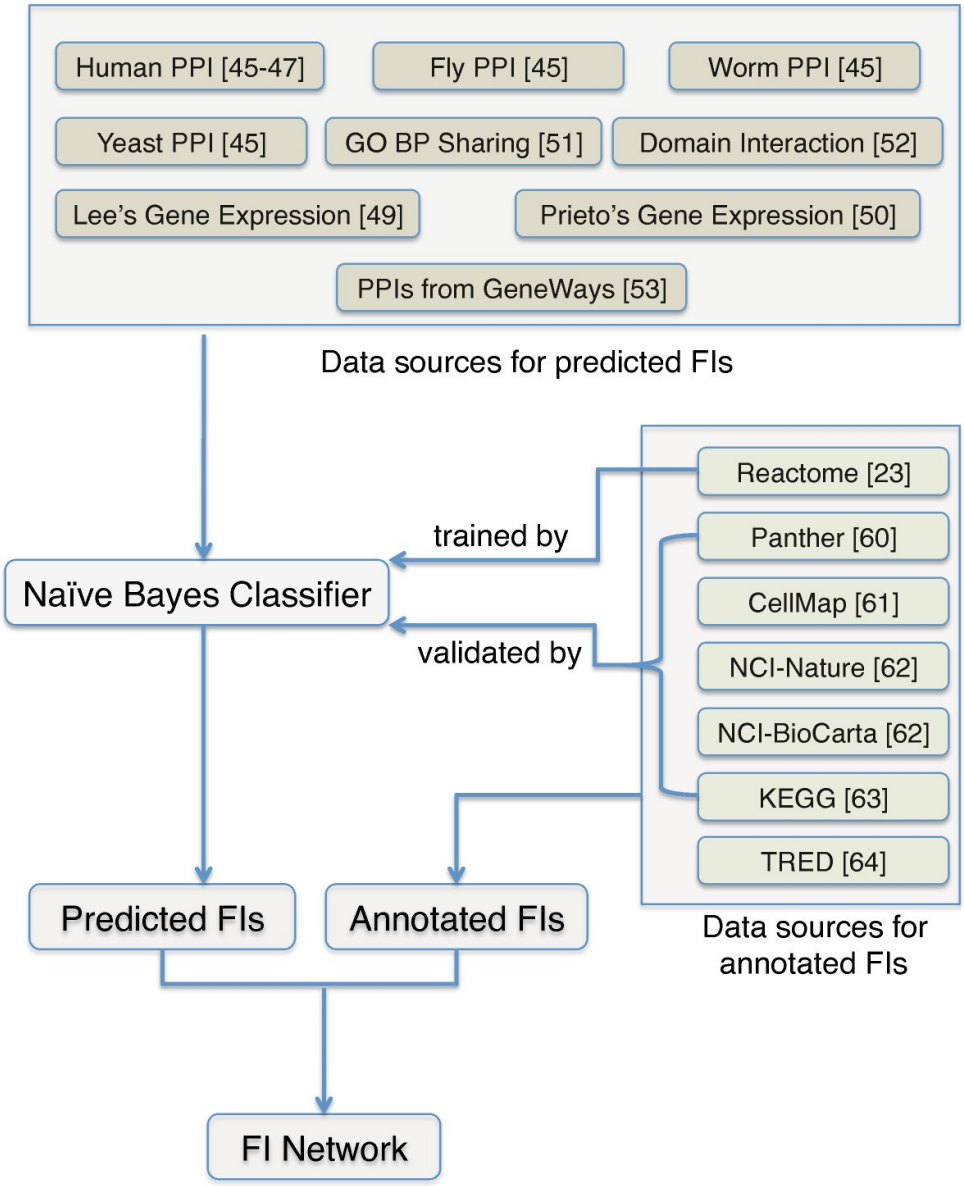
Analysing enrichment of putative driver genes in protein networks

Various studies have also analysed gene functionality by mapping genes onto protein networks, which describe the functional relationships of the genes to one another. This enables a more comprehensive representation of a gene list, to identify sets of gene nodes which are functionally coherent. Wu *et al* [269] constructed a functional interaction (FI) network based on data from protein-protein interactions, as described in figure 3.5.

This comprised 10,956 proteins and included functional relationships predicted using a trained naïve Bayes classifier. It was used to analyse cancer genes associated with Glioblastoma Multiforme (GBM), taken from the TCGA, and occuring in 2 or more samples. The majority of cancer causing genes from GBM were found to be closer in

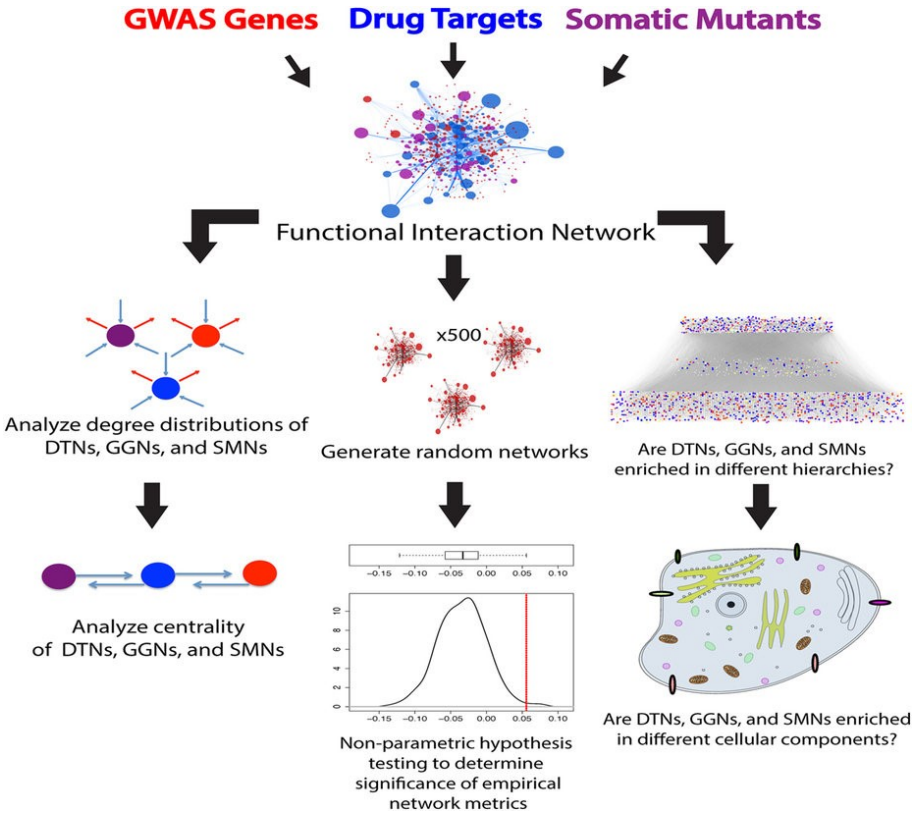
the network than expected by chance, forming clusters of functionally related genes. Enrichment studies revealed two main cellular locations of the cytoplasm and plasma membrane. Many pathways were enriched within these clusters, four of which had an FDR $\leq$ 8.0x10-4, including pathways involved in focal adhesion, signalling by PDGF and p53, and cell cycle regulation.

Figure 3.5: Functional Interaction network of genes, taken from [269].



Studies by Ung *et al* [146] used gene networks to study the topological relationships between three gene groups; genes containing germ-line variants from GWAS studies (GGNs), genes containing cancer somatic mutations (SMNs), and genes which were known drug targets(DTNs). The analysis pipeline used in this study is shown in figure 3.6. To elucidate any differences in the cellular components between the three classes, gene lists were analysed based on their GO cellular component annotations, using the DAVID web tool [49]. Gene centrality measures were also calculated to determine the control, closeness and betweenness scores.

Figure 3.6: Flowchart analysis of three gene sets using gene network measures. GWAS studies (GGNs), genes containing cancer somatic mutations (SMNs), and genes which were known drug targets (DTNs). Left pipeline:Network analysis was performed to identify topological relationships between the genes, middle pipeline:random networks generated to measure statistics, right pipeline:hierarchical analysis was performed between node classes, which were then subject to cellular component analysis. Taken from [146].



Control centrality reflects the ability of a node to control other nodes in a directed weighted network, which is arranged hierarchically. A high control centrality node controls many other downstream nodes (e.g a signalling receptor at the plasma membrane). Closeness centrality measured the connectivity of a gene to other genes in a network, whereby genes of high closeness are referred to as “hubs” and are often crucial to many cellular processes. Betweenness centrality identifies genes which are between subnetwork communities, and therefore play a role in communications between biochemical pathways. All of the network centrality measures were then compared to a random null distribution.

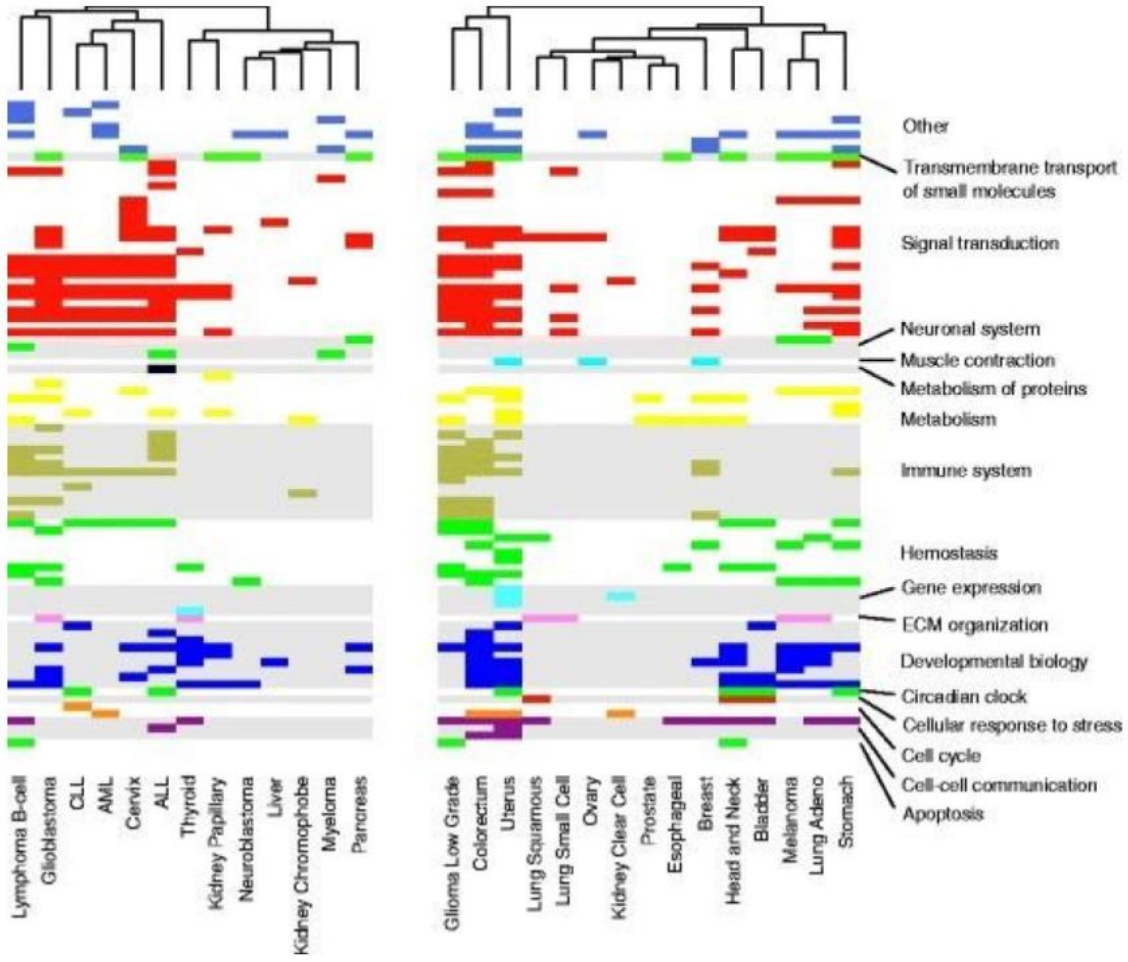
This analysis showed that the drug target genes (DTNs) have the highest control and closeness centrality, and exert greater control over the entire gene network by acting as “hubs” and controlling downstream genes. This was consistent with their cellular location being primarily on the cell surface, where the majority of drug targets are receptors. The somatic and GWAS variants both showed a high betweenness centrality, suggesting that they act to communicate signals across biological pathways. This was further supported by both gene sets being enriched in the nucleoplasm and cytosol cellular locations.

Other methods which have used a functional network to analyse candidate genes include the protocol of Vihinen *et al* [168], who analysed mutation datasets including those from 30 different cancers, COSMIC [15], ClinVar [122], and from the Database of Curated Mutations (DoCM) [8]. The mutations were mapped to proteins and analysed on their pathogenicity scores using PON-P2 [167], to prioritise harmful mutations. Candidate genes were identified based on frequencies of pathogenic mutations, which were normalised to the corresponding gene length. Proteins with the top 5% of harmful mutation ratios were analysed using the network developed by Wu *et al* [269], for network edge centrality and pathway enrichments, using the ReactomeFVIZ tool. The enriched pathways for each cancer are shown in figure 3.7.

The authors found that proteins containing harmful mutations were centrally located in protein interaction networks. Enriched pathways included those involved in the cell cycle, apoptosis, and growth factor signalling by EGFR, MAPK, MTOR and PI3K. All of

these are consistent with hallmarks of cancer [92]. Other pathways enriched included those involved in developmental signalling, such as WNT and NOTCH pathways, particularly for specific cancers, such as those in the colorectum and uterus.

Figure 3.7: Summary of the enriched pathways for each cancer. Each row represents a pathway, where each colour represents the different parent pathways within the reactome hierarchy. Taken from [92].

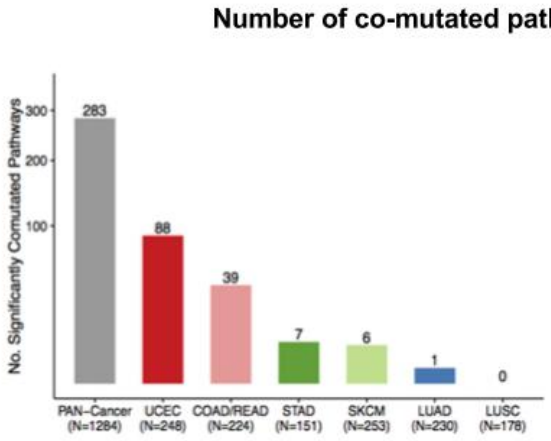


Co mutated pathway network analysis

More recently, Jiang *et al* [245] analysed the tendencies of pathways to be co-mutated within cancers. Mutation data were taken from 1284 tumour samples from a range of cancers in The Cancer Genome Atlas (TCGA). Namely, uterine corpus endometrial carcinoma, colon/rectum adenocarcinoma, stomach, lung and skin carcinomas. Each so-

matic mutation was predicted as deleterious, either by the predictor CADD [145] or if they were insertions or deletions. Pathways were taken from the EnrichmentMap pathway repository [151]. For any pair of pathways, the mutated frequencies were subject to bayesian conditional probability modelling. The significance of pathway pair co-mutation was compared to a simulated random model, using a one sided fisher's test. For the random permutation model, a subset of the EnrichmentMap repository was used which consisted of 217 pathways. The number of co-mutated pathways in each cancer and the combined pan-cancer dataset is shown in figure 3.8.

Figure 3.8: The number of co-mutated pathways in each cancer mutation dataset. Taken from [245]. N=number of mutations. The cancer names and abbreviations are in the table under the graphs.



Cancer	Description
UCEC	Uterine Corpus Endometrial Carcinoma
COAD/READ	Colon/ Rectum adenocarcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
STAD	Stomach adenocarcinoma
SKCM	Skin Cutaneous Melanoma

Analysis of the pan-cancer data from all 6 cancers, showed that pathways involved in IP3 metabolism and PI3K were significantly co-mutated, the most heavily mutated gene being PTEN. Other gene mutations were detected, where different gene pairs were co-mutated in different patients. Both IP3 and PI3K pathways have been shown to increase the second messenger PIP3, leading to activation of downstream signalling to the cell survival pathways involving MTOR and Akt. Interestingly, it was also found that the number of significantly co-mutated pathways did not correlate with the size of the mutation dataset for a given cancer.

In order to determine the co-mutated hallmarks, the Molecular Signature database was used [133]. A pair-wise enrichment score was computed for the pan-cancer mutations, by normalising the number of significant co-mutated pairs of pathways by the total number of pathways in each hallmark category. Consistent with their pathway analysis, the main hallmarks were those involved in signalling and cellular component categories, and effects related to DNA damage and proliferation were less frequent. In terms of the mutations in colorectal cancer, it was found that co-mutated pairs of pathways were enriched for cellular hallmarks within the proteosome and apoptosis pathways. Another interesting observation from the study was the lack of co-mutated pathways within lung cancer.

Comparing cancer and non-cancer gene sets using network analysis

As well as analysing genes implicated in cancer, other studies have also included other disease genes, and considered their cellular affects. Studies by Pinero *et al* [87] performed a systematic analysis of 4 different disease gene sets by analysing their network properties. Cancer and complex disease genes were extracted from the DisGeNET database based on the phenotypes of neoplastic process, congenital abnormalities, and mental or behavioural dysfunction. Other cancer genes were taken from the cancer genome census (CGC) [15]. The mendelian disease genes were retrieved from OMIM [10].

For each of these gene sets, 10,000 randomly selected samples of gene sets were extracted of the same size as the gene sets being studied, as a control. For each gene mutated in disease, different non-disease mutations from EXAC were also considered; non synonymous and synonymous mutations, stop and start site mutations, frameshift and splice site mutations. Mutations in the disease genes were predicted deleterious according to CADD. In order to assess the tolerance of genes to deleterious mutations, a High-impact to Synonymous ratio (HS) was calculated for each gene, based on the ratio of deleterious variants and the non-deleterious variants to the synonymous mutations.

To construct the protein network, experimentally derived interactions were retrieved

from 5 major resources; HIPPIE [102], BIANA [77] , BioGRID [224], IntAct [114], IrefIndex, andS [159]. In addition, 2 further experimental protein interaction datasets were included – from a yeast two hybrid assays, and an affinity capture using mass spectrometry (ACMS) study. To make the gene network, Pinero *et al* [87] partitioned the protein interaction networks using community detection. This analysis showed that all disease genes possessed a higher degree and betweenness centrality compared to the random control. Specifically, cancer genes showed the highest scores, reflecting their more central role in the network and their role in bridging network communities. It was also shown that for each of the cancer and complex disease gene sets, their tolerance to germ line disease variants was inversely correlated to their centrality within a protein network. Examples of intolerant genes, include VHL, PIK3CA, TP53, and proteins related to the RAS family.

In this chapter, driver gene lists were identified in the MutFams described in chapter 2 and compared to driver gene lists identified by Miller *et al* [154] and Yang *et al* [271] using a Pfam based approach. Analysis was performed for enrichments in Gene Ontology (GO) annotations in the driver gene lists, using a well-known network based method developed by [269]. GO Biological process enrichments demonstrated that the unique MutFam and Pfam derived driver genes by Miller *et al* [154] and Yang *et al* [271] affect different steps within common biological pathways, which are implicated in cancer. Unique pathways identified for the MutFam and Pfam driver genes could be explained by tissue bias in the cancers analysed. A cancer hallmark analysis, using ACSN, found that the MutFam and Pfam driver genes were enriched in survival and EMT cell motility hallmarks.

In order to see if the MutFam driver genes captured specific functional events within different stages of glioma, the MutFam driver genes for early stage low grade glioma (LGG) and more advanced Glioblastoma multiforme (GBM) were analysed for their GO and pathway enrichments and compared to their phenotypes, as reported in the literature.

Materials and methods

Identifying putative cancer driver genes

The MutFam protocol, described in chapter one, identified significantly mutated FunFam domain families. In order to assess the value of these MutFam functional families in understanding disease mechanisms associated with cancer, we contrasted the genes in the MutFams with known cancer genes in the Cancer Gene Census [15], and with predicted cancer driver genes identified using alternative approaches based on Pfam families.

For each MutFam, putative human driver gene sets were derived by taking the top 25% of mutated genes in the MutFams for the 22 cancers in table 3.1. For a control, neutral mutations were taken from the UniProt HUMSAVAR resource [11], where proteins having “polymorphism” mutations were selected.

Driver genes obtained from the literature

In order to compare the driver genes identified in MutFams with those identified by other groups, we compared our genes with predicted driver genes reported in the study of Miller *et al* [154] and Yang *et al* [271]. The Miller driver genes were identified using a Pfam based protocol (see chapter 2, section 1.1.2.2) and obtained from the MutationAligner website [79] which provides information on all mutationally enriched Pfam domains. Significantly mutated Pfam domain families were selected based on a Bonferroni corrected P value ≤ 0.05 and the top 2 genes mutated were used for subsequent analysis. This gave a set of 271 genes -referred to here as Miller genes. We also compared our predicted driver genes to those identified by Yang *et al* [271], based on a method that also looked for mutationally enriched Pfam domains, and then used for subsequent analysis. Yang driver genes were those which only contained mutations which had a “functional” effect according to the IntOgen tool predictor. These are referred to as Yang genes, of which there was 94, which were extracted from the supplementary material given in Yang *et al* [271].

Table 3.1: Cancers included in the MutFam driver gene set, where the number of MutFams and samples are shown.

Code	Description	Num. Samples	Num. MutFams
BLCA	Bladder cancer	364	17
BRCA	Breast invasive carcinoma	1142	13
COAD	Colon adenocarcinoma	371	36
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	161	4
ESCA	Esophageal carcinoma	286	4
GBM	Glioblastoma multiforme	492	18
GLI	Gliomas	870	27
KIRC	Kidney renal clear cell carcinoma	603	18
LAML	Acute Myeloid Leukemia	276	18
LGG	Low grade gliomas	378	7
LIHC	Liver hepatocellular carcinoma	891	15
LUAD	Lung adenocarcinoma	606	29
LUSC	Lung squamous cell carcinoma	287	18
OV	Ovarian serous cystadenocarcinoma	516	6
PAAD	Pancreatic adenocarcinoma	667	7
PRAD	Prostate adenocarcinoma	244	4
READ	Rectum adenocarcinoma	145	13
SKCM	Skin Cutaneous Melanoma	633	87
STAD	Stomach adenocarcinoma	319	15
THCA	Thyroid carcinoma	428	11
UCEC	Uterine Corpus Endometrial Carcinoma	248	36
UCS	Uterine Carcinosarcoma	23	2

Known cancer genes from the Cancer Genome Census (CGC)

The known driver genes were those given in the Cancer Gene Census from Wellcome-Sanger [15] [257]. We selected CGC genes that harbour missense mutations (232 genes).

Hallmark enrichments

To analyse enrichment within hallmark processes, we used hallmark information taken from the Atlas of cancer signalling networks (ACSN), which have been curated based on studies reported in the literature (Hanahan2011). ACSN provides a specialised GSEA tool, which we used to study gene enrichment in these hallmark processes. Genes from MutFams, associated with a cancer, were subjected to gene set enrichment analysis for different hallmark processes using the GSEA online tool of The Atlas of Cancer Signalling Networks (ACSN) [120]. Significance was determined using the whole human genome as a background. We also considered enrichment in hallmark processes. 'Processes' are pathways, which can be grouped together into a particular hallmark. An example of a hallmark process is 'survival: MAPK', where the hallmark is survival and a specific process within this hallmark is the MAPK pathway. Correction for multiple testing was also performed and significant hallmarks were those with a P-Value ≤ 0.01.

GO slim enrichments

GOslim was used here since it provides a high level GO annotation and a very broad view of protein functions and has been widely used and cited in the literature by groups performing similar studies of GO term enrichment [175]. GOslim enrichments for the gene sets were executed using the PANTHER online tool using statistical over representation test [153]. Significant GOslims were those with a P value ≤ 0.01 for each gene set considered.

Reactome GO biological process enrichments

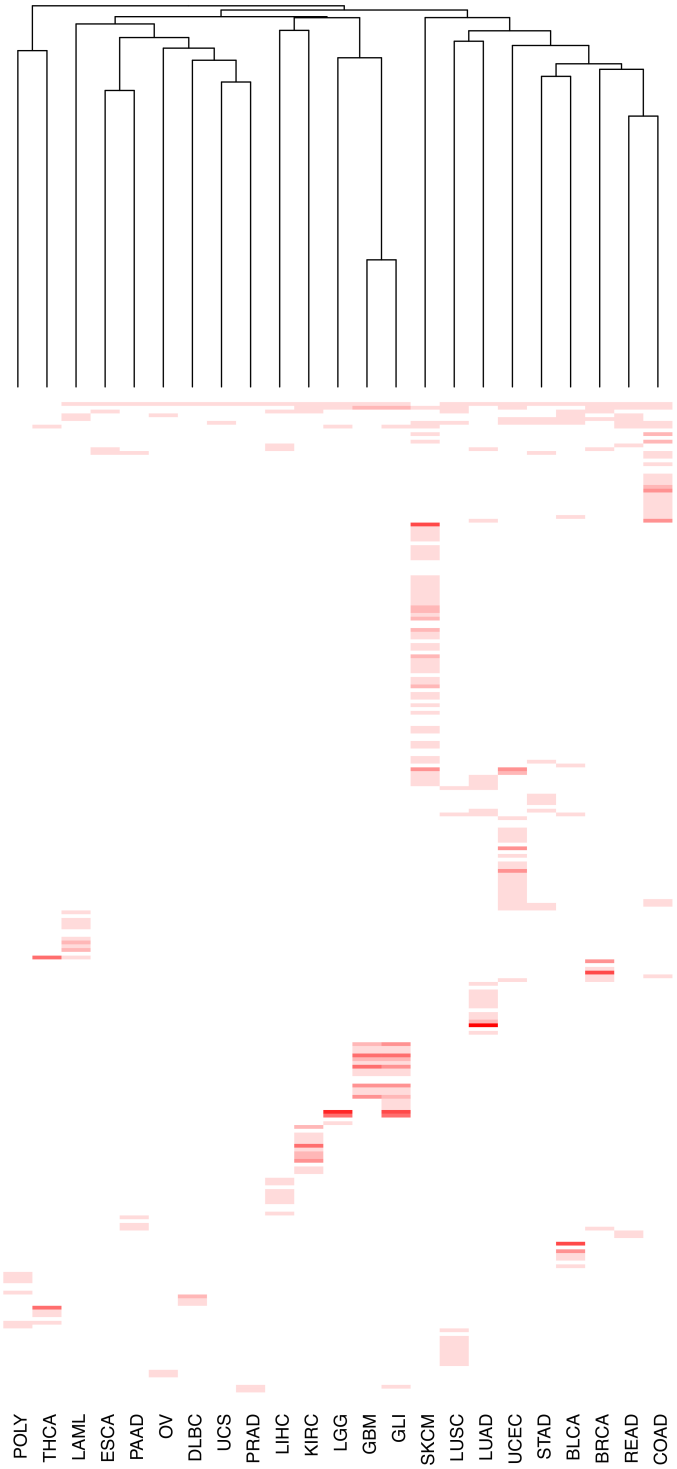
For the GO biological process annotations and analyses, the ReactomeFVIZ was used as this enabled the clustering of genes into more functionally coherent modules, and also the analysis of pathways within Reactome. This functionality is not provided by GOslims. The tool Reactome FVIZ takes an input set of genes and maps these onto the human protein network built by Wu *et al* [269], described in section 3.1.4.2. This network comprises functional associations between genes based on known and predicted data, using curated pathway databases such as Reactome and KEGG, known and predicted protein-protein interactions from yeast, worm and human, and Gene ontology (GO) annotations. Putative cancer genes were clustered into sub-network modules, using community detection methods [82].The modules were then subject to gene ontology biological process (GO-BP) enrichment studies, and were selected if they had an FDR ≤ 0.005 compared to a random distribution within Reactome FVIZ. This strict threshold ensured that only most significant GO-BP were included.

Results

Mutationally enriched FunFams (MutFams) in different cancer types

For each of the 22 cancer types considered (see Material and methods, table 3.1), mutations were accumulated across CATH FunFam domains, and statistically significant FunFams were selected, referred to as MutFams (see chapter 2, section 2.2.1.2). The number of MutFams identified for a cancer were not dependent on the number of mutations recorded for that cancer, both values provided in table 3.1. According to the heatmap in figure 3.9, the most common MutFam is the cellular tumour antigen P53, which contains the well-known cancer gene p53. This MutFam also possesses the largest additive enrichment factor. In contrast, the neutral polymorphisms from UniProt show the fewest enriched MutFams, with relatively low enrichment factors, consistent with their neutrality. The most enriched MutFam (with an enrichment factor of 2.499) for the neutral mutations is the MHC Class II antigen FunFam, whose members undergo somatic hyper-mutation within the adaptive immune response. SKCM (Skin cutaneous melanoma) harbours the highest number of enriched MutFams, which is likely to be due to the skin being exposed to a higher number of external carcinogens, but further studies would be needed to investigate this. It can also be seen in figure 3.9 that thyroid carcinoma (THCA) is near to the POLY in the dendogram, due to it not having the p53 enriched MutFam. Since p53 is known to be effected in THCA and almost all cancers, the literature has suggested that altering p53 function in THCA is engendered by other means than missesne mutations, such as truncation mutations, and mutations in transcription factors which in turn regulation p53 function [141].

Figure 3.9: Heatmap of the 22 cancers and their enriched MutFams. The cancer types are along the x axis, and the MutFams are along the horizontal where one MutFam is a coloured bar. More enriched MutFams are coloured a darker shade of red. The heatmap clustering is not based on mutation number, but on the common enriched MutFams in the cancers.



Considering cancer types that are from the same tissue, it can be seen that the gliomas (i.e found in brain tissue) cluster together, since they share a similar set of enriched MutFams. Other related cancer types include LUSC and LUAD which are the 2 main histological subtypes of non-small cell lung cancer (NSCLC). There are some common enriched MutFams between them, but they differ in their overall MutFam enrichment profiles, in that LUAD contains more enriched MutFams (29) than to LUSC (18). This difference has been shown in previous studies by Bruin and Swanton [46] [47], which showed that LUAD harbours 20% more driver mutations compared to LUSC, in the context of APOBEC mutations.

Assessing the functional relevance of predicted driver genes and known driver genes from the Cancer Genome Census (CGC)

Analysis of gene overlaps between predicted driver and CGC genes

We compared the MutFam predicted driver genes and the driver genes predicted by Miller *et al* and Yang *et al* (based on Pfam) to known cancer genes with missense mutations in CGC. The overlaps of the gene sets are shown in figure 3.10 below. It can be seen that the MutFam gene set has the most genes in common with CGC set accounting for 21.2% of the CGC gene set.

The Yang and Miller driver genes showed a relatively modest overlap of CGC genes, of 12.9% and 14.7% respectively. Although each of the driver gene methods produces a modest coverage of CGC genes, when we look at the coverage of genes from all 3 methods, this produces a greater percentage overlap with CGC genes, of 38.3%. This complementarity between driver gene methods has also been highlighted in studies by Karchin and co-workers who evaluated the different ways of assessing driver gene potential [247].

As with our approach here, Karchin and co-workers assessed the validity of predicted driver genes by examining their overlap to known cancer genes from the Cancer Genome Census (CGC). In agreement with our results, they showed that the combination of pre-

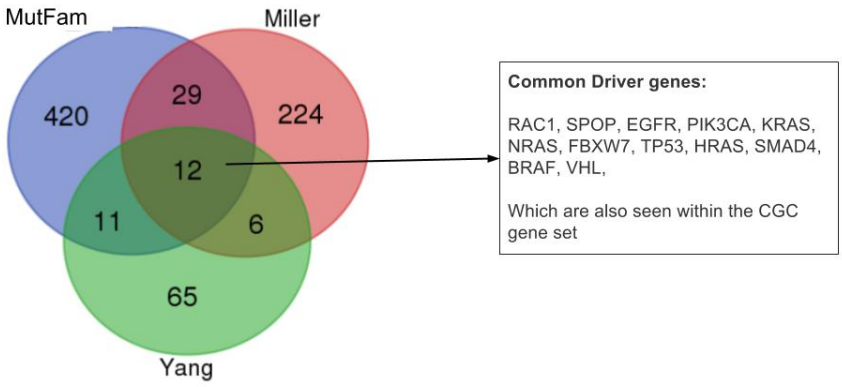
dicted genes from 8 driver gene predictors, considered in their analysis, showed significant enrichment of CGC genes, whereas the individual methods did not. It must be noted that the CGC gene list contains manually curated genes, included because experimental studies have suggested that these are drivers, and also genes suspected to be drivers but with no clear evidence that they induce cancer.

Figure 3.10: Comparison of predicted driver genes with known missense driver genes in CGC. a) The overlap of the driver methods with CGC genes b) comparison of the gene overlaps between driver gene methods, where the consensus driver genes are shown in a box to the right of the venn diagram.

a)

DATASETS	Number of total unique genes	Number of DATASET genes within CGC	% of DATASET genes in CGC to total CGC
MutFam	472	49	21.12
Miller	271	30	12.9
Yang	94	34	14.7
CGC	232	-	-

b)



In order to statistically elucidate the enrichment of predicted driver genes within the CGC genes, the MutFam and Miller genes were analysed on their overlaps within CGC genes relative to a random background of all human genes within SwissProt. Both MutFam and Miller driver gene sets showed a significant odds ratio of enrichment of 12.4 and 12.2 respectively, which were assessed using a Fisher's exact test, where both gene sets showed a P-value of test 2.2x10-16.

Enrichments in GO slim processes were performed to see if there was any convergence of non-overlapping driver genes from MutFam, Miller, and Yang sets, on particular cellular processes and how these processes compared to the consensus driver gene set. The Venn diagram in figure 3.10 shows the gene overlaps between the 3 driver gene methods considered. There are 420 MutFam only genes, 224 Miller only genes, 65 Yang only genes, and 12 consensus genes (all of which are within the CGC gene set). These 4 gene sets will be referred to as MutFam only, Miller only, and Yang only, and consensus driver genes respectively. Each was subjected to GOslim term enrichment analysis.

GOslim term analysis of the putative driver genes

The results of the GO slim analysis of the non-overlapping driver genes, and the 12 consensus driver genes are shown below. The 420 MutFam only genes were enriched in 14 GO slim terms, which were then manually categorised into 8 different cellular events as shown in table 3.2; Embryonic development, differentiation, cellular signalling/membrane and cytosolic transport, metabolic processes, DNA related processes, cell division, cell adhesion/migration, and stress response. The 224 Miller driver genes were enriched in 28 GO slims, within 6 different cellular event categories, including processes involved in biosynthetic processes and metabolism, shown in table 3.3.

There were common processes between the Miller and MutFam only genes, including terms implicated in adhesion/morphogenesis (light purple) and development (light blue). The 65 Yang only genes were enriched in 1 GOslim term, which was intracellular transduction. The consensus genes were enriched in GOslim terms involved in cytosolic signalling and vesicular mediated transport, where one of the terms specifically affected immune related signalling events of “I-KappaB Kinase/NF-KappaB cascade” shown in table 3.4. Although this analysis provides a general depiction of the functional attributes of the driver gene sets, further analysis on the specific biological processes involved was carried out, in order to characterise the processes in more detail.

Table 3.2: GOslim enrichments of the MutFam only driver genes. All GOslim enrichments have a P-value <0.01.

MutFam only genes (420)
endoderm development
developmental process
nervous system development
system development
ectoderm development
cell-cell adhesion
cell adhesion
biological adhesion
cellular component morphogenesis
anatomical structure morphogenesis
cell differentiation
response to stress

Table 3.3: GOslim enrichments of the Miller only driver genes. All GOslim enrichments have a P-value $<$0.01.

Miller only genes (224)
Sensory response to sound
mesoderm development
developmental process
system development
I-kappaB kinase/NF-kappaB cascade
intracellular signal transduction
Signal transduction
cell communication
Cellular process
receptor-mediated endocytosis
endocytosis
vesicle-mediated transport
Intracellular protein transport
Biosynthetic process
Metabolic process
Primary metabolic process
RNA - Metabolic process
Nitrogen compound metabolic process
regulation of transcription from RNA polymerase II promoter
regulation of nucleobase-containing compound metabolic process
nucleobase-containing compound metabolic process
Transcription, DNA-dependent
Cytokinesis
Mitosis
cellular component movement
cellular component morphogenesis
Muscle contraction

Table 3.4: GOslim enrichments of the consensus driver genes. All GOslim enrichments have a P-value $<$0.01.

Consensus driver genes (12)
I-kappaB kinase/NF-kappaB cascade
intracellular signal transduction
receptor-mediated endocytosis
endocytosis
Unclassified

Network analysis and enrichment of GO biological processes for the unique driver genes from the MutFam, Miller,Yang, and common gene sets

In order to filter out noise within the driver gene sets, the genes within MutFam, Miller, and Yang driver gene based methods were clustered based on functional associations to identify modules. This analysis was done using a method developed by Wu *et al* [269], whereby genes are mapped onto a network, in which genes are connected based on their functional associations (See section 3.2.6 Materials and Methods). The network is then clustered into modules of functionally related gene groups, which are in turn subject to a GO biological process enrichment analysis. The module-based GO biological process enrichments were then compared and contrasted between the MutFam, Miller, Yang unique genes, and the consensus genes.

These process are enriched with an FDR $<$0.0001, and are within modules comprising 3 or more genes. These modules are small nevertheless, where one or more putative driver genes are sufficient to cause a statistically significant enrichment given the background. Therefore we only show the putative driver genes within the tables 3.5 to 3.10, and tables 3.12 to 3.16, and have not listed all the genes in the module.

Module comparison between MutFam, Miller, Yang, and common gene sets

The modules identified for the unique driver gene sets can be seen in table 3-6, where modules enriched in GO biological processes are highlighted in red. It can be seen that the Yang only genes in table 3.7 have no modules with enriched processes at this FDR ($<$0.001), whereas there are 5 out of 7 modules and 12 out of 21 modules with enriched GO biological processes, within the MutFam and Miller genes respectively within tables 3.9 and 3.8 respectively. The common driver genes form 2 modules (table 3.6), both of which have enriched GO biological processes. It is important to note that not all driver gene sets mapped to functional modules, either because of their lack of functional associations, or because they could not be mapped to the network itself. The percentage

of mapped genes for each set are summarised in table 3.5.

Table 3.5: Percentage of common and non-overlapping driver genes mapped in network modules.

Driver gene set	total genes	mapped to modules	percentage in network modules
MutFam	420	193	46
Miller	224	119	53.1
Yang	65	20	31
Common drivers	12	9	75

Table 3.6: Gene module mapping for the common driver genes.

Module	Nodes in module	Node list	Significantly enriched GO biological processes (FDR >0.001)
0	5	EGFR,KRAS,NRAS,PIK3CA,RAC1	9
1	4	BRAF,FBXW7,HRAS,TP53	1

Table 3.7: Gene module mapping for the Yang driver genes

Module	Nodes in module	Node list	Significantly enriched GO biological processes (FDR >0.001)
0	6	CDKN2A,IRF4,KRT6A,MYC,PBRM1,PIM1	none
1	3	ITPR3,PRKACB,PRKCB	none
2	3	CTNNA3,PCDH11X,PCDH11Y	none
3	2	BCL2,PPP2R1A	none
4	2	MT-ND1,MT-ND6	none
5	2	PRPF19,U2AF1	none
6	2	BLNK,CD79B	none

Table 3.8: Gene module mapping for the Miller driver genes

Module	Nodes in module	Node list	Significantly enriched GO biological processes (FDR >0.001)
0	63	ACVRL1,AKT2,ANXA1,ANXA6,AR,BMPR1B,CDK8,CEBPA,CREBBP,CSNK1D, CSNK1E,CSNK1G1,CSNK2A2,DBP,DMBT1,DYRK1A,EEF1A1,EGR2,EP300,EPHA2, ERBB3,ERBB4,FGFR1,FLT4,GEM,GSK3B,HCK,HLF,IKBKE,LATS2,MAP3K4, MAPK11,MAPK8,MARK2,MYNN,MYO5A,MYO9A,MYT1,NF1,NFIL3,PDGFC,PDGFRA, PPIE,RAB27A,RAP1A,RAP1B,RASA1,RASA2,RASAL1,RHOA,RHOB,RHOG,RHOJ,RRAS, RUNX2,SMAD3,SNAI2,SOX17,TEK,TP63, ZAP70,ZBTB17,ZBTB24	18
1	37	ZFP2,ZNF117,ZNF180,ZNF181,ZNF184,ZNF208,ZNF250,ZNF260,ZNF28,ZNF286A,ZNF286B, ZNF320,ZNF34,ZNF345,ZNF461,ZNF468,ZNF470,ZNF483,ZNF527,ZNF540,ZNF546,ZNF554, ZNF568,ZNF571,ZNF572,ZNF583,ZNF616,ZNF620,ZNF624,ZNF681,ZNF70,ZNF71,ZNF768, ZNF836,ZNF845,ZNF883,ZNF92	2
2	9	CDK12,HNRNPA3,HNRNPF,HNRNPH1,HNRNPH2,MAP2K5,MAPK7,MAPKAPK2,RBM5	20
3	3	KCNB1,KCNB2,KCND2	16
4	3	AZIN1,BAAT,ODC1	12
5	2	ATP5B,ATP6V1B2	none
6	2	CDC42BPA,MYO18A	none

Table 3.9: Gene module mapping for the MutFam driver genes.

Module	Nodes in module	Node list	Significantly enriched GO biological processes (FDR >0.001)
0	33	ANGPT1,CA2,CA9,CBL,CYP4A11,FOXA3,GCK,HGF,HIF1A,HK2, HK3,HKDC1,HLA-DQA1, HLA-DQB1,INSR,MET,NKX2-2,NTRK3,PI4KB,PIK3C3,PIK3CG,PIK3R1,PIP4K2A,PIP4K2B, PRKD1,PTPN11,RET,SEC24A,SEC24B,SH3GL1,SLC4A2,TNC,TPTE	7
1	27	AKAP13,ARHGEF12,ARHGEF18,CDC42,DCC,DES,EPHA7,EPHB1, EPHB3,GRIK1,GRIK2,GRIK4,ITSN1,ITSN2,KALRN,NET1,NGEF, OBSCN,RHOF,RHOV, ROBO1,TGFBR2,TTN,UNC5C,UNC5D,VAV2,VAV3	21
2	25	ACTA1,ACTB,ACTC1,COL14A1,COL19A1,COL4A1,COL4A3,COL4A5, COL5A3,COL8A1,DNAJC5B,DNAJC6,DSC2,DSP,EVPL,FLNB,FOXC1, HSPA8,IRF6,MYH1,MYH13,MYH2,PLEC,RALA,SF3B1	9
3	24	CDH1,CDH5,CELSR2,PCDHA1,PCDHA13,PCDHA2,PCDHA3,PCDHA4, PCDHA5,PCDHA6,PCDHA8,PCDHA9,PCDHB10,PCDHB12,PCDHB13, PCDHB14,PCDHB16,PCDHB2,PCDHB3,PCDHB8,PCDHGA1,PCDHGA2, PCDHGA3,PCDHGA6	6
4	22	ACSL4,AQP7,ATRX,BCL6,CHD7,CTCF,DUSP1,DUSP22,EZH2,FOXA1, FOXA2,FOXE1,FOXF1,FOXQ1,IRF2,KMT2C,NKX2-1,RXRA,SOX1, SOX15,SOX3,TBK1	4
5	17	ZNF100,ZNF138,ZNF267,ZNF429,ZNF43,ZNF430,ZNF431,ZNF479,ZNF492, ZNF506,ZNF585A,ZNF585B,ZNF676,ZNF708,ZNF714,ZNF83,ZNF98	2
6	6	CALML6,CAMK1,CNGA2,MYLK,MYLK3,PYGM	none
7	4	MARK1,RPS6KA3,SIK3,STK11	1
8	4	C6,C7,C8A,C8B	7
9	4	KRT13,KRT15,KRT6C,KRT77	none
10	3	ABCC1,ABCC3,PIK3C2G	none
11	3	POU3F4,POU4F3,SOX9	2
12	3	NOTCH2,NOTCH4,PCSK5	3
13	3	PRDX1,PRDX2,PRDX6	5
14	3	PLXNA1,PLXNA2,PLXNA4	14
15	2	POU4F1,POU4F2	none
16	2	CPE,PCSK1	none
17	2	RPL8,RPS16	none
18	2	DUSP5,DUSP6	none
19	2	SYT3,SYT9	none
20	2	BMPR1A,BMPR2	none

Enriched GO biological processes in the network modules identified for the common driver genes

The common driver genes mapped to 2 modules within the protein network, which were enriched in 10 biological processes, as shown in table 3.10. The processes include those in cell signalling linked to various growth factor receptors, specifically “vascular epidermal growth factor receptor signalling”, “fibroblast growth factor receptor signalling”, “epidermal growth factor signalling”, and ”Insulin signalling”. These processes are all consistent with the GO slim processes implicated in intracellular signal transduction, where the mutated genes NRAS, PIK3CA, KRAS, and RAC1 form a signalling hub between these pathways. The diverse effects of this set of mutated genes is further demonstrated in other processes in Module 1, which are implicated in immunity, such as the “Fc-epsilon receptor signalling pathway”. All of these enriched processes are implicated in sustaining proliferative signalling, and inducing angiogenesis, both of which are known hallmarks in cancers [92].

Table 3.10: The enriched GO biological processes for the network modules identified for the common driver genes

Module	GeneSet	Nodes
1	vascular endothelial growth factor receptor signaling pathway	NRAS,PIK3CA,KRAS,RAC1
1	nerve growth factor receptor signaling pathway	NRAS,PIK3CA,KRAS,RAC1
1	Fc-epsilon receptor signaling pathway	NRAS,PIK3CA,KRAS,RAC1
1	blood coagulation	NRAS,PIK3CA,KRAS,RAC1
1	leukocyte migration	NRAS,PIK3CA,KRAS
1	fibroblast growth factor receptor signaling pathway	NRAS,PIK3CA,KRAS
1	insulin receptor signaling pathway	NRAS,PIK3CA,KRAS
1	epidermal growth factor receptor signaling pathway	NRAS,PIK3CA,KRAS
1	innate immune response	NRAS,PIK3CA,KRAS,RAC1
2	positive regulation of ERK1 and ERK2 cascade	FBXW7,BRAF,HRAS

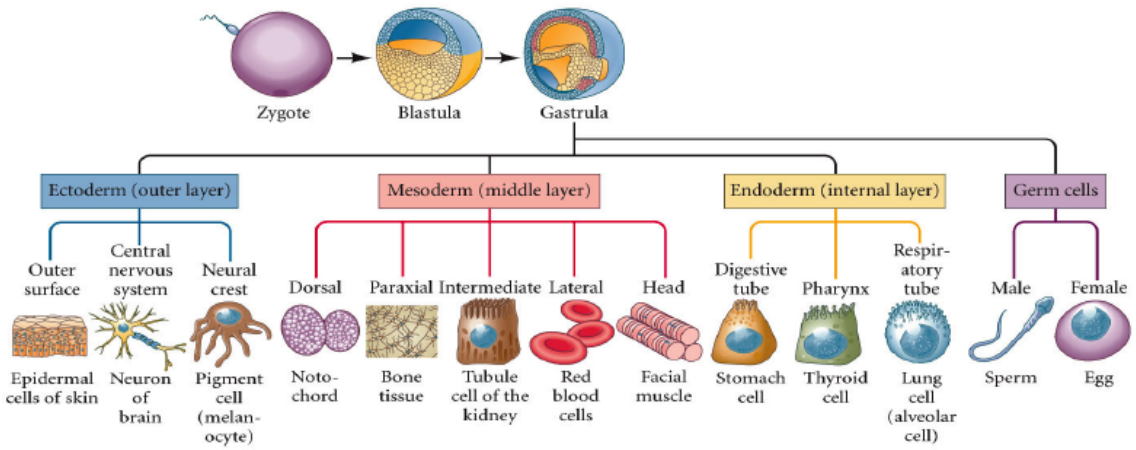
Enrichment of GO biological process in the network modules identified for the unique MutFam and Miller gene sets

The MutFam and Miller studies cover different cancers (as shown in table 3.11) and this is likely to affect the mutationally enriched families identified. However, it is also interesting to see that the genes identified also map to distinct gastrulation layers. The mutated genes from the Miller and MutFam datasets are derived from various cancer types, some of which occur in different tissues. This difference may affect the enriched biological processes within each driver set, by enriching tissue specific processes. Therefore, in order to examine this tissue bias, the cancer types of each driver gene set (MutFam and Miller) were compared with respect to their tissue origins in gastrulation, see table 3.11. The 3 gastrulation layers, ectoderm, mesoderm, and endoderm are formed during embryogenesis, and are shown in the image under figure 3.11. The cancers that are common and unique to the Miller and MutFam studies are listed in table 3.11. The number of cancers found in the 3 gastrulation layers are also shown.

Table 3.11: Comparing the cancers within the MutFam and Miller driver genes. The 3 different gastrulation layer columns contain the numbers of the different cancers in each group. The cancer abbreviations are listed in Materials and methods.

Datasets	Cancer count	Cancers	Ectoderm	Mesoderm	Endoderm
Miller MutFam	17	SKCM, LIHC, GBM, STAD, UCS LUSC, BLCA, OV, LGG, UCEC, LUAD, COADREAD, THCA, KIRC, LAML, BRCA, PRAD,	3	6	8
Miller	12	THCA,LUSC,KICH, OV, KICH, SKCMHNSC,CESC, ACC, UCS, LUAD, KIRP	2	6	4
MutFam	4	DBLC, ESCA, GLI, PAAD,	2	1	1

Figure 3.11: The 3 gastrulation layers within the embryo form distinct tissue types. Taken from [1].



The cancers in the Miller study show a relative bias towards mesodermal tissues compared to cancers in the MutFam study which are more evenly spread across the 3 layers with a slight relative bias towards ectoderm tissues. Therefore, there is a difference in tissue bias between the two driver gene sets, which may partly explain the low overlap in driver genes, and should be accounted for when comparing cellular processes.

Processes implicated in cellular development and differentiation in MutFam and Miller modules

More detailed analysis of modules involved in cellular development was performed to see if there was convergence on common or related biological processes for the MutFam and Miller gene sets. This showed that both Miller and MutFam modules were enriched in biological processes implicated in cellular development, but within different gastrulation layers.

MutFam modules For the MutFam genes, there are three gene modules which are enriched in developmental processes within the endoderm layer (which eventually forms the lung, thyroid, and digestive tract) and the ectoderm layer (which forms the skin epidermis, neurons in the brain, and pigment cells). This tissue bias in the MutFam cancer set explains the specific enriched processes of “nervous system development”, “axon guidance”, and “synapse assembly”, to mention a few. The full list of developmental pathways enriched are shown in table 3.12. Processes affecting neuronal and system development have been implicated in some cancers, where changes in neuronal guidance, migratory contacts, and Ephrin signalling contribute towards breast cancer carcinogenesis [93][163].

Table 3.12: Enriched GO processes in the network modules identified for the MutFam genes in cellular development. Processes common to the Miller genes are highlighted in bold.

Module	GO Biological process	Gene list
1	ephrin receptor signaling pathway	ITSN1,KALRN,EPHB1,EPHB3,EPHA7,VAV2,NGEF
1	**axon guidance**	**ITSN1,KALRN,ROBO1,EPHB1,EPHB3,EPHA7,DCC, UNC5C,UNC5D,VAV2,NGEF**
1	retinal ganglion cell axon guidance	EPHB1,EPHB3,EPHA7
1	nerve growth factor receptor signaling pathway	ITSN1,KALRN,ARHGEF18,VAV2,NET1,OBSCN,NGEF
1	glutamate receptor signaling pathway	GRIK4,GRIK1,GRIK2
1	ionotropic glutamate receptor signaling pathway	GRIK4,GRIK1,GRIK2
1	synaptic transmission, glutamatergic	GRIK4,GRIK1,GRIK2
1	anterior/posterior axon guidance	DCC,UNC5C
1	negative regulation of collateral sprouting	EPHA7,DCC
1	peptidyl-tyrosine phosphorylation	TTN,EPHB1,EPHB3,EPHA7
1	dendritic spine development	EPHB1,EPHB3
1	regulation of cell-cell adhesion	EPHB3,EPHA7
1	regulation of GTPase activity	EPHB3,VAV2,NGEF
1	dendritic spine morphogenesis	EPHB1,EPHB3
1	central nervous system projection neuron axonogenesis	EPHB1,EPHB3
3	nervous system development	PCDHB12,PCDHA1,PCDHB2,PCDHA5, PCDHA4,PCDHA3,PCDHA2,PCDHB3,PCDHA8, PCDHA6
3	synapse assembly	PCDHB14,PCDHB13,PCDHB10,PCDHB2, PCDHB16,PCDHB3
3	synaptic transmission	PCDHB14,PCDHB13,PCDHB10,PCDHB2, PCDHB16,PCDHB3
14	regulation of axon extension involved in axon guidance	PLXNA2,PLXNA1,PLXNA4
14	branchiomotor neuron axon guidance	PLXNA2,PLXNA1,PLXNA4
14	semaphorin-plexin signaling pathway	PLXNA2,PLXNA4
14	**axon guidance**	**PLXNA2,PLXNA1,PLXNA4**
14	chemorepulsion of branchiomotor axon	PLXNA4
14	cerebellar granule cell precursor tangential migration	PLXNA2
14	postganglionic parasympathetic nervous system development	PLXNA4
14	vagus nerve morphogenesis	PLXNA4
14	anterior commissure morphogenesis	PLXNA4
14	dichotomous subdivision of terminal units involved in salivary gland branching	PLXNA1
14	glossopharyngeal nerve morphogenesis	PLXNA4
14	trigeminal nerve structural organization	PLXNA4

Miller modules In contrast, the Miller genes were implicated in developmental processes within the mesoderm layers. These form the cardiac, smooth and skeletal muscle tissues, kidney tubules, and red blood cells, shown in table 3.13. The specific GO biological processes of the Miller genes are consistent with mesodermic processes affecting “cardiac rhythm” and “regulation of smooth muscle contraction”, and other processes. Manipulation of the cellular cytoskeleton, using contractile elements such as those used in muscle contraction, is a common event within cell migration – specifically the amoeboid mode of migration - which enables invasion through the surrounding extracellular

matrix [174].

The Miller genes also include RhoGTPases (see table 3.13). In the context of cellular development, the RhoGTPases are downstream effectors of Ephrin receptor signalling, where RhoGTPase activation leads to actin monomer assembly into larger cytoskeletal structures, imperative for cellular adhesion and migration, and in turn axon guidance, which has been shown to be implicated in cancer [203].

Table 3.13: Enriched GO processes in the network modules identified for the Miller genes in cellular development. Processes common to the MutFam genes are highlighted in bold.

Module	GO Biological process	Gene list
1	fat cell differentiation	GSK3B,AKT2,EP300,CEBPA,EGR2
1	**axon guidance**	**GSK3B,PDGFRA,RHOG,RHOB,RASA1,RASA2, EPHA2,RASAL1,RRAS,ERBB4,FGFR1**
1	circadian rhythm	GSK3B,DBP,EP300,DYRK1A,NFIL3
3	regulation of smooth muscle contraction	KCNB2
3	locomotor rhythm	KCND2

Both the MutFam and Miller genes showed enrichment in processes affecting the Notch signalling pathway involved in cellular differentiation, but via contrasting mechanisms, shown in table 3.14. The MutFam genes are enriched in the “Notch receptor processing” process, and the mutations are within the receptors themselves, NOTCH2 and NOTCH4. Whereas the Miller genes are enriched in the “Notch signalling pathway”, affecting upstream and downstream regulators of the Notch pathway such as EP300 (Histone acetylase in chromatin modification), CENPA (Histone H3 alternative involved in kinetochore assembly), and the transcriptional regulators CDK8 and SNAI2. It has been suggested that altering cellular development in cancer, implicated by both the MutFam and Miller genes, plays a role in maintaining a pluripotent cellular state, which is more adaptable and genetically flexible upon changing conditions, contributing to cancer cell survival **??**. Therefore mutations in proteins involved in such developmental pathways may act to maintain this pluripotent state, contributing to the carcinogenic phenotype.

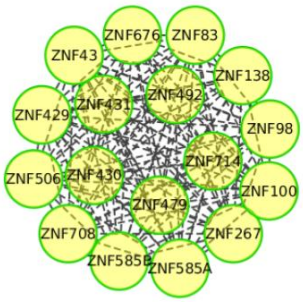
Table 3.14: Enriched GO processes in the network modules identified for both the MutFam and Miller genes, involved in cellular differentiation.

Module	**GO biological process (MutFam)**	**Gene list**
12	cell fate determination	NOTCH2,NOTCH4
12	Notch receptor processing	NOTCH2,NOTCH4
12	hemopoiesis	NOTCH2,NOTCH4
Module	**GO biological process (Miller)**	**Gene list**
1	keratinocyte differentiation	TP63,ANXA1,LATS2,EPHA2
1	Notch signaling pathway	EP300,TP63,CEBPA,CREBBP,CDK8,SNAI2

Common processes in the MutFam and Miller modules implicated in DNA binding and transcriptional regulation

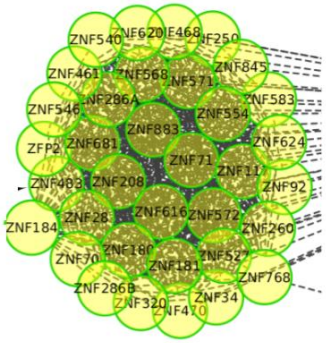
In addition to cellular development and differentiation, there are other common GO processes between MutFam and Miller-unique gene sets including those affecting DNA transcription. Both the MutFam and Miller modules include proteins with a zinc-finger domain (as shown in figure 3.12, which enables DNA binding, and is known to be implicated in cancer (Jen2016). Effect on genome stability and regulation is one of the cancer hallmarks, where further genetic errors arise as a result of compromised DNA quality control [92]. Although the MutFam and Miller genes map to different modules shown as shown in figure 3.12, these modules are enriched in the same processes of regulation of transcription, DNA-dependent. This result further highlights the convergence of GO biological processes in the different driver gene sets. In this case they contain a related DNA binding domain, specifically the zinc finger binding domain.

Figure 3.12: Unique zinc finger DNA binding genes in the Miller and MutFam sets map to network modules, which are associated with common GO biological processes. The figures are made in cytoscape.



MutFam genes

Module	GO biological process	Gene list
5	regulation of transcription, DNA-dependent	ZNF430,ZNF100,ZNF492,ZNF98,ZNF43,ZNF708,ZNF83,ZNF429,ZNF506, ZNF714,ZNF138,ZNF479,ZNF676,ZNF585B,ZNF267,ZNF431,ZNF585A
5	transcription, DNA-dependent	ZNF430,ZNF100,ZNF492,ZNF98,ZNF43,ZNF83,ZNF429,ZNF506,ZNF714, ZNF138,ZNF479,ZNF676,ZNF585B,ZNF267,ZNF431,ZNF585A



Miller genes

Module	GO biological process	Gene list
1	regulation of transcription, DNA-dependent	ZNF28,ZNF572,ZNF571,ZNF250,ZNF470,ZNF208,ZNF527,ZNF845,ZNF768,ZNF286B,ZNF624,ZNF568,ZNF546,ZNF468,ZNF620,ZNF883,ZNF286A,ZNF320,ZNF583,ZNF681,ZNF483,ZNF461,ZNF184,ZNF260,ZFP2,ZNF181,ZNF180,ZNF34,ZNF92,ZNF70,ZNF71,ZNF836,ZNF616,ZNF117,ZNF554
1	transcription, DNA-dependent	ZNF28,ZNF572,ZNF571,ZNF470,ZNF208,ZNF527,ZNF845,ZNF286B,ZNF624,ZNF568,ZNF546,ZNF468,ZNF620,ZNF883,ZNF286A,ZNF320,ZNF540,ZNF583,ZNF681,ZNF483,ZNF184,ZNF260,ZFP2,ZNF181,ZNF34,ZNF92,ZNF70,ZNF71,ZNF836,ZNF616,ZNF117,ZNF554

Different GO biological processes identified in the network modules identified for the unique driver genes

Cell adhesion: Unique MutFam modules Unique to the MutFam modules, there is enrichment of genes in the Proto-cadherin family, which are involved in cell adhesion and migration (see table 6, module 3). The specific processes enriched for this module are “homophillic cell adhesion” and “calcium dependent cell adhesion” shown in table 3.15, both of which have been implicated in cancer. Proto-cadherin genes are believed to act as a chemical conduit between intracellular signalling and migratory processes during cancer cell invasion [100] [19].

This is further supported by studies that show ECM reorganisation and alteration of such cell-cell junction proteins give an increased invasion in carcinomas [92] [203]. These effects (i.e activation of focal adhesions and the break-down of cell-cell adherences such as cadherin-based adheren’s junctions) are all reminiscent of a phase that cancer cells undergo to increase invasion and metastasis called the Epithelial to Mesenchymal Transition (EMT) [92].

Table 3.15: Unique GO biological processes identified in the MutFam modules involved in cellular adhesion

Module	GO biological process	Gene list
3	homophilic cell adhesion	PCDHGA6,PCDHGA3,PCDHGA2,PCDHGA1,PCDHB14, PCDHB13,PCDHA13,PCDHB12,CELSR2,PCDHB10,CDH5, PCDHA1,PCDHB2,CDH1,PCDHB16,PCDHA5,PCDHA4, PCDHA3,PCDHA2,PCDHB3,PCDHA9,PCDHA8,PCDHB8, PCDHA6
3	calcium-dependent cell-cell adhesion	PCDHB14,PCDHB13,PCDHB10,PCDHB2,PCDHB16,PCDHB3
3	cell adhesion	PCDHB12,PCDHA1,PCDHB2,PCDHA5,PCDHA4,PCDHA3, PCDHA2,PCDHB3,PCDHA8,PCDHA6

RNA splicing: Unique Miller modules The GO biological processes unique to the Miller driver gene set include those implicated in RNA splicing, RNA processing, and gene expression, all of which have been implicated in cancer [205]. Specifically, the affected genes include various Heterogeneous Nuclear Ribonucleoproteins (HNRNP), cyclin dependent kinase 12 (CDK12), and RNA Binding Motifs (RBM), as shown in table 3.16. All genes play a crucial role in facilitating transcription by preventing RNA secondary structures, and aiding mRNA pre-processing and nuclear exportation steps, recently been shown to be implicated in tumour progression [244] .

Table 3.16: Unique GO biological processes identified in the MutFam modules, involved in RNA splicing

Module	**GO biological process**	**Gene list**
2	RNA splicing	HNRNPA3,HNRNPH1,HNRNPF,HNRNPH2,CDK12, RBM5
2	nuclear mRNA splicing, via spliceosome	HNRNPA3,HNRNPH1,HNRNPF,HNRNPH2,RBM5
2	regulation of RNA splicing	HNRNPH1,HNRNPF,CDK12
2	gene expression	HNRNPA3,HNRNPH1,HNRNPF,MAPKAPK2,HNRNPH2, RBM5
2	RNA processing	HNRNPH1,HNRNPF,RBM5

Analysis of enriched cancer hallmarks from the Atlas of Cancer Signalling Networks (ACSN)

To assess the validity and clinical relevance of MutFams for identifying driver genes, the sets of MutFam genes for the 22 cancers were analysed for their enrichment in different ACSN hallmark processes using the ACSN online tools [120]. The hallmark analysis was also performed on the driver cancer genes from the Miller and Yang studies, identified using Pfam domains [154] [271]. Hallmarks for the mutated genes for the MutFams, Miller, and Yang sets were also compared to known cancer genes within the Cancer Gene Census (CGC) [15].

In order to see which cancer hallmarks were enriched for the predicted and known driver gene sets, the online tool for the Atlas of Cancer Signalling Networks (ACSN) [120] was used. The ACSN resource contains 4600 reactions covering 564 genes, which are then grouped into 5 main cellular processes; Apoptosis, Survival, Epithelial-mesenchymal transition (EMT), cell cycle, and DNA repair. To see the hallmarks enriched for the predicted and known cancer genes, a gene set enrichment analysis was performed for the MutFam, Miller, Yang genes, and CGC genes. The results of this are shown in figure 3.13.

Figure 3.13: Summary of mapped ACSN hallmark modules for the predicted and known cancer genes. The colour blocks indicate an enriched process within the hallmark categories of Survival, EMT cell motility, DNA repair.

	Survival			EMT cell motility			DNA repair			
Genesets	**MAPK**	**PI3K/AKT_MTOR**	**master**	**master**	**Cell-cell adhesion**	**Regulators**	**master**	**Fanconi**	**G1/S checkpoint**	**G1/CC checkpoint**
MutFam	**Y**	N	N	**Y**	**Y**	N	N	N	N	N
Miller	**Y**	**Y**	**Y**	**Y**	N	N	N	N	N	N
Yang	**Y**	**Y**	**Y**	N	N	**Y**	N	N	N	N
CGC	**Y**	N	N	N	N	N	**Y**	**Y**	**Y**	N

This analysis showed that the MutFam, Miller, and Yang genes show enrichment of hallmarks for 2 main cellular processes: Survival, and EMT Cell motility, which is consistent with the enriched GOslim processes in table 3.4. Specifically, the enriched hallmark processes comprised MAPK and PI3K, AKT and MTOR growth factor signalling genes, which affect processes involved in cell to cell adhesions and polarity. These results are consistent with studies that show PI3K and MTOR pathways are highly mutated in pan-cancer mutations [245] [29].

The CGC genes mainly affect DNA repair events, with only one affecting a survival process. Unlike the predicted driver genes sets, there was no enrichment in hallmarks affecting EMT and cell motility. It must be noted that some of the CGC genes, whilst mainly containing missense mutations, also contain other mutation types such as insertions and deletions [257].

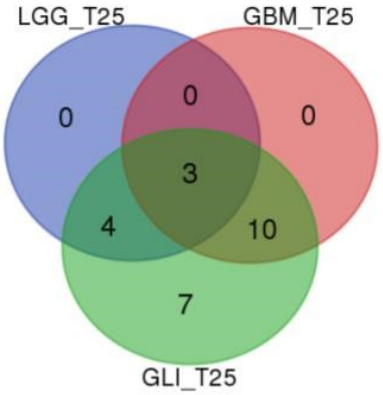
Detailed analysis of MutFam driver genes in brain cancers

Analysing the driver gene overlaps between glioma subtypes

Specific analysis of predicted MutFam cancer genes, within a subset of brain cancers, was performed, namely for the gliomas; low grade glioma (LGG), glioblastoma multiforme (GBM) and glioma (GLI). LGG is a grade II glioma and is the least invasive of the subtypes, but can progress into later stage glioma subtypes, including the more aggressive grade IV glioblastoma multiforme (GBM)[36]. Glioma (GLI) encapsulates mutations from both LGG and GBM, along with other subtypes of different grades and differing brain cell types [36]. GBM is one of the most mature (stage IV) of invasive gliomas with a poor prognosis [259]. LGG, GBM, and GLI were chosen as they share many MutFams. The driver genes analysed for each glioma subtype, were the top quartile mutated genes in each cancer MutFam. The overlap between the MutFam genes for each cancer is shown in figure 3.14.

It can be seen from figure 3.14 that 3 well-known cancer genes, TP53, PTEN, and CHEK2, are common to all three glioma types. The genes within LGG are either common to GBM (3 genes), or are within GLI (4 genes), and all but CHEK2 are within the Cancer Genome Census (CGC). GBM and GLI do contain other mutated genes not in LGG, and GLI contains genes neither in LGG nor GBM. To investigate the GO biological processes of the three glioma gene sets, network analyses for each of the driver gene sets were performed.

Figure 3.14: Venn diagram comparing MutFam genes between LGG, GBM, and GLI gliomas. Overlaps to CGC genes are shown in the table below. Genes in bold are those common to the CGC genes.



Groups	Total genes	elements
GBM LGG GLI	3	**TP53 PTEN** CHEK2
LGG GLI	4	**CIC BRAF NOTCH1 IDH1**
GBM GLI	10	ZNF646 **HIF1A** HSPA8 ZNF429 GSTM5 **EGFR PIK3R1 LZTR1** FKBP9 TBK1
GLI	7	**PIK3CA** PCDHA1 KRT13 **ATRX** KRT15 CACNA1S PCDHA3

Enrichment of GO biological processes and network analysis of glioma MutFam driver genes

To enable a deeper analysis of GO biological processes and pathways represented by the low grade glioma (LGG), glioblastoma multiforme (GBM), and glioma (GLI) genes, a gene network was constructed using ReactomeFVIZ [269] for the LGG and GBM MutFam gene sets respectively, as described in the materials and methods section (3.2 Materials and methods).

A total of 3, 8, and 12 genes from the LGG, GBM, and GLI gene sets could be mapped to the functional network. These were then clustered into gene modules for further analysis, shown in table 3.17. There were a total of 1, 2, and 3 gene modules in LGG, GBM, and GLI respectively, which parallels the progression of the glioma cancers from an early stage LGG to later stage gliomas, such as GBM which harbours more

mutated genes. To get an insight into what processes are affected within these modules, and whether this reflects the distinct clinical phenotypes of the cancer subtypes, further analyses of the GO biological process enrichments were carried out on the gene network modules. For clarity, figures including networks for just LGG and GBM are shown.

Table 3.17: Network modules of the LGG, GBM, and GLI MutFam (Top quartile) genes.

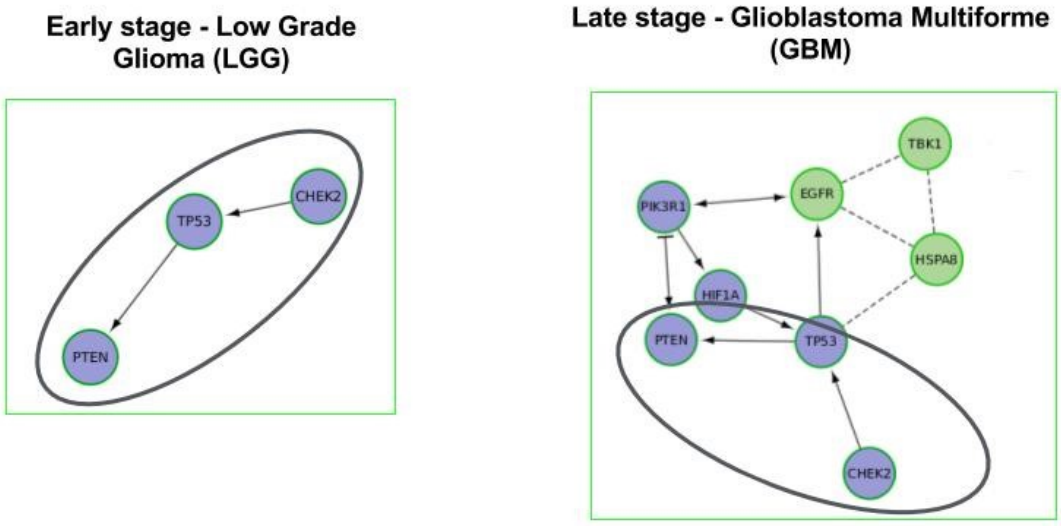
Cancer	**Module**	**Nodes in modules**	**Node list**
LGG	1	3	CHEK2,PTEN,TP53
Cancer	**Module**	**Nodes in modules**	**Node list**
GBM	1	5	CHEK2,HIF1A,PIK3R1, PTEN,TP53
GBM	2	3	EGFR,HSPA8,TBK1
Cancer	**Module**	**Nodes in modules**	**Node list**
GLI	1	6	HIF1A,NOTCH1,PIK3CA, PIK3R1,PTEN, TP53
GLI	2	4	CHEK2,EGFR,HSPA8, TBK1,
GLI	3	2	PCDHA1,PCDHA3

Common gene modules detected in all gliomas: DNA repair and cell cycle checkpoint There is one gene network module (module 1 in table 3.17) common to all gliomas; LGG, GBM, and GLI, containing the core genes of TP53, CHEK2, and PTEN. All genes are known tumour-suppressors [257], and are involved in DNA repair pathways, regulation within cell cycle checkpoint, and the regulation of growth factor signalling. This result is consistent with the biological pathways identified, i.e those implicated in p53 downstream events involved in DNA repair and the cell cycle checkpoint mediated by CHEK2, whereby p53 alters the DNA damage response and appropriate entry to the cell cycle [51].

Interference with DNA quality control and repair pathways would be expected to lead to further genomic errors, and is one of the core hallmarks of all cancers [92]. This common module is circled in figure 3.15 and reflects the fact that these cancers, which come

from the same tissue type, have shared mechanisms, although GBM has developed into a more aggressive and invasive form, which contains other mutated genes.

Figure 3.15: Network modules in LGG and GBM. The common module 1 is in blue, and a GBM/GLI specific module is in green. Common genes between the two cancers are circled.



Gene modules unique to GBM and GLI: Developmental signalling Genes within module 1, which are both in GBM and GLI are HIF1A and PIK3CR1. GLI has an additional gene involved in developmental signalling – NOTCH1 (see table 3.17. These extra genes mediate various extracellular nutrient and growth factor signals involved in the intracellular hypoxia response and engendering angiogenesis and vasculature formation for tumour growth. This observation is consistent with studies by Gavalas *et al* [166]. It has been shown that these processes predominate in later stage carcinomas, and represent an adaptive response which promotes increased circulation and absorption of nutrients to support the neoplastic growth of the tumour [92] [12]. These results are also consistent with the reported biological processes for these GBM and GLI cancers shown by GO enrichments analysis (see table 3.18), which includes processes effecting "hypoxia response via HIF activation" and the "HIF-1 signalling pathway".

Table 3.18: Enriched biological processes for the LGG, GBM, and GLI cancers.

Cancer	Total biological process	Biological process
LGG GBM GLI	15	AP-1 transcription factor network(N) p53 pathway feedback loops 2(P) Prostate cancer(K) Central carbon metabolism in cancer(K) Melanoma(K) Small cell lung cancer(K) p53 pathway(N)] Hepatitis B(K) PLK3 signaling events(N) p53 pathway(P) p53 signaling pathway(K) Endometrial cancer(K) Sphingolipid signaling pathway(K) Glioma(K) DNA Double Strand Break Response(R)
LGG	3	Cell Cycle Checkpoints(R) Cell cycle(K) Direct p53 effectors(N)
GBM	3	Hypoxic and oxygen homeostasis regulation of HIF-1-alpha(N) hypoxia and p53 in the cardiovascular system(B) HTLV-I infection(K)
GBM GLI	21	Hypoxia response via HIF activation(P) HIF-1 signaling pathway(K) Phosphatidylinositol signaling system(K) Proteoglycans in cancer(K) Thyroid hormone signaling pathway(K) Pancreatic cancer(K) Choline metabolism in cancer(K) mTOR signaling pathway(K) Apoptosis(K) Renal cell carcinoma(K) Colorectal cancer(K) Chronic myeloid leukemia(K) PIP3 activates AKT signaling(R) PI Metabolism(R) p75(NTR)-mediated signaling(N) VEGFR1 specific signals(N) Pathways in cancer(K) Non-small cell lung cancer(K) PI3K-Akt signaling pathway(K) Signaling events mediated by Stem cell factor receptor (c-Kit) (N) BCR signaling pathway(N)

Gene modules unique to GBM and GLI: immune response and protein folding GBM and GLI also contain gene modules implicated in the immune system and protein folding (TBK1, HSPA6), shown in figure 3.15 (Green circles). TBK1 is a regulatory signalling kinase which lies directly upstream from the main inflammatory pathway involving NFKB, and mutations in this protein have been found in various cancers [212]. This is consistent with other studies reporting pathways involved in immunity, such as "B cell receptor signalling" in table 3.18 .

HSPA6 is a molecular chaperone within the HSP70 chaperone family, and acts in response to cellular stress to maintain proteostasis and prevent protein mis-folding. It also plays a role in antigen presentation in immunity. The role of chaperones in cancer is controversial, since although they ensure correct protein folding, they can also mask internal errors within mutated proteins and therefore contribute to their abnormal functioning [106]. Abnormal activity in HSPA6 and other Hsp70 members has been seen in cancer, and is associated with tumour progression and poor prognosis [162]. These effects are consistent with the reported increased involvement of the mature immune response in late stage cancers such as GBM [92] [12], and are often associated with increased drug resistance due to the recruitment of immune cells which form a protective barrier around the tumour, and are often associated with a poorer prognosis [259] .

Unique to GLI, there is a gene module that contains a range of enriched processes involving proto-cadherin genes, which are implicated in cellular adhesion and cell signalling. This module has already been discussed in the analysis of the MutFam genes, within the section 3.3.3.6. The unique genes within the GLI glioma are associated with different stages of tumour progression and occur within different types of brain cells.

Conclusion

The heterogeneous nature of cancer complicates the analysis of the mutational landscape, whereby driver genes and their mutations are hidden amongst the noise of less functionally significant passenger mutations. The MutFam protocol has been developed for prioritising cancer driver genes, based on the enrichment of missense mutations in a common functional domain, a CATH FunFam. The performance of the MutFam protocol compared to Pfam domain based prediction methods, shows that MutFams identify the highest coverage of known cancer genes in the CGC. i.e. 21.2% compared with other studies: 12.9% for Miller and 14.7% for Yang. The remaining CGC genes, not found in the MutFam, Miller, or Yang genes, may contain higher proportions of other types of variants i.e. deletions and splice variants, and are enriched in DNA repair processes.

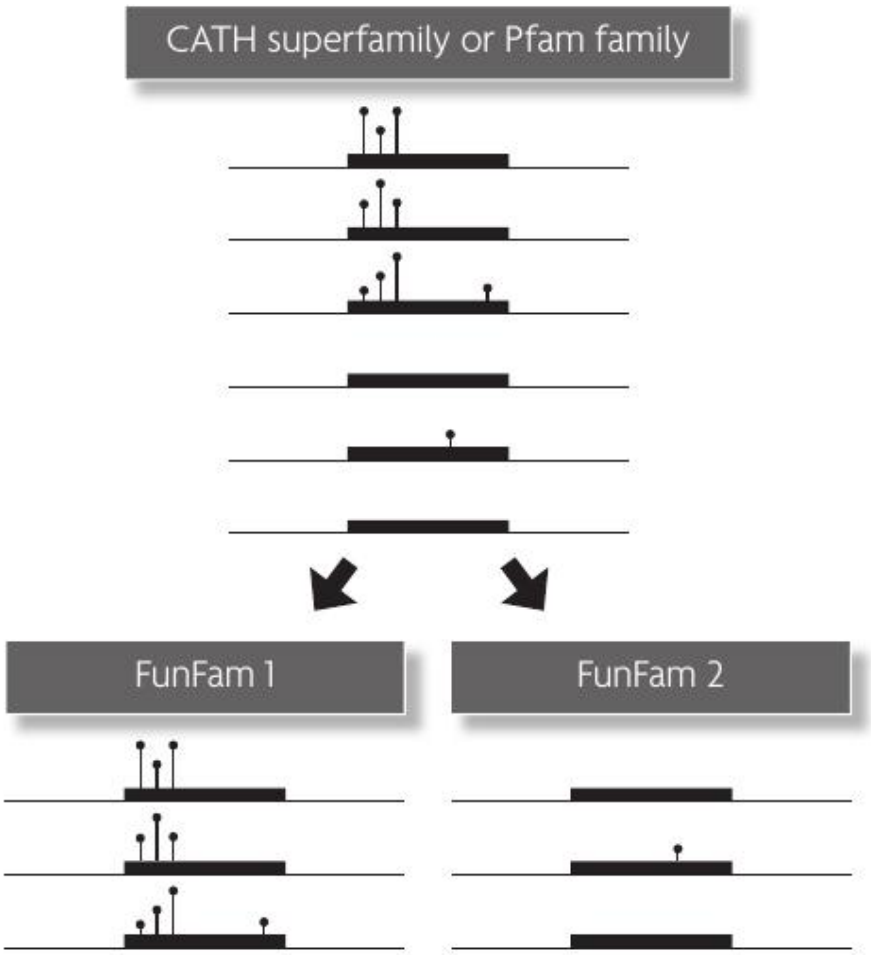
Although protein structures are intensively studied, there needs to be at least one protein structure to identify the family. In contrast, Pfam can identify families purely from sequence data and this resource therefore contains more protein entries and more families and accounts for a higher proportion of human domains than CATH families. For our analysis we used functional subfamilies from CATH to aggregate mutations across the family. We have demonstrated that CATH FunFams are functionally more coherent than Pfam families. This means that using CATH FunFams for mutation aggregation is more likely to identify a statistically significant signal of enriched mutations than using Pfam families, in which some relatives will have different function and therefore may not be mutationally affected for a particular disease. Figure 3.16 schematically illustrates this concept. In this context, it is not surprising that CATH FunFams identify more mutationally enriched domain families than Pfam families. Whilst, some of these may represent false positives, it is not possible to compare the level of false positives between FunFams and Pfams, but it is reassuring that genes from mutationally enriched families identified by both methods, map to similar pathways and processes as demonstrated in the ReactomeFVIZ analyses.

Figure 3.16 shows the benefit of sub-classifying evolutionary relatives into functional families for the purposes of aggregating mutation data. The top half of the figures shows

a set of protein domains in a family – this can be a broad CATH superfamily or a more closely related set of domains in a Pfam family. Mutations in each domain are shown by a vertical bar, the height which reflects the number of mutations. It can be seen that the top 3 relatives have many disease associated mutations in a number of sites common to the relatives. However, the bottom 3 relatives have very few mutations. The lower half of the figure shows the sub-classification of these relatives into functional families (FunFam1 and FunFam2). Mutation enrichment studies analysing the set of relatives in the top half of the figure (eg across a CATH superfamily) would return a low enrichment score as only half the relatives have significant mutations. In contrast, enrichment analyses of the functional families, FunFam1 and FunFam2, would indicate that FunFam1 is significantly enriched in disease associated mutations.

It is possible that CATH contains more families comprising domains that are highly ordered, than Pfam, since each family has at least one solved structure. However, recent work by the Pfam team have identified families for most proteins having structures deposited in the PDB. Therefore it is unlikely that this feature would represent a large enough difference in the number of families with ordered structures, to account for the differences in the number of mutationally enriched families between CATH FunFams and Pfam.

Figure 3.16: Comparing mutation aggregation within a CATH superfamily, a Pfam family, and a CATH FunFam.



Analysis of the GO biological processes demonstrated that the non-overlapping MutFam and Miller driver genes show convergence on common biological processes, and encapsulate mutated proteins in different parts of these pathways. Many of these processes have been reported in the literature, in connection with cancer. The unique pathways for each driver gene set could be rationalised by considering the gastrulation layer tissue biases of the cancer types analysed by each driver gene detection method. We used a very strict background for enrichment studies as we used the whole human genome. Since human genes only map to 70% of human genes map to CATH domains, by using the whole human genome as the background rather than the smaller set of human genes mapping to CATH families, we effectively applied an overly stringent background. Any enrichment would have been more significant had we used the smaller background of human genes mapping to CATH as a background.

The driver genes were also analysed on their enrichment for cancer hallmarks using The Atlas of Cancer Signalling networks (ACSN). This showed that the putative driver genes from all the domain based methods were enriched in the two hallmark modules of survival and EMT cell motility. Specific analyses of MutFam driver genes within LGG and GBM revealed that both the common and unique MutFam genes reflect well known biological processes present in each cancer, such as processes implicated in growth factor signalling and P53 related events. GBM driver genes also affect advanced immune related processes, hypoxia, and angiogenesis. These processes reflect the more advanced nature of GBM relative to the more benign glioma subtype, LGG, thereby demonstrating the ability of the MutFam driver genes to capture the specific functional events within different stages of cancer.

Chapter 4

Structural and Stability Analyses of Cancer Mutations in FGFR3

Introduction

In contrast to the large scale analyses performed within chapters 2 and 3, In this chapter, various structural and functional analyses of cancer mutations within FGFR3 were conducted. Below is an introduction to FGFR3 kinase structure and function, followed by a summary of the work performed in this chapter.

Chapter 2 and 3 presented general trends detected by analyses of large scale data covering many protein families. This chapter reports the impacts of mutations in a specific family of interest, the Fibroblast Growth Factor Receptor(FGFR) family, which contains the FGFR1-4 receptor tyrosine kinases, involved in cell signalling pathways, which are heavily implicated in diseases such as cancer.

The FGFR receptor tyrosine kinases are located in the cell membrane, and are activated by ligand binding (fibroblast growth factor) to the extracellular domain leading to conformational changes within the receptor, that are transmitted to the intracellular kinase domains within the cytosol. Ligand binding results in activation of the kinase domains, which in turn results in their auto-phosphorylation and initiation of kinase activity leading to the phosphorylation of tyrosine residues on substrates, and recruitment of downstream adaptor proteins and signalling messengers. These act together to cascade the ligand binding event down to the transcription factors within the nucleus, engendering cellular growth.

Studies were performed on the FGFR kinase superfamily as mutations in these domains are implicated in bladder cancer, a disease being studied by our collaborator, Dr Matilda Katan Muller. Computational studies were performed to determine the distance of mutations from functional sites using the MutDist program. Clusters of mutations (MutClusters) were also identified within the FGFR kinase domain using the MutClust

programme. MutDist and MutClust are described in chapter 2. In addition to this, other bioinformatics tools were used, which measured the impacts of the mutations on protein stability, and structural and functional features. The results of these methods were compared to experimentally characterised effects on FGFR3 activation, measured by members of the Katan-Muller lab at UCL.

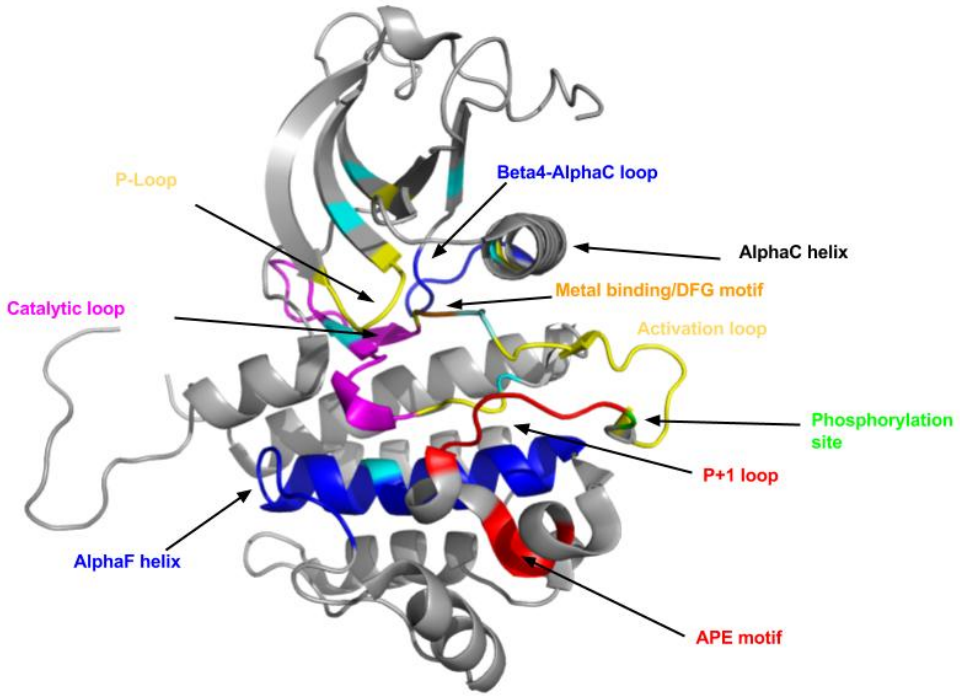
Structural and functional features of FGFR3 Kinase

All FGFR receptor kinases contain an intracellular kinase domain, which has a conserved structure between family members. This consists of 2 lobes, the C-lobe and N-lobe and the catalytic site resides in the cleft between them. Surrounding the central catalytic cavity, there are other residues which control the catalysis, regulate kinase activity, and provide a scaffolding role which supports the various elements in performing kinase activation. A summary of the main functional regions in FGFR kinase are described in a review by Taylor *et al* [242]. A PYMOL structure of FGFR3 and the main functions of the residues are given in figure 4.1.

One of the first steps in kinase activation is phosphorylation of a tyrosine residue in the activation loop in the C-lobe (A-loop). This event includes initiation of a hydrogen bonding network between residues within the activation loop and residues in the N-lobe, causing subsequent conformational shifts, including the outward extension of the A-loop, facilitating active state progression [117]. Such effects can be mimicked by mutations, referred to as phosphomimetic mutations, including the cancer mutation K650E in FGFR3, which triggers the formation of this hydrogen bonding cascade, promoting active state formation [98].

Establishment of the hydrogen bonding network within the activation loop also promotes the rotation of the upstream DFG motif, causing the inward flipping of the aspartate and glycine residues, and the outwards movement of phenylalanine away from the active site within the catalytic loop. The resulting conformation of the DFG motif is known as the ”DFG in” state and is conducive to kinase activation, since this enables metal binding by the aspartate residue, and increases accessibility to the active site,

Figure 4.1: Summarising the main functional regions of FGFR3, based on [242]. The FGFR3 structure is taken from the Protein Data Bank (PDBcode 4K33)



COLOUR	KINASE REGIONS	MAIN ROLE
BLUE	alpha-F helix (C-lobe) and beta-4/alphaC loop (N-lobe)	Scaffolding kinase elements/Molecular brake
RED	P+1 loop and APE motif	Substrate binding and anchoring kinase elements
YELLOW	P loop and activation loop	Phosphate coordination and catalysis
CYAN	hydrophobic residues spanning the lobes	Lobe coordination and active state stabilisation
MAGENTA	catalytic loop and gatekeeper residue	ATP binding
ORANGE	The D in the DFG motif	Metal binding
PINK	adjacent to the APE motif	Allosteric site
GREEN	between the activation loop and P+1 loop	Site for autophosphorylation

priming the active site for catalysis.

In the N-lobe, there is a β4-αC loop which contains residues which form hydrogen bond contacts within the loop and the adjacent αC helix, leading to the formation of the molecular brake and preventing the progression into the kinase active state [32]. This region regulates the 30 degree inward rotation of the downstream αC loop, towards the central catalytic core. Release of this molecular brake is evoked by activation loop phosphorylation, where the phosphate group withdraws the hydrogen bonds of the αC helix from the β4-αC loop, triggering the inwards rotation of the αC helix towards the catalytic cleft where it forms vital salt bridge contacts that support catalysis.

In addition to this, ATP binding takes place in the cleft upstream of the catalytic loop, and is regulated by an adjacent gatekeeper residue in the N lobe. The ATP phosphates are coordinated by residues within the adjacent P-loop in the N-lobe. The adenine ring of the ATP, along with the outwards facing DFG - phenyalanine in the C-lobe, forms a hydrophobic strip referred to as the C-spine, which spans the 2 kinase lobes across the catalytic site, further stabilising the active state [117].

After priming the kinase elements for catalysis, a substrate is docked onto the P+1 loop which is responsible for substrate binding, while also providing a scaffolding role for the kinase active site. The P+1 loop lies directly downstream of the activation loop, and is in turn supported by the C-lobe scaffolding helices including the αF helix. The αF helix is a scaffolding element for the catalytic core and is essential for anchoring various parts of the kinase elements for catalysis.

Mutations within or near to a functional region, can result in a loss or gain of kinase activity if the mutation affects functional site residues [83] [226]. Loss of function is easier to characterise since this can involve structural effects which also have an effect on protein stability - for example, a mutation causing a steric clash, impairing function. Furthermore, many mutations that are within functional sites often impair protein function since natural selection favours certain residue combinations that perform the functions efficiently.

However, there are cancer mutations which act via more subtle means, and are tol-

erated within the protein structure because they occur in residues under less negative selection pressures. These include mutations which cause a gain of function by altering the electrostatics of a functional site enhancing its catalytic efficiency, or by exerting a stabilising effect on the active form, thereby leading to an increase in activity.

It is important to characterise these mutations according to their effects, to provide insights into mechanisms altering protein function and to ultimately determine which therapeutic intervention will be most effective.

Methods for assessing changes in protein stability

There are many different methods to determine the impacts of mutation on protein stability, ranging from the use of empirical free energies, to use of probabilistic functions and graph based signatures. These are complementary approaches and were therefore used in parallel in the analyses of FGFR3 mutations.

FoldX was used for the empirical calculations reported in this chapter. This was developed by Schymkowitz *et al* in 2005 [208] and predicts protein stability using a simulated force-field in silico. FoldX calculates the free energy changes for each atom, based on a range of features to determine the effects on the stability of the whole protein [208]. These features include electrostatic energies, van der waals contacts, residue hydrophobicity, hydrogen bonding energies, and solvent interaction energies, and are derived from experimental data for each atom type. Other features include calculations related to the free energy changes on restricting backbone conformations, based on the permitted dihedral angles of adjacent amino acids pairs, and the permitted side chain angle conformations of each amino acid.

In addition to empirical methods, stability predictors based on probabilistic functions and graph based signatures were used. Site Directed Mutator (SDM) is a statistical potential method developed by Worth *et al* in 2011 [268], which considers the structural environment of the mutation and compares this with the wild type. This method captures the structural propensities of residues in the environments of the mutation, in the form of an environment-specific-substitution table (ESST) based on hydrogen bonding, sec-

ondary structure and solvent accessibility features. The ESST then gives a probability of an amino acid within an environment being substituted for another amino acid, which is proportional to its tolerability and inversely proportional to its potential effects on protein structure and stability [264] [130].

Like SDM, mCSM [185] captures the structural environment of a mutation, but via an alternative description using structural signatures generated as a graph representation of the mutated residue environment, accounting for geometric and physico-chemical patterns, using pharmocaphores. Instead of using probability functions, mCSM uses these structural signatures in a machine learning predictor to predict the effects of a mutation on protein stability. More recently the stability predictor, DUET, uses SDM in an optimised combination with mCSM [187]. DUET was used in the analysis of FGFR cancer associated mutations, in conjunction with FoldX.

Analysing the effects of mutations on protein structure and predicting mutation pathogenicity

Chapter 1 gave an overview of the many mutation predictors available, which use sequence based and structure based features to predict mutation pathogenicity. In the analyses presented in this chapter, a subset of mutation pathogenicity predictors (SAAPpred, CONDEL) [9][84] were used to study cancer associated mutations in FGFR3. One of the main advantages of SAAPpred compared to other analysis tools is that it explicitly reports the structural effects of a mutation within a protein. For example the impacts of the mutation on hydrogen bonding, and steric clashes. CONDEL was also used in this study, as it is currently regarded as one of the best meta-predictors of mutation pathogenicity, and uses a combination of sequence and structure based predictors including SIFT, Mutation Assessor, and PolyPhen-2. A more detailed summary of mutation predictors of pathogenicity are described in chapter 1.

Materials and methods

Cancer mutations in FGFR3

A set of 57 mutations within FGFR3 were obtained from the COSMIC and TCGA databases [15] [249] and analysed. 12 of these mutations had undergone preliminary analysis by our collaborator Dr Katan-Muller at UCL, and were suspected to result in a gain of function (GOF) effect on FGFR3, and so were referred to as suspected driver mutations. The experimental work included assays measuring FGFR3 auto-phosphorylation and substrate phosphorylation. Further experimental details are given in Patani *et al* [177]. Mutation clusters (MutClusters) in FGFR3 were identified using the MutClust method described in chapter 2.

Structures used in the analysis

The structure of the active FGFR3 form was used (PDB code 4K33). For the FoldX analysis, the PDB structure was stripped of its ions, ligands, solvent atoms, and all other non-mutated chains of the FGFR3 kinase were removed from the PDB file, so only the stability of the mutated chain A was measured.

The structure for the FGFR3 active form already contains mutations, including the known cancer mutation K650E, and two mutations made to facilitate protein purification S482C, and A582C. Therefore it was necessary to backmutate to give an unmutated structure, which was made using the PYMOL mutator plug-in [48]. There are no structures for active FGFR3 with CSA information. Therefore, for analysing the proximity of mutations to catalytic sites, a close homologous structure of active FGFR1 (83% sequence identity) was used which contains known catalytic residues from CSA (1AGW).

To assess the relative impacts of the mutations on FGFR3 protein stability, both the active and inactive forms were used. Since there is no structure of the FGFR3 inactive form, a 3D model was made based on the FGFR3 sequence and the FGFR1 inactive structure (PDB code 4UWY). This was generated using MODELLER [265] by Dr Nethaji Thiyagarajan in Dr Matilda Katan's group. Various tests measuring structure similarity

were used to assess for model quality, including 1) measuring BLAST sequence identity between the FGFR1 and FGFR3 sequences [137] 2)SSAP [173] 3) MolProbity [33] 4) ProsA-web [266] and 5) ModEval [184] as shown in table 4.1. All tests validated that the model was of good quality. SSAP was used to align the FGFR3 inactive model and the inactive FGFR1 structure. The alignment was used to generate a superposition of the structures by PROFIT, which uses the fitting algorithm developed by McLachlan *et al* [149].

Table 4.1: Summary of the model quality checks for the 3D model of the FGFR3 inactive form

Model tests	**Assesses**	**Value**	**Range**
BLAST	Sequence Equivalence	83%	Good Sequence identity >40% is good
SSAP	Structural Equivalence	98.54	SSAP score from 0-100 >80 is good
MolProbity	Sterio-chemistry and topology	1.98	Combined score of the clashes, rotomers, and ramachandran scores. Acceptable scores are those <3.
ProsA-Web	Statistical Energy Score	Z-score within range	A good z- score is within the Z-scores of PDBs included in the ProsA dataset.
ModEval	native likeness, template, compactness, RMSD	DOPE= -1.438 RMSD=1.735	Good DOPE score <-1 , Good RMSD score is <2 Å

Predicting mutation pathogenicity and reporting structural effects

To test for pathogenicity, two mutation predictors were used, CONDEL [84] and SAAP-pred [9]. The input for SAAPpred was the active FGFR3 structure, PDB code 4K33. Mutations were modelled using the PYMOL Mutator plug-in [48], and the mutated residue conformation with the fewest rotomeric clashes was selected. Within SAAPpred, there is an analysis pipeline, SAAPdap, that reports the structural and functional effects of a mutation on a protein, which are then incorporated and used in the predictor SAAPpred. Effects of the mutation include; effects on hydrogen bonding, salt bridges, charge and hydophobicity change, steric clash and void formation and co-location in functional and conserved sites.

Methods for analysing the impacts of the FGFR3 mutations on stability

FoldX Method: The FGFR3 active structure (4K33) and the FGFR3 inactive model were subject to energy minimization in FoldX, using the Repair PDB option, to alleviate any unfavourable clashes and interactions that may be present within the structure. To take into account the different rotameric conformations, three runs were executed. Mutations of each residue to every possible amino acid were made for that codon, using the PositionScan option in FOLDX [208]. This was to create a tolerance landscape of the protein, which was used as a background to compare with the FGFR3 cancer mutation energies.

To assess the effects of a mutation within a single protein structure, the energy differences for each mutation were calculated with respect to the wild type form, to determine the free energy change using equation 1, which is for an individual structure (FGFR3 active structure or FGFR3 inactive model).

Equation 1:

$$\delta\delta G(FreeEnergyChange)) = \delta G(mutant) - \delta G(wildtype)$$

To assess the effects of a mutation on the equilibrium between the two protein states, the relative stability of the inactive to the active form was calculated using equation 2. These calculations are based on those used in Hashimoto *et al* [95].

Equation 2:

$$\delta\delta G(Stability)) = \delta G(Inactive) - \delta G(Active)$$

The energy differences were categorised according to their effect relative to the wild type as follows; most stabilising (top 5%), highly stabilising (top 10%), stabilising (top 20%) and slightly stabilising (30%) and obtained by summing all the mutation energies in the structure. Stabilising effects are associated with a negative free energy difference.

DUET Method - Another measure of how a mutation effects protein stability was performed using DUET. Within DUET, SDM calculates a statistical measure which gives an indication of the impact on stability [187], based on the propensity of the native and

mutated residues to occur within a particular structural environment. A structural environment here is defined by secondary structure, hydrogen bonding, and solvent accessibility patterns. DUET combines SDM, and graph based signatures using mCSM. A more detailed description of this is provided within the Introduction section 4.1.2.

Results

Analysing the 57 cancer mutations from COSMIC by effects on protein structure, function, and stability

Analyses of mutation effects using SAAPdap, SAAPpred and CONDEL

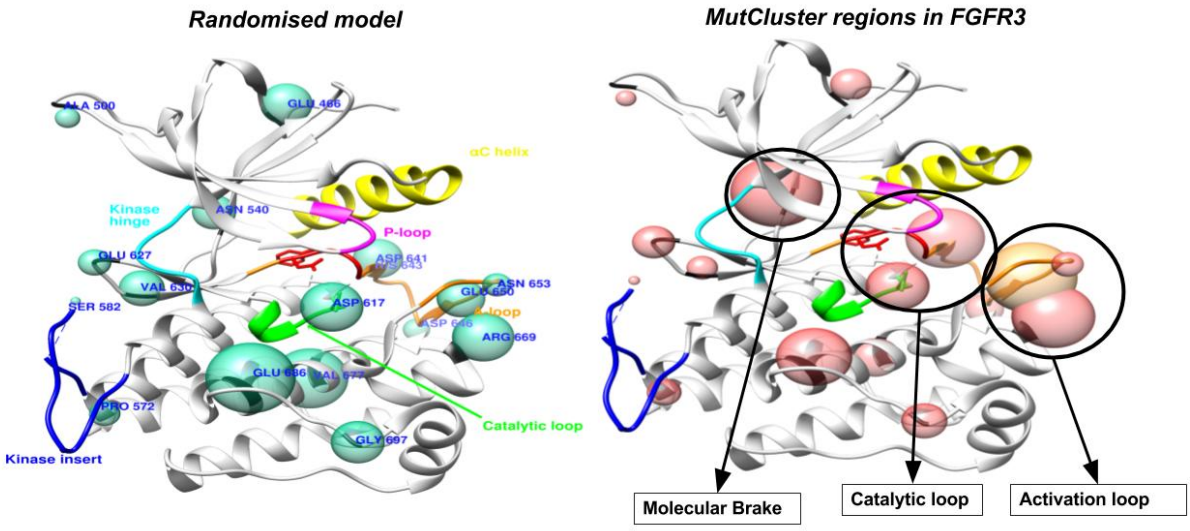
With regards to the predicted effects reported by SAAPdap analyses, 37/57 cancer mutations were associated with effects ranging from the occurrence in conserved sites, causing buried charges, and disturbing hydrogen bonding. The remaining 20 cancer mutations in FGFR3 showed no structural or functional effects reported by SAAPdap. For each of the 57 mutations in FGFR3, pathogenicity predictions were also made using CONDEL and SAAPpred in table 4.2. Only a few of the mutations are predicted to be pathogenic by both CONDEL and SAAPpred, which is interesting since these mutations are implicated in cancer and some have been experimentally characterised as activating. Only 9/57 mutations were predicted pathogenic by SAAPpred, 8/9 of these also showed significant pathogenicity scores reported by CONDEL. Most of the CONDEL effects related to the mutations being in conserved sites, consistent with its methodology in primarily assessing the impacts of mutations on conserved sites.

Identification of regions significantly enriched for cancer mutations in FGFR3 - MutClusters

Clusters of cancer mutations were identified using the MutClust program (see section 2.2.1.4). This was done to restrict the set of mutations to those most likely to be driving the cancer, and is a technique many studies have used to infer mutation pathogenicity. In addition to using the mutations in FGFR3, the MutClust programme also includes other cancer mutations taken from the FGFR relatives that belong to the same CATH FunFam as the FGFR3 kinase domain.

This analysis identified 3 significantly mutated clusters in the 3D structure of the FGFR3 kinase domain (circled in black in figure 4.2) compared to a random model, en-

Figure 4.2: The MutCluster regions identified within FGFR3 are located in the 3 main functional regions involved in kinase activity and regulation



capsulating 20/57 cancer mutations. The 3 MutClusters were located in the 3 main functional regions involved in kinase activity and its regulation. The first region includes the residues of the molecular brake and the pharmacologically relevant gatekeeper residue (V555) of the ATP binding site. The second region is within the catalytic loop, and the third is within the activation loop, as illustrated in figure 4.2.

Analysing the effects of cancer mutations on FGFR3 stabilty

Energy changes were calculated for all 57 cancer mutations by FoldX performed in the structure of the active FGFR3 form, and in the modelled inactive FGFR3 form, as shown in table 4.2. The comparison of energies between the 2 forms enabled insight into how the cancer mutations effected the equilibrium between the 2 states. From the 57 cancer mutations considered, it can be seen from table 4.2, that 26 of these have a stabilising effect on the FGFR3 active form relative to the inactive form. Interestingly, the cancer mutations which had no effects reported by SAAPpred, were predicted as having a stabilising effect by FoldX.

In summary, analyses of the 57 mutations by SAAPdap showed effects for 37 of the mutations. MutClust identified clusters comprising 20 mutations. 26 of the mutations had effects on stability as reported by FoldX, and 9 were predicted as pathogenic by SAAPpred. It can be seen from table 4.2 that 84% were associated with at least one characterised impact, and 58% with two impacts, and 42% with 3 impacts.

According to table 4.2, it can be seen 28 out of the 57 mutations are reported by both SAAPdap analyses and stability analyses using FoldX. Of these, 79% were shown to have a stabilising effect on the active FGFR3 form, and were associated with a range of SAAPdap effects. An example of this includes the mutation N540K, which exerts a very high stabilising effect on FGFR3, and is associated with effects including impacting hydrogen bonding. However, not all mutations affecting stability were associated with SAAPeffects, including K650E.In terms of mutations occurring in MutClusters, 8/19 were reported by SAAPdap to be near a conserved or functional site. For example, mutations at the D646N position are reported by SAAPdap to occur near a conserved site.

Table 4.2: Summary of the analyses for the 57 putative driver mutations in FGFR3. The FOLDX stability change is between the active and inactive FGFR3 forms.

Residue number	WT	MUT	Frequency (COSMIC)	MutCluster	Features	FoldX stability	CONDEL	SAAPdap effect	SAAPpred prediction
466	GLU	LYS	7			Neutral	PD	No structural effects identified	SNP
469	ARG	GLN	4			Destabilising - low	PD	No structural effects identified	SNP
490	GLU	GLY	2			Neutral	SNP	No structural effects identified	SNP
500	ALA	THR	1			Stabilising - low	SNP	No structural effects identified	SNP
505	VAL	ILE	1			Stabilising - high	PD	Conserved site	SNP
507	VAL	MET	3			Destabilising - low	PD	Clash—Conserved site	PD
538	ILE	PHE	4	1	Molecular brake	Stabilising - very high	PD	No structural effects identified	SNP
538	ILE	VAL	4	1	Molecular brake	Destabilising - low	SNP	No structural effects identified	SNP

Table 4.2 continued from previous page

Residue number	WT	MUT	Frequency (COSMIC)	MutCluster	Features	FoldX stability	CONDEL	SAAPdap effect	SAAPpred prediction
540	ASN	LYS	61	1	Molecular brake	Stabilising - very high	SNP	HBonds—Conserved site	SNP
540	ASN	SER	61	1	Molecular brake	Stabilising - low	SNP	Conserved site	SNP
555	VAL	MET	2		Gatekeeper	Stabilising - low	SNP	Conserved site	SNP
569	ALA	VAL	1			Neutral	SNP	No structural effects identified	SNP
572	PRO	ALA	1			Stabilising - medium	SNP	No structural effects identified	SNP
576	ASP	ASN	2			Neutral	SNP	No structural effects identified	SNP
582	CYS	PHE	4			Stabilising - high	SNP	Surface Phobic	SNP
603	ARG	GLN	6			Neutral	SNP	Conserved site	SNP
608	LEU	MET	2			Stabilising - medium	PD	Conserved site	SNP

Table 4.2 continued from previous page

Residue number	WT	MUT	Frequency (COSMIC)	MutCluster	Features	FoldX stability	CONDEL	SAAPdap effect	SAAPpred prediction
614	ILE	ASN	1	1	Catalytic loop	Destabilising - low	PD	Core Philic—Conserved site	PD
616	ARG	GLY	4	1	Catalytic loop,HRD motif	Stabilising - medium	PD	Buried Charge—HBonds—Conserved site	PD
617	ASP	GLY	1		Catalytic loop,HRD motif	Stabilising - very high	PD	Buried Charge—HBonds—SProtFT	PD
621	ARG	HIS	3		Catalytic loop	Destabilising - low	PD	HBonds—Conserved site	SNP
627	GLU	ASP	6			Neutral	SNP	No structural effects identified	SNP
627	GLU	GLY	6			Neutral	SNP	No structural effects identified	SNP
627	GLU	LYS	6			Neutral	SNP	No structural effects identified	SNP
627	GLU	VAL	6			Neutral	PD	Surface Phobic	SNP
630	VAL	ALA	2			Neutral	SNP	Conserved site	SNP

Table 4.2 continued from previous page

Residue number	WT	MUT	Frequency (COSMIC)	MutCluster	Features	FoldX stability	CONDEL	SAAPdap effect	SAAPpred prediction
630	VAL	MET	2			Neutral	SNP	Conserved site	SNP
636	PHE	LEU	1		Activation loop, DFG motif, Regulatory spine	Destabilising - low	PD	Conserved site	SNP
637	GLY	TRP	2		Activation loop, DFG motif	Stabilising - very high	PD	Conserved site	SNP
640	ARG	TRP	3	1	Activation loop	Destabilising - low	PD	Conserved site—Surface Phobic	PD
641	ASP	ASN	5	1	Activation loop	Neutral	SNP	No structural effects identified	SNP
641	ASP	GLY	5	1	Activation loop	Destabilising - low	SNP	HBonds	SNP
643	HIS	ARG	4		Activation loop	Stabilising - medium	SNP	No structural effects identified	SNP
643	HIS	ASP	4		Activation loop	Destabilising - low	SNP	No structural effects identified	SNP

Table 4.2 continued from previous page

Residue number	WT	MUT	Frequency (COSMIC)	MutCluster	Features	FoldX stability	CONDEL	SAAPdap effect	SAAPpred prediction
646	ASP	ASN	7	1	Activation loop	Stabilising - low	SNP	Conserved site	SNP
646	ASP	GLY	7	1	Activation loop	Stabilising - high	SNP	Conserved site	SNP
646	ASP	TYR	7	1	Activation loop	Stabilising - high	SNP	Conserved site	SNP
647	TYR	CYS	0		Activation loop, Phosphorylated tyrosines	Neutral	PD	Conserved site	SNP
650	LYS	ASN	210	1	Activation loop	Stabilising - very high	SNP	Buried Charge	SNP
650	LYS	GLN	210	1	Activation loop	Stabilising - very high	SNP	Buried Charge	SNP
650	LYS	GLU	210	1	Activation loop	Stabilising - very high	SNP	No structural effects identified	SNP
650	LYS	MET	210	1	Activation loop	Stabilising - very high	PD	Buried Charge—HBonds	PD

Table 4.2 continued from previous page

Residue number	WT	MUT	Frequency (COSMIC)	MutCluster	Features	FoldX stability	CONDEL	SAAPdap effect	SAAPpred prediction
650	LYS	THR	210	1	Activation loop	Stabilising - medium	SNP	Buried Charge	SNP
653	ASN	HIS	2		Activation loop	Stabilising - low	SNP	Conserved site	SNP
653	ASN	SER	2		Activation loop	Neutral	SNP	Conserved site	SNP
669	ARG	GLN	5	1		Neutral	SNP	No structural effects identified	SNP
669	ARG	GLY	5	1		Neutral	SNP	No structural effects identified	SNP
677	VAL	ILE	4			Neutral	SNP	Conserved site	SNP
679	SER	PHE	2			Stabilising - high	PD	HBonds—Conserved site	PD
686	GLU	LYS	4			Stabilising - low	PD	Buried Charge—HBonds—Conserved site	PD
689	THR	MET	1			Destabilising - low	PD	HBonds	SNP

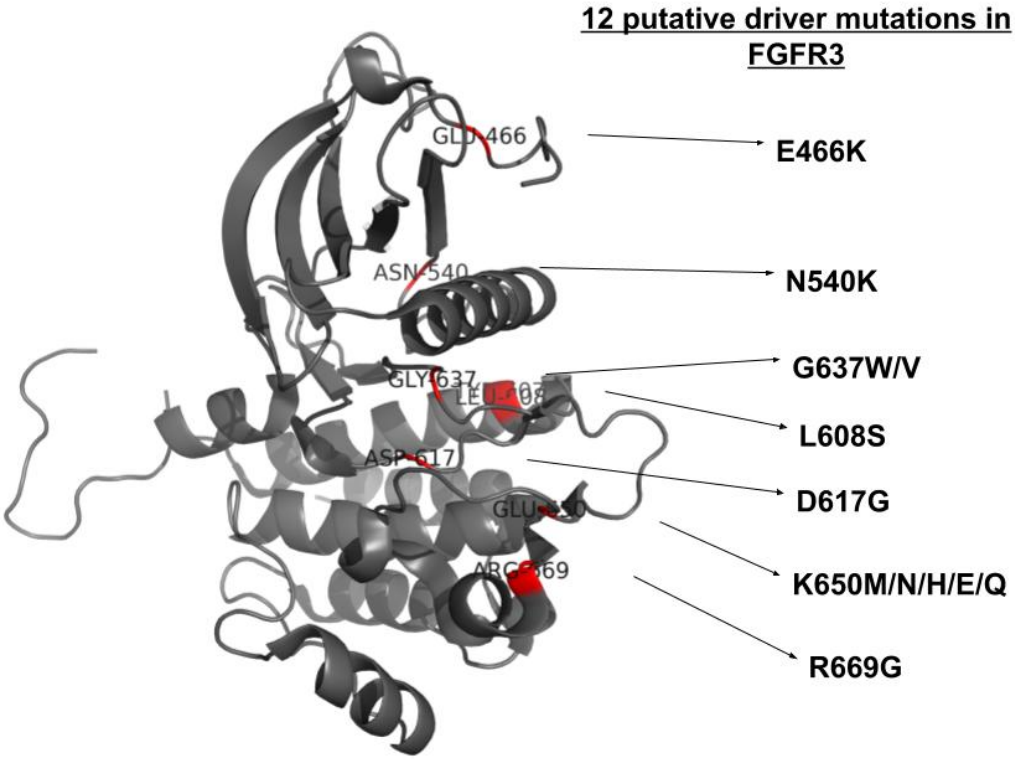
Table 4.2 continued from previous page

Residue number	WT	MUT	Frequency (COSMIC)	MutCluster	Features	FoldX stability	CONDEL	SAAPdap effect	SAAPpred prediction
696	PRO	LEU	1				PD	Surface Phobic	SNP
697	GLY	CYS	47			Destabilising - low	PD	Conserved site	SNP
700	VAL	ALA	1			Neutral	SNP	No structural effects identified	SNP
715	LYS	MET	2			Stabilising - low	SNP	HBonds—Conserved site	SNP
716	PRO	HIS	4			Stabilising - very high	PD	Buried Charge—Core Philic	PD
725	MET	ILE	1			Neutral	SNP	No structural effects identified	SNP

A more detailed analysis of the structural and functional effects of the 12 putative driver mutations within FGFR3

From the 57 cancer mutations analysed in FGFR3, 12 mutations were selected as suspected drivers, because they were shown to be activating by experimental assays, which are shown in the FGFR3 structure in figure 4.3. In addition to the SAAPdap, MutCluster, and FoldX analyses that had already been conducted on these, they were also analysed on their proximity to known catalytic and protein-protein interaction sites, their effects on the stability of the local protein environment using DUET, and their effects on the folding rate of the FGFR3 active form using FoldingRaCe. A more detailed discussion of the analyses conducted for these 12 mutations is given below - summarising both the previous studies and the additional studies undertaken for these 12.

Figure 4.3: The 12 putative driver mutations within the FGFR3 active structure. The mutated positions are highlighted in red



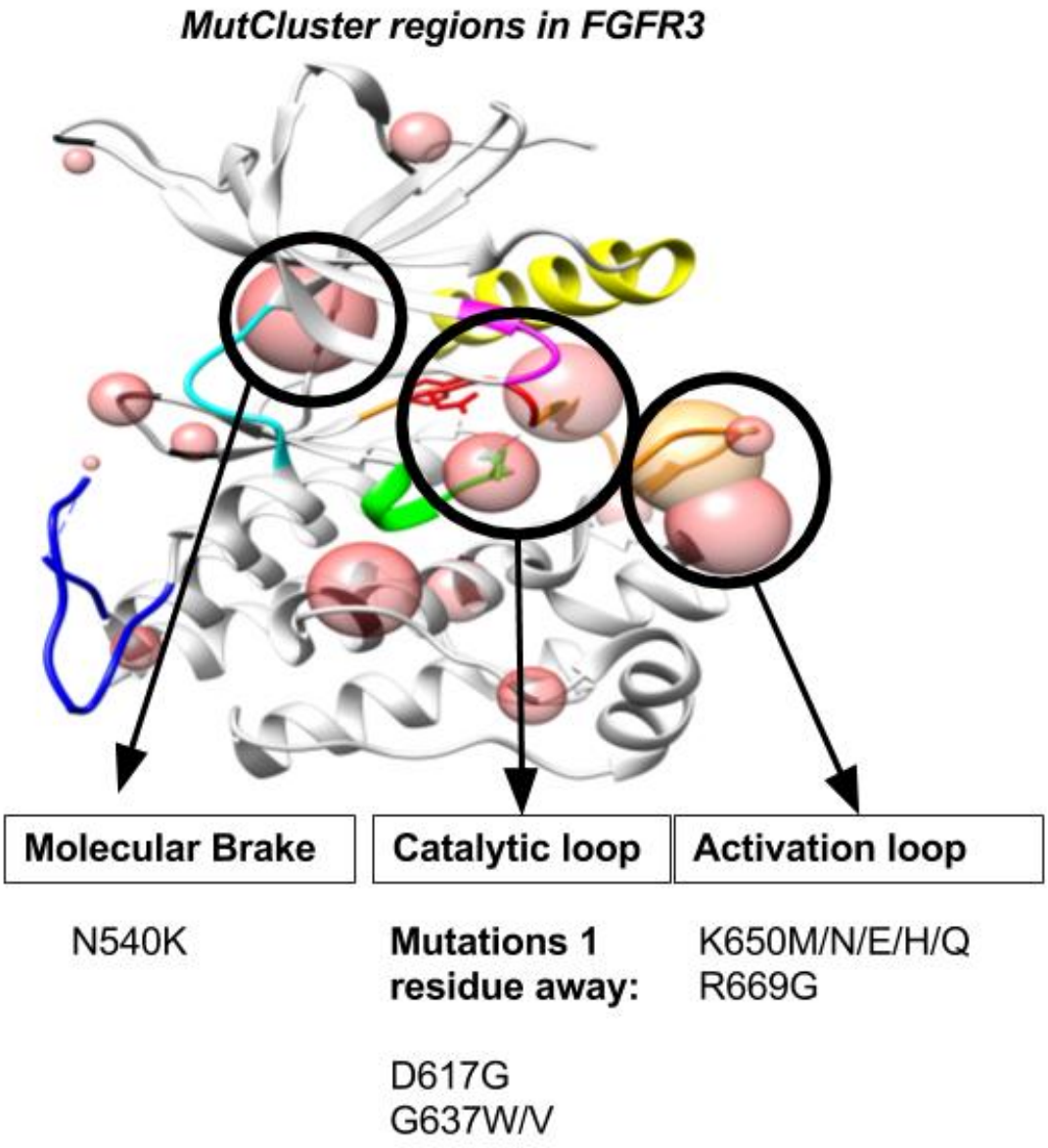
Mutation pathogenicity

Table 4.4 shows that 4/12 putative driver mutations were predicted pathogenic according to CONDEL – specifically the mutated positions L608S, G637V/W, and D617G. Although SAAPdap analysis reported these mutations as occurring within a conserved site, only the D617G was also predicted as pathogenic by SAAPpred, since this also disrupted hydrogen bonding and caused buried charge. In terms of their experimental effects, all 4 of these mutations were shown to decrease both auto-phosphorylation of FGFR3, consistent with their predicted pathogenicity. However, the remaining putative driver mutations at positions N540K, K650M/N/H/E/Q, and R669G were not predicted as pathogenic by either CONDEL or SAAPpred.

Colocation analysis to identified MutCluster regions in FGFR3

The MutClust analysis had identified 3 main regions enriched for cancer somatic mutations in FGFR3 compared to a random permutation (see figure 4.4). 7/12 of the driver mutations occurred within the clusters, while another 3 mutations were 1 residue away in the sequence from the MutCluster residues, as shown in table 4.4. As well as capturing well characterised cancer driver mutations that have a high mutation frequency at the position, such as K650 (210 mutations in COSMIC), the MutClusters also include rare mutations such as those at position R669 (5 mutations in COSMIC), which were also shown experimentally to be activating.

Figure 4.4: The occurrence of 10 of 12 putative driver mutations in or near the MutClusters identified for the FGFR3 kinase domain. MutClusters were identified using all COSMIC mutations for the FGFR3 kinase FunFam



Analysing the effects of the 12 putative driver mutations on FGFR3 stability

FOLDX stability analyses reported that all but one mutation exerted a stabilising effect on the FGFR3 active form, shown in table 4.4. The exception - R669G - had a neutral effect.

Analysing the effects of the 12 putative driver mutations on the folding rate of FGFR3

It has been proposed that fitness of a protein is governed by both its stability and concentration within a cell [207]. Therefore as well as analysing the stability equilibrium between inactive and active states using FoldX, FoldingRaCe was used to analyse the effects of cancer mutations on the folding rate of FGFR3 [31], to provide a proxy for the amount of folded FGFR3 present in the cell. This is based on the logic that if a mutation increases the folding rate, this would in turn increase the amount of folded protein, and hence concentration available to transition to the active state.

FoldingRaCe assesses how a mutation effects the rate of folding of a protein [31]. It applies a knowledge based technology which uses information from the single point mutation in the protein folding database, for proteins with 2 states. This was performed on the 12 putative driver mutations shown in table 4.4. The input for this method was the active FGFR3 structure and the residue mutation.

It can be seen in table 4.3, that some mutants cause an increase in folding rate, which could contribute to their activating effects in vitro. One mutation, N540K, which was found to be stabilising by FoldX was also found to have an accelerating effect on folding rate and also found to have an activation effect experimentally. Similarly, E466K also has an accelerating effect on folding rate. Further studies would be needed to validate this. The FoldingRaCe results suggest that some mutations decrease folding and therefore the effective concentration of the FGFR3 form, and may contribute to loss of kinase function in the mutations L608S, G637W and D617G.

Table 4.3: The effects of mutations on the folding rate of FGFR3, measured using FoldingRaCe. The increased folding rates are in bold

Residue_label	Native	Mutant	Folding RaCe Score: logarithmic rate of folding (/s)
608	L	S	-1.67
637	G	W/V	-3.66
617	D	G	-2.18
540	N	K	**5.68**
650	K	M	**0.32**
650	K	N	-1.88
650	K	E	-2.22
669	R	G	-0.83
466	E	K	**3.4**

Proximity analysis to catalytic sites and protein-protein interaction sites

A proximity analysis of the putative driver mutations to known functional sites (CSA and IBIS-PPI sites) was performed using the MutDist method (chapter 2, section 2.2.2) The 12 putative drivers were analysed for their proximity to functional sites,including catalytic residues from the Catalytic Site Atlas (CSA) or protein-protein interface residues taken from IBIS (Inferred Biological Interactions Server) database. This was performed using the MutDist method which calculates the closest atomic distance between the mutation and functional site residue. It was found that all but one of the mutations were within an IBIS protein-protein interaction site and the remaining mutation lay 5Å away from an IBIS-PPI site, as shown in table 4.4. This is consistent with the role of the FGFR3 kinase, since it is heavily involved in cellular signalling, and mutations within these residues may have effects on protein recruitment affinity and specificity [37] [169]. In contrast, only 2 mutations co-located to a catalytic residue, G617V and G617W, which caused a decrease and abolishment of FGFR3 activity respectively. Table 4.4 summarises the various analyses conducted on the 12 putative driver mutations. More detailed discussion is given below for each of the mutations.

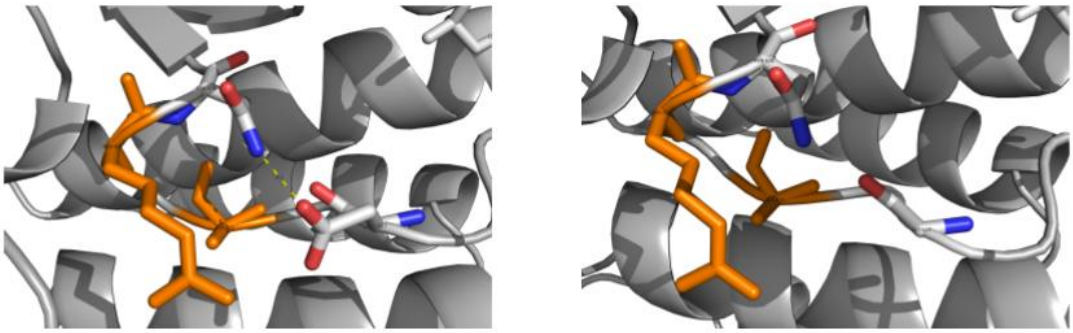
Table 4.4: Summary of the analyses for the 12 putative driver mutations in FGFR3. CONDEL and SAAPpred predictions of pathogenic (PD) and neutral (SNP) mutations are shown. For the proximity analysis to known functional sites, a mutation is annotated if it is within 5 Åto either a CSA or IBIS protein protein site.

Mutation	Frequency (COSMIC)	MutCluster	Kinase functional feature	FoldX stabilty effect	SAAPdap effect	SAAPpred prediction	CONDEL prediction	Within 5 Angstroms to functional sites (CSA and IBIS)
E466K	7	NO		Neutral	No structural effects identified	SNP	PD	IBIS
N540K	61	YES	Molecular brake	Stabilising - very high	HBonds—Conserved site	SNP	SNP	IBIS
L608S	2	NO		Stabilising - medium	Conserved site	SNP	SNP	IBIS
D617G	1	NO	Catalytic loop	Stabilising - very high	Buried Charge—HBonds—SProtFT	PD	PD	IBIS
G637V	2	NO	Activation loop, DFG motif	Stabilising - very high	Conserved site	PD	PD	IBIS/CSA
G637W	2	NO	Activation loop, DFG motif	Stabilising - very high	Conserved site	SNP	PD	IBIS/CSA
K650N	210	YES	Activation loop	Stabilising - very high	Buried Charge	SNP	SNP	IBIS
K650Q	210	YES	Activation loop	Stabilising - very high	Buried Charge	SNP	SNP	IBIS
K650H	210	YES	Activation loop	Stabilising - very high	No structural effects identified	SNP	SNP	IBIS
K650M	210	YES	Activation loop	Stabilising - very high	Buried Charge—HBonds	PD	SNP	IBIS
K650E	210	YES	Activation loop	Stabilising - medium	Buried Charge	SNP	SNP	IBIS
R669G	5	YES		Neutral	No structural effects identified	SNP	SNP	IBIS

Discussion of the possible impacts of the 12 putative driver mutations

D617G: According to the effects reported by SAAPdap, the predicted pathogenic mutant D617G affects a hydrogen bond and disrupts a buried charge, as shown in figure 4.5. It is also co-located to an active site residue listed in SwissProt, which agrees with the MutDist analysis showing co-location to a catalytic residue in CSA (see table 4.4). Since polar residues commonly mediate active site chemistry, loss of charge at this position will likely have an impact on function.

Figure 4.5: The predicted pathogenic mutation, D617G, loses a hydrogen bond (shown as a yellow dashed line) in the catalytic site. Images were made in PYMOL, using the Mutator tool.



L608S: The L608S mutation occurs within an alpha helix in the C-lobe, and is not co-located to a known catalytic residue or an IBIS interaction site. Nevertheless, this residue was reported as conserved by SAAPdap analysis. This mutation may have an effect on the scaffolding and stabilisation of the active site, much like the effect of the adjacent alphaF helix (see figure 4.1). Another possible explanation is that the mutation is affecting protein-protein binding site since this mutation lies 4.49Å from an IBIS-PPI site.

G637V/W: SAAPdap analysis reports that the G637 residue is a conserved site. G637 is within the C-lobe and forms part of the conserved DFG motif within protein kinases (see figure 4.1). The inherent flexibility of the native glycine residue is known to be important in communicating conformational changes within the activation loop to

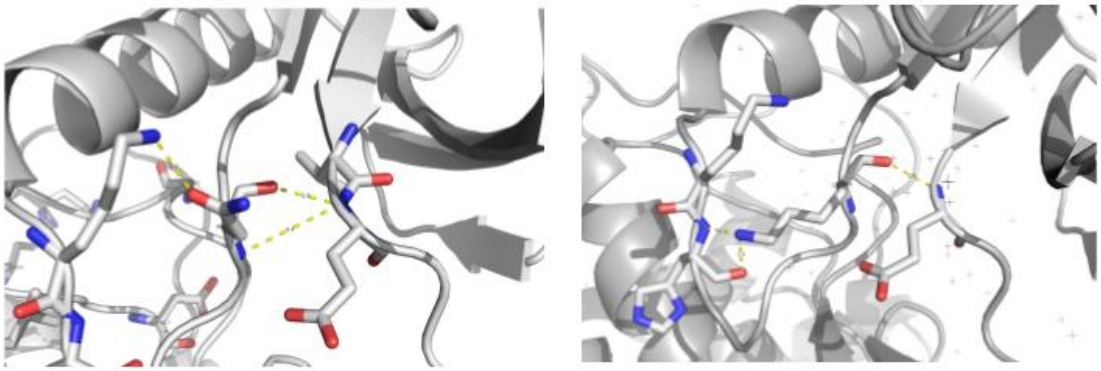
trigger rotation of the upstream phenylalanine and aspartate residues, enabling access to the catalytic site, and to the metal ion for additional co-ordination by other residues in the N-lobe. Therefore mutations to any amino acid at this residue, such as the observed mutants G637W/V, would inhibit this rotatory movement of the glycine, resulting in the phenylalanine lying in an inwards conformation, and pointing into and blocking the active site, and preventing metal binding functions by the aspartate within the ATP binding site. This would impact on active kinase function [117]

N540K: The N540K mutation is also reported by SAAPdap analysis to be within a conserved region. It is within the β4-αC helix loop in the kinase N-lobe, known as the molecular brake region of the kinase (see figure 4.1) and involved in the functional coupling of the 2 kinase lobes and regulating the movement of the adjacent N-lobe αC helix during catalysis [250] [32] [242]. The native asparagine at this position forms a hydrogen-bond triad with a conserved glutamate in the adjacent αC helix in the N-lobe and a catalytic lysine in the catalytic site [156], restricting entry into the active state, shown in figure 4.6. Therefore, mutations that disturb or alter hydrogen bonding at this conserved residue triad in the molecular brake, would release the molecular brake and lead to kinase activation.

Closer examination of the N540K mutation in the β4-αC loop in the active FGFR3 structure, shows that the native residue makes 3 hydrogen bond contacts of 2.7Å, 4.4Å and 2.8Å to residues on the adjacent loop regions, and not to residues in the molecular brake triad. These distances are smaller when the residue is mutated to lysine i.e. 2.1Å, 2.4Å and 2.8Å respectively (see figure 4.6). This is consistent with the SAAPdap analysis report, that hydrogen bonding is affected (see table 4.4). The shorter bond lengths suggest that this mutation creates stronger hydrogen bonding contacts within this loop region, which would enhance the rigidity of this part of the structure, creating a more appropriate position for the αC helix to establish the active state. Other studies have also suggested that mutations which enhance the rigidity of the β4-αC helix loop are likely to facilitate active state formation, by regulating the position of the αC helix between an out

(inactive) and an in (active) conformation [55].

Figure 4.6: Hydrogen bonding contacts of the wild type and N540K mutant amino acid to adjacent loop regions. (left) Native N540, (right) Mutant K540. The N40 residue is the residue in the middle of the image. Images were made in PYMOL.

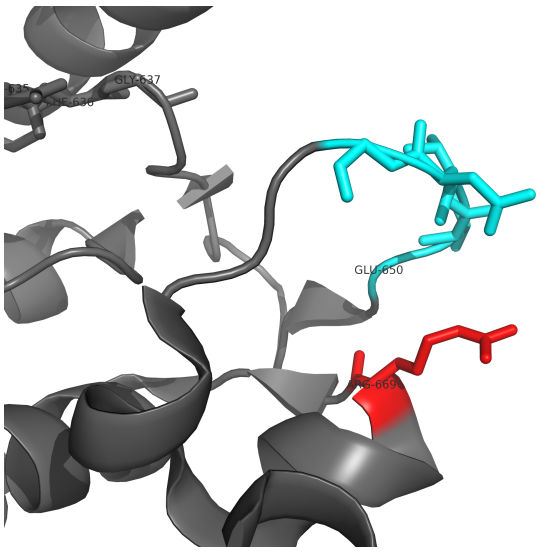


K650M/N/E/H/Q: Mutations at the position K650 were the only mutations predominantly affecting electrostatics, as mutations change the charge on the residue. The literature has reported that the residue K650 harbours many cancer mutations, which cause kinase activation [98], and is an established driver position in cancer. SAAPdap reports a disruption of hydrogen bonding and buried charge within the core respectively. K650E is close to the auto-phosphorylation site within the activation loop, a tyrosine, which when phosphorylated leads to kinase activation. As mentioned already, activation is caused by the phosphyl group forming hydrogen bond contacts with activation loop residues leading to a cascade of hydrogen bonding with residues along the activation loop, DFG motif, and catalytic site - priming residues for kinase activity [117] [98]. Since a mutation from lysine to glutamate would create a negative charge at this position, this would mimic the hydrogen bonding networks akin to those evoked by the phosphate group in the auto-phosphorylation site, leading to establishment of the active state and abnormal kinase activation.

E466K: The E466K mutant showed no reported effects by SAAPdap, but was co-located to an IBIS-PPI site according to the MutDist analysis.

R669G: The mutation at residue R669G, produced no effects reported by SAAPdap, nor is it predicted pathogenic by either CONDEL or SAAPpred, but it is nevertheless shown as experimentally activating. It does however occur within an identified MutCluster. The MutDist analysis also shows this mutation to occur within an IBIS-PPI site, directly downstream of the activation loop within the APE motif as shown in figure 4.7. Various studies have shown this region to be involved in allosteric signalling and such sites have been shown to be affected by cancer mutations [54] [242] [117], and so a possible mechanism of R669G pathogenicity is by having an effect on allosteric signalling. As regards stability, the mutation of Arginine to Glycine at position 669 had no effect on FGFR3 stability on the active form. Although, according to the experimental studies, this mutation activates the native FGFR3 active form. Therefore, it has been suggested that the stabilising effects of mutations at R669 are dependent on the presence of a phosphate group, an ATP molecule, or a ligand already bound to the kinase [177].

Figure 4.7: The R669G mutation (red) in FGFR3 lies just beneath the activation loop (cyan) within the APE motif



Furthermore, since there were discrepancies between the computational and experimental analyses, further structural analyses were carried out. The R669G mutation was studied using an alternative stability method, the predictor DUET, that combines SDM and mCSM [187]. These use probabilistic functions and graph based approaches respectively. DUET provides a more local stability measure, since it accounts for changes in the local structural environment of the mutation, rather than the global protein stability that FoldX considers. SDM reports the local stability of the mutated region by using the propensity of a given residue to occur with a given structural environment of secondary structure, solvent accessibility, and hydrogen bonding. It was hypothesised that, the subtle impacts of R669 mutations may act more locally, and could be overlooked by global stability measures. According to table 4.5, it can be seen that DUET-SDM predicts the R669 mutation (R669G) to exert a local stabilising effect in both the FGFR3 active and FGFR1 active structure, whereas the mCSM method predicted R669G to be slightly destabilising on the active form.

Table 4.5: Effects of the predicted pathogenic and activating mutations in FGFR3 active form.

Position	**Native**	**Mutant**	**SDM (Kcal/mol)**	**mCSM(Kcal/mol)**	**FoldX effect**
669	Arg	Gly	0.6	-1.28	Neutral

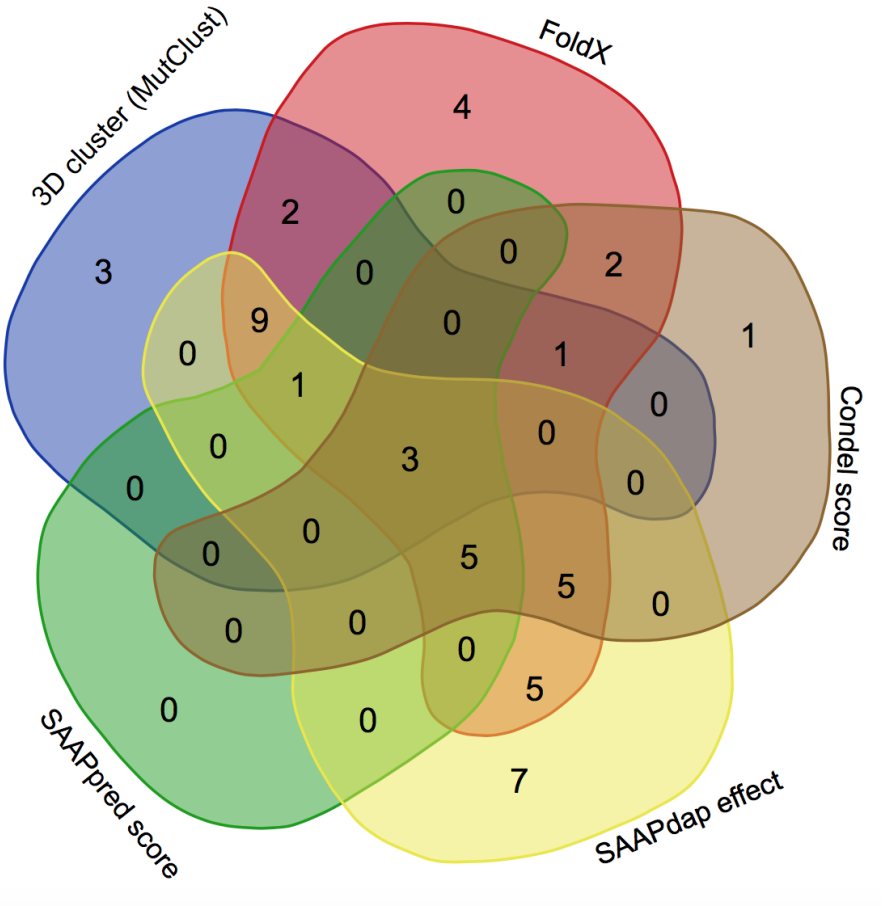
Conclusion

In this chapter, a comparative analysis of 57 cancer mutations in FGFR3 was performed, using both experimental and computational approaches. The use of both approaches enabled a more comprehensive characterisation of cancer mutations, highlighting the importance of complementary methods in mutation analysis.

The in-house program MutClust was run to identify regions enriched for cancer mutations in FGFR3, based on mutations from COSMIC within FGFR3 and other FGFR domains in the same functional family. This clustering resulted in the identification of 3 main regions - all of which coincided with functional regions known to be involved in kinase activity and regulation. These included a cluster within the molecular brake involved in regulation of the active state, the catalytic loop which performs the kinase chemistry, and thirdly the activation loop of the FGFR3 kinase. As well as detecting frequently mutated residues, the MutClusters also encapsulated rare cancer mutations known to cause activation, including the putative driver, R669G.

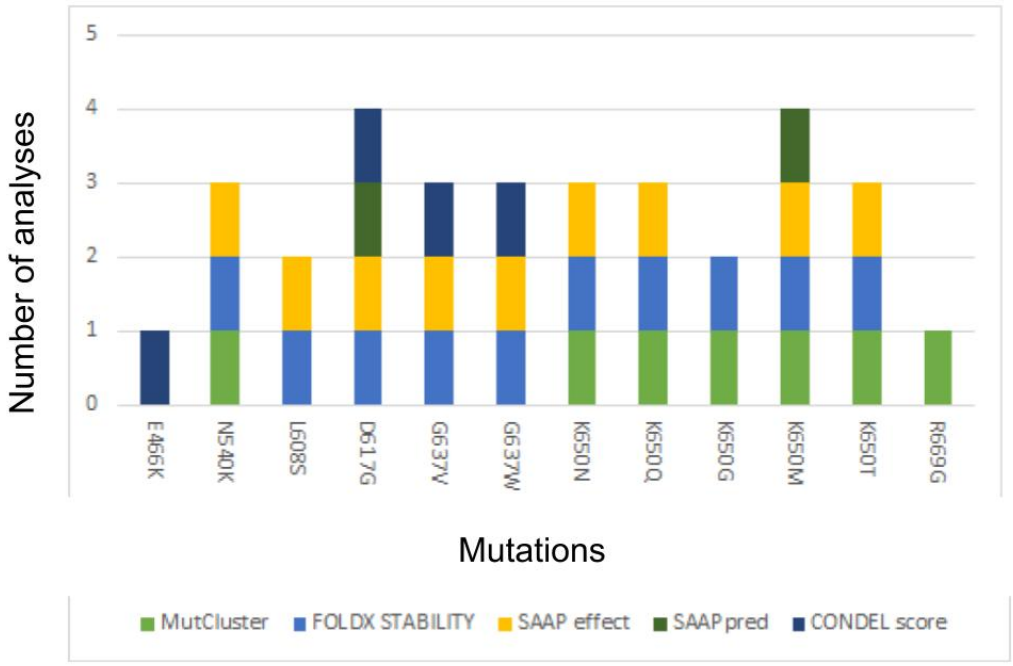
The effects of cancer mutations on FGFR3 stability were measured using FoldX analysis, which showed that 46% cancer mutations exerted a preferentially stabilising effect on the FGFR3 active form. These included the putative driver mutations that were within identified MutClusters within the activation loop and molecular brake region. In parallel to this, the structural effects of the mutations were analysed using SAAPdap, which reported 60% of the cancer mutations to have structural impacts, including the putative drivers K650E, R669G, and E466K. A summary of the 57 cancer mutations and their characterised effects are illustrated in figure 4.8, where 84% of mutations could be explained by at least 1 method. All mutations predicted by more than one method are putative drivers, but the more methods that predict pathogenic effects, the more confidence there is in it being a driver mutation.

Figure 4.8: Venn diagram summarising the analyses performed for the 57 cancer mutations in FGFR3.



A subset of mutations exerting an activating effect on FGFR3 in the experimental assays, were analysed in further detail - referred to as the putative driver mutations. It can be seen from figure 4.9 that all the putative drivers can be explained by at least 1 method.

Figure 4.9: The number of analyses reporting mutation pathogenicity and/pr structural and/or functional effects for the 12 putative driver mutations in FGFR3.



Further characterisation of the mutations effect on folding rate was analysed using FoldingRaCe, which provided another means of describing the pathogenicity of the putative drivers, N540K and E466K. Further analysis on the effects of folding rate could be undertaken using in vitro analyses. Such experiments could include performing a phi-value analysis [67], which measures the contribution of a mutation to protein folding. Alternatively, performing a hydrogen-deuterium exchange assay upon mutation would enable elucidation of the in vitro folding rate, using NMR spectroscopy.

In summary, where very detailed characterisation studies could be performed for 12 of the mutations, there were often multiple plausible contributions to pathogenicity.

Chapter 5

Conclusion

Cancer is a genetically heterogeneous disease, where many mutation events occur throughout the evolutionary trajectory, leading to the common phenotype of uncontrolled cellular growth. In order to filter out genetic noise within cancer, and select putative driver events from passenger mutations. In this thesis methods using protein domains from CATH were developed to prioritise more functional cancer mutations, and to identify driver genes.

Two main approaches were used to prioritise and filter cancer mutations, both of which use CATH FunFams. The first of which involves identifying statistically enriched CATH FunFam domains (MutFams) for a range of cancers, in order to detect driver genes. These were subject to biological process and cancer hallmark analysis to characterise their clinical relevance. The second approach detects driver mutations within these enriched MutFams using 3D clustering within the protein structure, referred to as MutClusters. These predicted driver mutations were then in turn compared to other disease mutations on their proximity to known and predicted functional sites.

In chapter 2, MutFams were identified based on the statistical enrichment of cancer missense mutations within CATH FunFams. Subsequently, significiantly enriched 3D clusters of mutations were identified (MutClusters). A large scale analysis of proximity to functional sites was then performed for the MutClusters and for mutations within germline non cancer cases, and mutations within a large cancer dataset from COSMIC. Overall, this showed that mutations in MutClusters showed a greater tendency to be proximal to both known and predicted functional sites compared to the unfiltered cancer mutations. In particular, the predicted FunSites showed the greatest enrichment for all disease causing mutations compared to other functional sites considered. Other functional sites which showed great mutation enrichment were protein-protein interaction sites from IBIS, in particular for the cancer mutations. Germline non cancer variants

however showed greater tendency to be proximal to betweenness centrality sites, which are more buried within the protein core.

For MutClusters, occuring in a FunFam with low information content, predicted FunSites could be inferred from a closely related FunFam.This analysis showed the functional relevance of MutCluster mutations, along with highlighting the use of CATH in predicting clinically relevant functional sites. Studies have shown greater benefits of 3D mutation clustering, for capturing mutations within both oncogenes and tumour suppressor genes, than using 1D hotspot methods. Therefore, further studies from this would be to analyse the MutClusters on their structural properties within tumour suppressor genes and oncogenes, such as solvent accessibility and stability, and compare this to unfiltered cancer and germline non cancer mutations. It would also be interesting to examine whether the putative driver genes identified in MutFams were closer together in a human protein network than expected from a random model. Our work suggested this as the genes tended to enrich in particular processes and pathways.

In chapter 3, the putative driver genes within 22 cancers were derived based on them containing a common MutFam domain, which were then analysed on their biologial and clinical relevance using Gene Ontology (GO) biological process enrichments, and enrichments in known cancer hallmarks from The Atlas of Cancer Signalling Networks (ACSN) respectively. Comparisons to other domain based methods - from Pfam - were also performed. This analysis showed convergence of the MutFam driver genes and Pfam derived genes on different steps of the same biological pathway - such as events involved in cellular development, and cell migration - many of which have been reported in the literature as being implicated in cancer, and are consistent with the enriched ACSN hallmarks. In cases, where there were different biological processes enriched, these unique pathways could be attributed to the tissue bias between the cancer types analysed by the Pfam and MutFam driver methods. Here the MutFam genes were associated with more processes seen in tissues from an ectodermic origin, and the Pfam genes were more associated with processes in tissues from a mesodermic origin.

In addition to analysing the MutFam driver genes across the 22 cancers, the MutFam driver genes were compared between two stages of glioma - low grade glioma (LGG) and glioblastoma multiforme (GBM). GO biological process enrichments showed that the MutFam driver genes captured the specific clinical phenotypes associated with the different stages of glioma, where the GBM driver genes were implicated in processes involved in immunity, angiogenesis, and proteostasis, which are more reminiscent of later stage carcinomas. Considerations for the future include analysing the biological pathways associated with individual cancers, to inspect any convergence of cancer specific MutFam genes on specific enriched cellular events. This in turn could highlight therapeutic redundancies, since many cancers could be targeted by considering a single pathway.

Chapter 4 presented a detailed study for the cancer mutations in FGFR3, implicated in bladder cancer. The use of computational approaches, in conjunction with experimental work, performed by collaborators, were complementary in characterising the structural and functional impacts of cancer mutations in FGFR3. Computational tools included analysing the effects on FGFR3 stability using FoldX, which reported that 37/57 of the analysed FGFR3 cancer mutations affected the stability of the active form of FGFR3, 70% of which exerted a stabilising effect on the FGFR3 active form. Other tools included analysing the specific structural and functional effects of mutations in the protein structure using SAAPdap. SAAPdap reported effects for 35/57 of the cancer mutations, ranging from impacting hydrogen bonding, and occurring in conserved sites to disturbing electrostatics. Predictions of mutation pathogenicity were also performed using CONDEL and SAAPpred, which showed 8 and 9 mutations to be pathogenic respectively.

In addition to known bioinformatic tools, an in-house method developed in chapter 2, MutClust, was used to identify driver mutations in FGFR3. The MutClust method detected 3 main clusters of cancer mutations in the FGFR3 kinase domain, which occurred within functional regions associated with the kinase activity; molecular brake, the catalytic loop, and the activation loop. As well as capturing high frequency mutants such as the known driver position K650, the MutClusters also included rare mutations which

were nevertheless shown as experimentally activating, such as the position R669. It was shown that the SAAPdap proved helpful in explaining how stability could be effected, and also in confirming the proximity of putative driver mutations to functional sites. Further work would be to study the putative driver mutations in FGFR3 on other experimentally derived features, such as the impacts on protein folding rate, using methods such as phi-value analysis and NMR spectroscopy.

Appendix A

Table A.1: Summary of the proximities of the MutCluster central residues to known and predicted functional sites. A cluster residues is annotated as having a 1 or 0 if it lies within or not within 5 Angsroms to a functional site respectively.

Superfamily ID	FunFam number	FunFam structure	Cluster residue	CSA	IBIS PPI	FunSite	Uniprot feature	Betweeness Centrality	STRESS surface
1.10.1520.10	4026	2eb1	B152	0	0	1	1	1	0
1.10.1520.10	4026	2eb1	B167	0	0	1	0	1	0
1.10.1520.10	4026	2eb1	B47	0	0	1	1	1	0
1.10.1540.10	596	1t77	A2342	0	0	1	0	1	1
1.10.1540.10	596	1t77	A2344	0	0	1	0	1	1
1.10.1540.10	596	1t77	A2353	0	0	1	0	1	1
1.10.1540.10	596	1t77	A2355	0	0	1	0	1	1
1.10.1540.10	596	1t77	A2392	0	0	1	0	1	1
1.10.196.10	1070	2gtp	C66	1	0	1	0	0	0
1.10.196.10	1070	2gtp	C67	1	0	1	0	0	0
1.10.196.10	1070	2gtp	C68	1	0	1	0	0	0
1.10.196.10	1070	2gtp	C70	1	0	1	0	0	0
1.10.220.60	71	1r4a	F2187	0	1	0	0	0	1

Table A.1 continued from previous page

Superfamily ID	FunFam number	FunFam structure	Cluster residue	CSA	IBIS PPI	FunSite	Uniprot feature	Betweeness Centrality	STRESS surface
1.10.220.60	71	1r4a	F2189	0	1	0	0	0	1
1.10.260.40	46829	1ic8	B125	0	0	1	0	0	1
1.10.418.10	4640	1dxx	C106	0	1	1	0	1	0
1.10.472.10	9821	2r7g	A755	0	1	1	0	0	1
1.10.510.10	78531	3ppj	A583	1	1	1	0	0	1
1.10.510.10	78531	3ppj	A589	1	1	1	0	0	1
1.10.510.10	78531	3ppj	A591	1	1	1	0	0	1
1.10.510.10	78531	3ppj	A593	1	1	1	0	0	1
1.10.510.10	78531	3ppj	A594	1	1	1	0	0	1
1.10.510.10	78531	3ppj	A595	1	1	1	0	0	1
1.10.510.10	78531	3ppj	A596	1	1	1	0	0	1
1.10.510.10	78531	3ppj	A615	1	1	1	0	0	1
1.10.510.10	79008	2ycf	A355	1	1	1	0	0	1
1.10.510.10	79008	2ycf	A392	1	1	1	0	0	1
1.10.510.10	79008	2ycf	A394	0	1	1	0	0	1

Table A.1 continued from previous page

Superfamily ID	FunFam number	FunFam structure	Cluster residue	CSA	IBIS PPI	FunSite	Uniprot feature	Betweeness Centrality	STRESS surface
1.10.510.10	79008	2ycf	A396	1	1	1	0	0	1
1.10.510.10	79298	3my0	I383	0	0	1	0	0	0
1.10.630.10	29426	3mdr	B143	0	0	1	0	1	0
1.10.630.10	29426	3mdr	B144	0	0	1	0	1	0
1.10.630.10	29426	3mdr	B244	0	0	1	0	1	0
1.10.630.10	29426	3mdr	B245	0	0	1	0	1	0
1.10.630.10	29426	3mdr	B248	0	0	1	0	1	0
1.10.630.10	29426	3mdr	B435	1	0	1	0	1	0
1.20.920.10	4440	3tlp	A491	0	1	1	0	0	1
1.20.920.10	4440	3tlp	A550	0	1	1	0	0	1
1.20.920.10	4440	3tlp	A551	0	1	1	0	0	1
1.20.920.10	4440	3tlp	A554	0	1	1	0	0	1
1.20.920.10	4440	3tlp	A588	0	1	1	0	0	1
1.20.920.10	4440	3tlp	A590	0	1	1	0	0	1
1.20.920.10	4440	3tlp	A591	0	1	1	0	0	1

Table A.1 continued from previous page

Superfamily ID	FunFam number	FunFam structure	Cluster residue	CSA	IBIS PPI	FunSite	Uniprot feature	Betweeness Centrality	STRESS surface
1.20.920.10	4440	3tlp	A594	0	1	1	0	0	1
2.10.110.10	5817	2xjz	B95	0	1	0	0	0	1
2.10.220.10	11249	1mox	A287	0	1	0	0	0	1
2.10.220.10	11249	1mox	A293	0	1	0	0	0	1
2.10.220.10	11249	1mox	A295	0	1	0	0	0	1
2.10.220.10	11249	1mox	A297	0	0	0	0	0	1
2.10.220.10	11249	1mox	A299	0	1	0	0	0	1
2.10.220.10	11249	1mox	A300	0	1	0	0	0	1
2.10.220.10	11249	1mox	A302	0	1	0	0	0	1
2.10.220.10	5834	1s78	A284	0	1	0	0	0	1
2.10.220.10	5834	1s78	A286	0	1	0	0	0	1
2.10.220.10	5834	1s78	A287	0	1	0	0	0	1
2.10.220.10	5834	1s78	A288	0	1	0	0	0	1
2.10.220.10	5834	1s78	A289	0	1	0	0	1	1
2.130.10.10	102894	2ovq	B2385	0	1	0	0	1	1

Table A.1 continued from previous page

Superfamily ID	FunFam number	FunFam structure	Cluster residue	CSA	IBIS PPI	FunSite	Uniprot feature	Betweeness Centrality	STRESS surface
2.130.10.10	102894	2ovq	B2391	0	1	0	0	1	1
2.130.10.10	102894	2ovq	B2399	0	1	0	0	1	1
2.130.10.10	102894	2ovq	B2400	0	1	0	0	1	1
2.130.10.10	102894	2ovq	B2420	0	1	0	0	1	1
2.130.10.10	102894	2ovq	B2423	0	1	0	0	1	1
2.130.10.10	102894	2ovq	B2424	0	1	0	0	1	1
2.130.10.10	102894	2ovq	B2425	0	1	0	0	1	1
2.130.10.10	102894	2ovq	B2426	0	1	0	0	0	1
2.130.10.10	102894	2ovq	B2436	0	1	0	0	1	1
2.130.10.10	102894	2ovq	B2437	0	1	0	0	1	1
2.130.10.10	102894	2ovq	B2438	0	1	0	0	1	1
2.130.10.10	102894	2ovq	B2441	0	1	0	0	1	1
2.130.10.10	102894	2ovq	B2443	0	1	0	0	1	1
2.130.10.10	102894	2ovq	B2460	0	1	0	0	1	1
2.130.10.10	102894	2ovq	B2462	0	1	0	0	1	1

Table A.1 continued from previous page

Superfamily ID	FunFam number	FunFam structure	Cluster residue	CSA	IBIS PPI	FunSite	Uniprot feature	Betweeness Centrality	STRESS surface
2.130.10.10	102894	2ovq	B2463	0	1	0	0	1	1
2.130.10.10	102894	2ovq	B2464	0	1	0	0	1	1
2.130.10.10	102894	2ovq	B2465	0	1	0	0	0	1
2.130.10.10	102894	2ovq	B2466	0	1	0	0	1	1
2.130.10.10	102894	2ovq	B2476	0	1	0	0	0	1
2.130.10.10	102894	2ovq	B2479	0	1	0	0	1	1
2.130.10.10	102894	2ovq	B2480	0	1	0	0	1	1
2.130.10.10	102894	2ovq	B2481	0	1	0	0	1	1
2.130.10.10	102894	2ovq	B2482	0	1	0	0	1	1
2.130.10.10	102894	2ovq	B2483	0	1	0	0	1	1
2.130.10.10	102894	2ovq	B2497	0	1	0	0	1	1
2.130.10.10	102894	2ovq	B2500	0	1	0	0	1	1
2.130.10.10	102894	2ovq	B2502	0	1	0	0	1	1
2.130.10.10	102894	2ovq	B2504	0	1	0	0	1	1
2.130.10.10	102894	2ovq	B2505	0	1	0	0	1	1

Table A.1 continued from previous page

Superfamily ID	FunFam number	FunFam structure	Cluster residue	CSA	IBIS PPI	FunSite	Uniprot feature	Betweeness Centrality	STRESS surface
2.130.10.10	102894	2ovq	B2518	0	1	0	0	1	1
2.130.10.10	102894	2ovq	B2542	0	1	0	0	1	1
2.130.10.10	102894	2ovq	B2543	0	1	0	0	1	1
2.130.10.10	102894	2ovq	B2544	0	1	0	0	1	1
2.130.10.10	102894	2ovq	B2580	0	1	0	0	1	1
2.130.10.10	103908	2uzx	B427	0	1	1	0	1	1
2.30.29.30	22238	2x18	B37	0	0	1	0	1	0
2.30.29.30	22238	2x18	B50	0	0	1	0	1	0
2.30.29.30	22238	2x18	B51	0	0	1	0	1	0
2.40.128.20	5800	3apu	A120	0	0	1	0	1	1
2.60.120.260	35252	3nru	D90	0	0	1	0	1	0
2.60.200.10	492	1dd1	A326	0	1	0	0	0	1
2.60.200.10	492	1dd1	A332	0	1	0	0	0	0
2.60.200.10	492	1dd1	A337	0	1	0	0	0	1
2.60.200.10	492	1dd1	A338	0	1	0	0	0	1

Table A.1 continued from previous page

Superfamily ID	FunFam number	FunFam structure	Cluster residue	CSA	IBIS PPI	FunSite	Uniprot feature	Betweeness Centrality	STRESS surface
2.60.200.10	492	1dd1	A352	0	1	0	0	0	1
2.60.200.10	492	1dd1	A355	0	1	0	0	0	1
2.60.200.10	492	1dd1	A492	0	1	0	0	0	0
2.60.200.10	492	1dd1	A496	0	1	0	0	0	0
2.60.200.10	492	1dd1	A504	0	1	0	1	0	1
2.60.200.10	492	1dd1	A505	0	1	0	1	0	1
2.60.200.10	492	1dd1	A506	0	1	0	1	0	1
2.60.200.10	492	1dd1	A524	0	1	0	1	0	1
2.60.200.10	492	1dd1	A533	0	1	0	0	0	0
2.60.200.10	492	1dd1	A536	0	1	0	0	0	0
2.60.200.10	492	1dd1	A537	0	1	0	0	0	0
2.60.210.10	2360	1d0a	A373	0	1	1	0	0	1
2.60.210.10	2360	1d0a	A374	0	1	1	0	0	1
2.60.40.10	134913	2edl	A9	0	1	1	0	1	1
2.60.40.720	232	3qyn	C195	0	1	1	0	0	1

Table A.1 continued from previous page

Superfamily ID	FunFam number	FunFam structure	Cluster residue	CSA	IBIS PPI	FunSite	Uniprot feature	Betweeness Centrality	STRESS surface
2.60.40.720	232	3qyn	C199	0	1	1	0	1	1
2.60.40.720	236	”1e50”	E146	0	0	1	0	1	0
2.60.40.720	236	”1e50”	E162	0	0	1	0	1	0
2.60.40.720	236	”1e50”	E164	0	0	1	0	1	0
2.60.40.720	236	”1e50”	E165	0	0	1	0	1	0
2.60.40.780	50	4b95	I117	0	0	1	0	1	0
2.60.40.780	50	4b95	I126	0	0		0	0	0
2.60.40.780	50	4b95	I127	0	0	1	0	0	0
3.10.320.10	2577	4gg6	B71	0	1	0	0	1	0
3.30.200.20	1872	2w96	B10	0	1	0	0	1	1
3.30.200.20	1872	2w96	B22	0	1	0	0	1	1
3.30.200.20	1872	2w96	B23	0	1	0	0	1	0
3.30.200.20	1872	2w96	B24	0	1	0	0	1	0
3.30.200.20	1872	2w96	B31	0	1	0	0	1	0
3.30.200.20	1872	2w96	B9	0	1	0	0	1	0

Table A.1 continued from previous page

Superfamily ID	FunFam number	FunFam structure	Cluster residue	CSA	IBIS PPI	FunSite	Uniprot feature	Betweeness Centrality	STRESS surface
3.30.200.20	3475	2rfn	B1111	0	0	0	0	0	0
3.30.200.20	3475	2rfn	B1112	0	0	0	0	0	0
3.30.200.20	64610	2vx3	B158	0	0	Psuedo Funsite	0	0	0
3.30.200.20	65851	2dyl	A168	0	1	1	0	0	1
3.30.200.20	65851	2dyl	A169	0	1	1	0	0	1
3.30.200.20	65851	2dyl	A170	0	1	1	0	0	1
3.30.200.20	65851	2dyl	A171	0	1	1	0	0	1
3.30.40.10	59876	1fbv	A396	0	1	0	0	0	1
3.30.40.10	59876	1fbv	A398	0	1	0	0	0	1
3.30.40.10	59876	1fbv	A400	0	1	0	0	0	1
3.30.40.10	59876	1fbv	A401	0	1	0	0	0	1
3.30.40.10	59876	1fbv	A404	0	1	0	0	0	1
3.30.40.10	59876	1fbv	A408	0	1	0	0	0	1
3.30.40.10	59876	1fbv	A416	0	1	0	0	0	1

Table A.1 continued from previous page

Superfamily ID	FunFam number	FunFam structure	Cluster residue	CSA	IBIS PPI	FunSite	Uniprot feature	Betweeness Centrality	STRESS surface
3.30.40.10	59876	1fbv	A417	0	1	0	0	0	1
3.30.40.10	59876	1fbv	A418	0	1	0	0	0	1
3.30.40.10	59876	1fbv	A419	0	1	0	0	0	1
3.30.40.10	59876	1fbv	A420	0	1	0	0	0	1
3.30.505.10	1631	3tl0	A61	0	1	0	0	1	1
3.30.505.10	1631	3tl0	A63	0	1	0	0	1	1
3.30.505.10	1631	3tl0	A71	0	1	0	0	1	1
3.30.505.10	1631	3tl0	A72	0	1	0	0	1	1
3.30.505.10	2646	2cr4	A25	0	1	0	0	1	1
3.30.505.10	2646	2cr4	A28	0	1	0	0	1	1
3.40.190.10	203444	1n84	A24	0	1	1	0	1	1
3.40.190.10	203444	1n84	A26	0	1	1	0	1	1
3.40.190.10	203444	1n84	A28	0	1	1	0	1	1
3.40.50.10140	4333	2js7	A22	0	0	1	0	1	1
3.40.50.10140	4333	2js7	A89	0	0	1	0	1	1

Table A.1 continued from previous page

Superfamily ID	FunFam number	FunFam structure	Cluster residue	CSA	IBIS PPI	FunSite	Uniprot feature	Betweeness Centrality	STRESS surface
3.40.50.300	630744	4dlt	A117	0	1	1	0	1	1
3.40.50.300	630744	4dlt	A12	0	1	1	0	1	1
3.40.50.300	630744	4dlt	A13	0	1	1	0	1	1
3.40.50.300	630744	4dlt	A14	0	1	1	0	1	1
3.40.50.300	630744	4dlt	A15	0	1	1	0	1	1
3.40.50.300	630744	4dlt	A61	0	1	1	0	1	1
3.80.10.10	105163	1xwd	C71	0	0	0	0	1	1
3.90.1290.10	1257	1lm7	A2364	0	1	1	0	0	1
3.90.1460.10	12	2dn4	A36	0	0	0	0	1	1
3.90.1460.10	12	2dn4	A38	0	0	0	0	1	1
3.90.1460.10	12	2dn4	A40	0	0	0	0	1	1
3.90.190.10	11605	3ezz	A221	0	0	1	0	0	1
3.90.70.10	17143	1fh0	B2175	0	0	1	0	0	0

References

[1] Gastrulation Figure Reference. 20, 159

[2] R. Abagyan, M. Totrov, and D. Kuznetsov. ICM - A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *Journal of Computational Chemistry*, 15(5):488–506, 1994. 56

[3] G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, and G. A. McVean. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73, 2010. 43

[4] L. A. Abriata, T. Palzkill, and M. Dal Peraro. How structural and physicochemical determinants shape sequence constraints in a functional enzyme. *PloS one*, 10(2):e0118684, 2015. 32

[5] M. S. Achary, A. B. M. Reddy, S. Chakrabarti, S. G. Panicker, A. K. Mandal, N. Ahmed, D. Balasubramanian, S. E. Hasnain, and H. A. Nagarajaram. Disease-causing mutations in proteins: structural analysis of the CYP1B1 mutations causing primary congenital glaucoma in humans. *Biophysical journal*, 91(12):4329–39, 2006. 34

[6] I. Adzhubei, D. M. Jordan, and S. R. Sunyaev. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. In *Current Protocols in Human Genetics*, volume Chapter 7, pages 7.20.1–7.20.41. 2013. 41

[7] J. Ahn, M. Urist, and C. Prives. The Chk2 protein kinase, 2004. 112

[8] B. J. Ainscough, M. Griffith, A. C. Coffman, A. H. Wagner, J. Kunisaki, M. N. Choudhary, J. F. McMichael, R. S. Fulton, R. K. Wilson, O. L. Griffith, and E. R. Mardis. DoCM: A database of curated mutations in cancer, 2016. 137

[9] N. S. Al-Numair and A. C. Martin. The SAAP pipeline and database: tools to analyze the impact and predict the pathogenicity of mutations. *BMC genomics*, 14 Suppl 3, 2013. 31, 36, 42, 184, 186

[10] J. Amberger, C. A. Bocchini, A. F. Scott, and A. Hamosh. McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic acids research*, 37(Database issue):D793—-796, jan 2009. 33, 40, 47, 48, 56, 81, 140

[11] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L.-S. L. Yeh. UniProt: the Universal Protein knowledgebase. *Nucleic acids research*, 32(Database issue):D115–D119, 2004. 39, 42, 47, 48, 50, 61, 81, 85, 133, 142

[12] D. Aran, A. Lasry, A. Zinger, M. Biton, E. Pikarsky, A. Hellman, A. J. Butte, and Y. Ben-Neriah. Widespread parainflammation in human cancer. *Genome biology*, 17(1):145, 2016. 172, 174

[13] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: Tool for the unification of biology, 2000. 52, 133

[14] H. Baeissa, G. Benstead-Hume, C. J. Richardson, and F. M. G. Pearl. Identification and analysis of mutational hotspots in oncogenes and tumour suppressors. *Oncotarget*, 8(13):21290–21304, 2017. 36, 72

[15] S. Bamford, E. Dawson, S. Forbes, J. Clements, R. Pettett, A. Dogan, A. Flanagan, J. Teague, P. A. Futreal, M. R. Stratton, and R. Wooster. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *British journal of cancer*, 91(2):355–8, 2004. 30, 33, 43, 48, 53, 64, 70, 77, 81, 83, 112, 137, 140, 142, 144, 167, 185

[16] A. Baresić, L. E. Hopcroft, H. H. Rogers, J. M. Hurst, and A. C. Martin. Compensated pathogenic deviations: analysis of structural effects. *Journal of molecular biology*, 396(1):19–30, feb 2010. 30, 63

[17] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. a. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V. Kryukov, D. Sonkin, A. Reddy, M. Liu, L. Murray, M. F. Berger, J. E. Monahan, P. Morais, J. Meltzer, A. Korejwa, J. Jané-Valbuena, F. a. Mapa, J. Thibault, E. Bric-Furlong, P. Raman, A. Shipway, I. H. Engels, J. Cheng, G. K. Yu, J. Yu, P. Aspesi, M. de Silva, K. Jagtap, M. D. Jones, L. Wang, C. Hatton, E. Palescandolo, S. Gupta, S. Mahan, C. Sougnez, R. C. Onofrio, T. Liefeld, L. MacConaill, W. Winckler, M. Reich, N. Li, J. P. Mesirov, S. B. Gabriel, G. Getz, K. Ardlie, V. Chan, V. E. Myer, B. L. Weber, J. Porter, M. Warmuth, P. Finan, J. L. Harris, M. Meyerson, T. R. Golub, M. P. Morrissey, W. R. Sellers, R. Schlegel, and L. a. Garraway. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–307, 2012. 37

[18] A. Baudot, F. X. Real, J. M. Izarzugaza, and A. Valencia. From cancer genomes to cancer models: bridging the gaps. *EMBO reports*, 10(4):359–366, apr 2009. 29

[19] G. Berx and F. van Roy. Involvement of members of the cadherin superfamily in cancer., 2009. 165

[20] M. J. Betts, Q. Lu, Y. Jiang, A. Drusko, O. Wichmann, M. Utz, I. a. Valtierra-Gutiérrez, M. Schlesner, N. Jaeger, D. T. Jones, S. Pfister, P. Lichter, R. Eils, R. Siebert, P. Bork, G. Apic, A.-C. Gavin, and R. B. Russell. Mechismo: predicting the mechanistic impact of mutations and modifications on molecular interactions. *Nucleic acids research*, 43(2):e10, 2015. 70

[21] M. Brylinski and J. Skolnick. FINDSITE-metal: Integrating evolutionary information and machine learning for structure-based metal-binding site prediction at the proteome level. *Proteins: Structure, Function and Bioinformatics*, 79(3):735–751, 2011. 59

[22] K. C. Bulusu, J. E. Tym, E. A. Coker, A. C. Schierz, and B. Al-Lazikani. canSAR: updated cancer research and drug discovery knowledgebase. *Nucleic acids research*, 42(Database issue):D1040—-D1047, jan 2014. 37

[23] T. D. Bunney, S. Wan, N. Thiyagarajan, L. Sutto, S. V. Williams, P. Ashford, H. Koss, M. A. Knowles, F. L. Gervasio, P. V. Coveney, and M. Katan. The Effect of Mutations on Drug Sensitivity and Kinase Activity of Fibroblast Growth Factor Receptors: A Combined Experimental and Theoretical Study. *EBioMedicine*, 2(3):194–204, 2015. 45

[24] B. M. Butler, Z. N. Gerek, S. Kumar, and S. B. Ozkan. Conformational dynamics of non-synonymous variants at protein interfaces reveals disease association. *Proteins*, dec 2014. 33, 62

[25] J. A. Capra and M. Singh. Characterization and prediction of residues determining protein functional specificity. *Bioinformatics*, 24(13):1473–1480, 2008. 52

[26] E. Capriotti and R. B. Altman. Improving the prediction of disease-related variants using protein three-dimensional structure. *BMC bioinformatics*, 12 Suppl 4:S3, 2011. 42

[27] M. Cargill, D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, N. Patil, N. Shaw, C. R. Lane, E. P. Lim, N. Kalyanaraman, J. Nemesh, L. Ziaugra, L. Friedland, A. Rolfe, J. Warrington, R. Lipshutz, G. Q. Daley, and E. S. Lander. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature genetics*, 22(3):231–238, jul 1999. 28

[28] H. Carter, S. Chen, L. Isik, S. Tyekucheva, V. E. Velculescu, K. W. Kinzler, B. Vogelstein, and R. Karchin. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer research*, 69(16):6660–7, 2009. 43, 71

[29] M. T. Chang, S. Asthana, S. P. Gao, B. H. Lee, J. S. Chapman, C. Kandoth, J. Gao, N. D. Socci, D. B. Solit, A. B. Olshen, N. Schultz, and B. S. Taylor. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nature biotechnology*, 34(2):1–11, 2015. 168

[30] D. Chasman and R. M. Adams. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *Journal of molecular biology*, 307(2):683–706, mar 2001. 30, 43

[31] P. Chaudhary, A. N. Naganathan, and M. M. Gromiha. Folding RaCe : A Robust Method for Predicting Changes in Pro- tein Folding Rates upon Point Mutations. *Bioinformatics*, pages 1–7, 2015. 202

[32] H. Chen, J. Ma, W. Li, A. V. Eliseenkova, C. Xu, T. A. Neubert, W. T. Miller, and M. Mohammadi. A Molecular Brake in the Kinase Hinge Region Regulates the Activity of Receptor Tyrosine Kinases. *Molecular Cell*, 27:717–730, 2007. 182, 206

[33] V. B. B. Chen, W. B. B. Arendall, J. J. J. Headd, R. J. Keedy, and R. R. D. C. MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography*, 66(1):12–21, 2010. 186

[34] O. H. Chen Chen, Natalia Gorlatova, Zvi Kelman. Structures of p63 DNA binding domain in complexes with half-site and with spacer-containing full response elements. *PNAS*, 108(16):6456–6461, 2011. 115

[35] D. Clarke, A. Sethi, S. Li, S. Kumar, R. W. F. Chang, J. Chen, and M. Gerstein. Identifying Allosteric Hotspots with Dynamics: Application to Inter- and Intra-species Conservation, 2015. 34, 50, 65, 87, 88, 109

[36] E. B. Claus, K. M. Walsh, J. K. Wiencke, A. M. Molinaro, J. L. Wiemels, J. M. Schildkraut, M. L. Bondy, M. Berger, R. Jenkins, and M. Wrensch. Survival and low-grade glioma: the emergence of genetic information. *Neurosurg Focus*, 38(1):1–10, 2015. 169

[37] P. Creixell, A. Palmeri, C. J. Miller, H. J. Lou, C. C. Santini, M. Nielsen, B. E. Turk, and R. Linding. Unmasking Determinants of Specificity in the Human Kinome. *Cell*, 163(1):187–201, sep 2015. 32, 39, 61, 62, 203

[38] A. L. Cuff, R. W. Janes, and A. C. R. Martin. Analysing the ability to retain sidechain hydrogen-bonds in mutant proteins. *Bioinformatics (Oxford, England)*, 22(12):1464–70, 2006. 30

[39] A. L. Cuff and A. C. Martin. Analysis of void volumes in proteins and application to stability of the p53 tumour suppressor protein. *Journal of molecular biology*, 344(5):1199–1209, dec 2004. 29

[40] S. Das, D. Lee, I. Sillitoe, N. L. Dawson, J. G. Lees, and C. A. Orengo. Functional classification of CATH superfamilies: a domain-based approach for protein function annotation. *Bioinformatics (Oxford, England)*, pages btv398–, 2015. 52

[41] S. Das, D. Lee, I. Sillitoe, N. L. Dawson, J. G. Lees, and C. A. Orengo. Functional classification of CATH superfamilies: a domain-based approach for protein function annotation. *Bioinformatics (Oxford, England)*, pages btv398–, 2015. 55, 87, 127

[42] A. David, R. Razali, M. N. Wass, and M. J. Sternberg. Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. *Human mutation*, 33(2):359–363, feb 2012. 47, 62, 95

[43] A. David and M. J. Sternberg. The contribution of missense mutations in core and Rim residues of protein-protein interfaces to human disease. *Journal of molecular biology*, 2015. 30, 33, 47, 62, 95

[44] J. P. Dawson, M. B. Berger, C.-C. Lin, J. Schlessinger, M. A. Lemmon, and K. M. Ferguson. Epidermal Growth Factor Receptor Dimerization and Activation Require Ligand-Induced Conformational Changes in the Dimer Interface. *MOLECULAR AND CELLULAR BIOLOGY*, 25(17):7734–7742, 2005. 116

[45] G. De Baets, L. Van Doorn, F. Rousseau, and J. Schymkowitz. Increased Aggregation Is More Frequently Associated to Human Disease-Associated Mutations Than to Neutral Polymorphisms. *PLoS computational biology*, 11(9):e1004374, sep 2015. 31

[46] E. C. De Bruin, N. McGranahan, R. Mitter, M. Salm, D. C. Wedge, L. Yates, M. Jamal-Hanjani, S. Shafi, N. Murugaesu, A. J. Rowan, E. Grönroos, M. A. Muhammad, S. Horswell, M. Gerlinger, I. Varela, D. Jones, J. Marshall, T. Voet, P. Van Loo, D. M. Rassl, R. C. Rintoul, S. M. Janes, S. M. Lee, M. Forster, T. Ahmad, D. Lawrence, M. Falzon, A. Capitanio, T. T. Harkins, C. C. Lee, W. Tom, E. Teefe, S. C. Chen, S. Begum, A. Rabinowitz, B. Phillimore, B. Spencer-Dene, G. Stamp, Z. Szallasi, N. Matthews, A. Stewart, P. Campbell, and C. Swanton. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science*, 346(6206):251–256, 2014. 148

[47] E. C. de Bruin, N. McGranahan, and C. Swanton. Analysis of intratumor heterogeneity unravels lung cancer evolution. *Molecular & Cellular Oncology*, 2(3):e985549, 2015. 148

[48] W. L. DeLano. The PyMOL Molecular Graphics System, Version 1.8. *Schrödinger LLC*, page http://www.pymol.org, 2014. 185, 186

[49] G. Dennis Jr, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, C. H. Lane, R. A. Lempicki, G. Dennis, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, and R. A. Lempicki. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology*, 4(5):P3, 2003. 136

[50] L. Di Paola, M. De Ruvo, P. Paci, D. Santoni, and A. Giuliani. Protein contact networks: An emerging paradigm in chemistry, 2013. 88

[51] F. Dietlein, L. Thelen, and H. C. Reinhardt. Cancer-specific defects in DNA repair pathways as targets for personalized therapeutic approaches, 2014. 171

[52] K. Ding and I. J. Kullo. Geographic differences in allele frequencies of susceptibility SNPs for cardiovascular disease. *BMC medical genetics*, 12, 2011. 28

[53] H. Dinkel, S. Michael, R. J. Weatheritt, N. E. Davey, K. Van Roey, B. Altenberg, G. Toedt, B. Uyar, M. Seiler, A. Budd, L. Jödicke, M. a. Dammert, C. Schroeter, M. Hammer, T. Schmidt, P. Jehl, C. McGuigan, M. Dymecka, C. Chica, K. Luck,

A. Via, A. Chatr-Aryamontri, N. Haslam, G. Grebnev, R. J. Edwards, M. O. Steinmetz, H. Meiselbach, F. Diella, and T. J. Gibson. ELM - The database of eukaryotic linear motifs. *Nucleic Acids Research*, 40(D1):1–10, 2012. 33, 34

[54] A. Dixit and G. M. Verkhivker. The energy landscape analysis of cancer mutations in protein kinases. *PloS one*, 6(10):e26071, 2011. 31, 34, 35, 64, 113, 208

[55] A. Dixit and G. M. Verkhivker. Structure-functional prediction and analysis of cancer mutation effects in protein kinases. *Computational and mathematical methods in medicine*, 2014:653487, 2014. 112, 207

[56] A. Dixit, L. Yi, R. Gowthaman, A. Torkamani, N. J. Schork, and G. M. Verkhivker. Sequence and structure signatures of cancer mutation hotspots in protein kinases. *PloS one*, 4(10):e7485, 2009. 31, 34, 113

[57] R. J. Dobson, P. B. Munroe, M. J. Caulfield, and M. A. A. Saqi. Predicting deleterious nsSNPs: an analysis of sequence and structural attributes. *BMC bioinformatics*, 7, 2006. 41

[58] G. P. Doss. Computational Methods in SNP Analysis. *Computational Methods in SNP Analysis*, 2014. 44

[59] Z. Dosztányi, V. Csizmok, P. Tompa, and I. Simon. IUPred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, 21(16):3433–3434, 2005. 33

[60] C. Douville, D. L. Masica, P. D. Stenson, D. N. Cooper, D. M. Gygax, R. Kim, M. Ryan, and R. Karchin. Assessing the Pathogenicity of Insertion and Deletion Variants with the Variant Effect Scoring Tool (VEST-Indel). *Human Mutation*, 37(1):28–35, 2016. 62

[61] H. B. Engin, J. F. Kreisberg, and H. Carter. Structure-Based analysis reveals cancer missense mutations target protein interaction interfaces. *PLoS ONE*, 11(4), 2016. 35, 62, 63, 65

[62] O. Espinosa, K. Mitsopoulos, J. Hakas, F. Pearl, and M. Zvelebil. Deriving a mutation index of carcinogenicity using protein structure and protein interfaces. *PloS one*, 9(1):e84598, 2014. 30, 43, 62, 63

[63] A. Fabregat, K. Sidiropoulos, P. Garapati, M. Gillespie, K. Hausmann, R. Haw, B. Jassal, S. Jupe, F. Korninger, S. McKay, L. Matthews, B. May, M. Milacic, K. Rothfels, V. Shamovsky, M. Webber, J. Weiser, M. Williams, G. Wu, L. Stein, H. Hermjakob, and P. D'Eustachio. The reactome pathway knowledgebase. *Nucleic Acids Research*, 44(D1):D481–D487, 2016. 53

[64] A.-M. Fernandez-Escamilla, F. Rousseau, J. W. H. Schymkowitz, and L. Serrano. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature biotechnology*, 22(10):1302–1306, 2004. 41

[65] M. Ferraiuolo, S. Di Agostino, G. Blandino, and S. Strano. Oncogenic Intra-p53 Family Member Interactions in Human Cancers. *Frontiers in oncology*, 6(March):77, 2016. 114

[66] C. Ferrer-Costa, M. Orozco, and X. de la Cruz. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *Journal of molecular biology*, 315(4):771–786, jan 2002. 29, 34

[67] A. R. Fersht and S. Sato. -Value analysis and the nature of protein-folding transition states. *Proceedings of the National Academy of Sciences*, 101(21):7976–7981, 2004. 212

[68] S. K. Fetics, H. Guterres, B. M. Kearney, G. Buhrman, B. Ma, R. Nussinov, and C. Mattos. Allosteric Effects of the Oncogenic RasQ61L Mutant on Raf-RBD. *Structure (London, England : 1993)*, 23(3):505–516, 2015. 34

[69] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. Sonnhammer, J. Tate, and M. Punta. Pfam: The protein families database, 2014. 51

[70] A. Fischer, C. Greenman, and V. Mustonen. Germline fitness-based scoring of cancer mutations. *Genetics*, 188(2):383–93, 2011. 29, 35

[71] S. A. Forbes, D. Beare, N. Bindal, S. Bamford, S. Ward, C. G. Cole, M. Jia, C. Kok, H. Boutselakis, T. De, Z. Sondka, L. Ponting, R. Stefancsik, B. Harsha, J. Tate, E. Dawson, S. Thompson, H. Jubb, and P. J. Campbell. COSMIC: High-resolution cancer genetics using the catalogue of somatic mutations in cancer. *Current Protocols in Human Genetics*, 2016:10.11.1–10.11.37, 2016. 48

[72] N. Furnham, G. L. Holliday, T. A. P. De Beer, J. O. B. Jacobsen, W. R. Pearson, and J. M. Thornton. The Catalytic Site Atlas 2.0: Cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Research*, 42(D1), 2014. 49

[73] N. Furnham, I. Sillitoe, G. L. Holliday, A. L. Cuff, S. A. Rahman, R. A. Laskowski, C. A. Orengo, and J. M. Thornton. FunTree: A resource for exploring the functional evolution of structurally defined enzyme superfamilies. *Nucleic Acids Research*, 40(D1), 2012. 124

[74] J. Gao, M. T. Chang, H. C. Johnsen, S. P. Gao, B. E. Sylvester, S. O. Sumer, H. Zhang, D. B. Solit, B. S. Taylor, N. Schultz, and C. Sander. 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. *Genome Medicine*, 9(1):4, 2017. 79

[75] M. Gao, H. Zhou, and J. Skolnick. Insights into Disease-Associated Mutations in the Human Proteome through Protein Structural Analysis. *Structure*, 23(7):1362–1369, jul 2015. 16, 59, 60, 81, 88, 89, 90, 92, 95, 124

[76] M. Gao, H. Zhou, and J. Skolnick. Insights into Disease-Associated Mutations in the Human Proteome through Protein Structural Analysis. *Structure*, 23(7):1362–1369, 2015. 47

[77] J. Garcia-Garcia, S. Schleker, J. Klein-Seetharaman, and B. Oliva. BIPS: BIANA Interolog Prediction Server. A tool for protein-protein interaction inference. *Nucleic Acids Research*, 40(W1), 2012. 141

[78] P. Gatti-Lafranconi and F. Hollfelder. Flexibility and Reactivity in Promiscuous Enzymes. *ChemBioChem*, 14(3):285–292, 2013. 32

[79] N. P. Gauthier, E. Reznik, J. Gao, S. O. Sumer, N. Schultz, C. Sander, and M. L. Miller. MutationAligner: a resource of recurrent mutation hotspots in protein domains in cancer. *Nucleic acids research*, pages 1–6, 2015. 69, 142

[80] L. Y. Geer, A. Marchler-Bauer, R. C. Geer, L. Han, J. He, S. He, C. Liu, W. Shi, and S. H. Bryant. The NCBI BioSystems database. *Nucleic Acids Research*, 38(SUPPL.1), 2009. 53

[81] B. Giardine, C. Riemer, T. Hefferon, D. Thomas, F. Hsu, J. Zielenski, Y. Sang, L. Elnitski, G. Cutting, H. Trumbower, A. Kern, R. Kuhn, G. P. Patrinos, J. Hughes, D. Higgs, D. Chui, C. Scriver, M. Phommarinh, S. K. Patnaik, O. Blumenfeld, B. Gottlieb, M. Vihinen, J. Väliaho, J. Kent, W. Miller, and R. C. Hardison. PhenCode: connecting ENCODE data with mutations and phenotype. *Human mutation*, 28(6):554–62, 2007. 42

[82] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821–6, 2002. 145

[83] S. Gong and T. L. Blundell. Structural and functional restraints on the occurrence of single amino acid variations in human proteins. *PloS one*, 5(2):e9186+, feb 2010. 29, 30, 35, 60, 63, 109, 123, 182

[84] A. González-Pérez and N. López-Bigas. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *American journal of human genetics*, 88(4):440–449, apr 2011. 39, 44, 184, 186

[85] A. Gonzalez-Perez and N. Lopez-Bigas. Functional impact bias reveals cancer drivers. *Nucleic Acids Research*, 40(21):1–10, 2012. 44

[86] A. Gonzalez-Perez, C. Perez-Llamas, J. Deu-Pons, D. Tamborero, M. P. Schroeder, A. Jene-Sanz, A. Santos, and N. Lopez-Bigas. IntOGen-mutations

identifies cancer drivers across tumor types. *Nature methods*, 10(11):1081–2, 2013. 44, 69

[87] C. A. P. J. B. A. Gonzalez-Perez A and F. LI. Uncovering disease mechanisms through network biology in the era of Next Generation Sequencing. *Sci Rep.*, 6(24570), 2016. 140, 141

[88] C. A. Gough, T. Gojobori, and T. Imanishi. Cancer-related mutations in BRCA1-BRCT cause long-range structural changes in protein-protein binding sites: a molecular dynamics study. *Proteins*, 66(1):69–86, 2007. 34

[89] J. Gough, K. Karplus, R. Hughey, and C. Chothia. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *Journal of molecular biology*, 313(4):903–19, 2001. 39

[90] K. O. Gress A, Ramensky V. Spatial distribution of disease-associated variants in three-dimensional structures of protein complexes. *Oncogenesis*, 6(9):e380, 2017. 60, 95

[91] C. J. R. Hanadi M. Baeissa, Graeme Benstead-Hume and F. M. Pearl. Mutational patterns in oncogenes and tumour suppressors. *Biochem Soc Trans*, 44(3):925–31, 2016. 38

[92] D. Hanahan and R. A. Weinberg. Hallmarks of cancer: The next generation, 2011. 19, 20, 114, 129, 130, 138, 157, 163, 165, 171, 172, 174

[93] G. C. Harburg and L. Hinck. Navigating breast cancer: Axon guidance molecules as breast cancer tumor suppressors and oncogenes. *Journal of Mammary Gland Biology and Neoplasia*, 16(3):257–270, 2011. 160

[94] S. Hashemi, A. Nowzari Dalini, A. Jalali, A. M. Banaei-Moghaddam, and Z. Razaghi-Moghadam. Cancerouspdomains: comprehensive analysis of cancer type-specific recurrent somatic mutations in proteins and domains. *BMC Bioinformatics*, 18(1):370, 2017. 74

[95] K. Hashimoto, I. B. Rogozin, and A. R. Panchenko. Oncogenic potential is related to activating effect of cancer single and double somatic mutations in receptor tyrosine kinases. *Human mutation*, 33(11):1566–75, 2012. 187

[96] A. Heravi-Moussavi and M. Anglesio. Recurrent Somatic DICER1 Mutations in Nonepithelial Ovarian Cancers. *New England Journal of Medicine*, 366(3):234–242, 2012. 119

[97] G. L. Holliday, C. Andreini, J. D. Fischer, S. A. Rahman, D. E. Almonacid, S. T. Williams, and W. R. Pearson. MACiE: exploring the diversity of biochemical reactions. *Nucleic Acids Research*, 40(Database issue):gkr799—-D789, nov 2011. 49

[98] Z. Huang, H. Chen, S. Blais, T. a. Neubert, X. Li, and M. Mohammadi. Structural mimicry of a-loop tyrosine phosphorylation by a pathogenic FGF receptor 3 mutation. *Structure (London, England : 1993)*, 21(10):1889–96, 2013. 33, 180, 207

[99] Z. Huang, L. Zhu, Y. Cao, G. Wu, X. Liu, Y. Chen, Q. Wang, T. Shi, Y. Zhao, Y. Wang, W. Li, Y. Li, H. Chen, G. Chen, and J. Zhang. ASD: a comprehensive database of allosteric proteins and modulators. *Nucleic acids research*, 39(Database issue):D663–9, 2011. 61

[100] P. Hulpiau and F. van Roy. Molecular evolution of the cadherin superfamily, 2009. 165

[101] J. M. Hurst, L. E. McMillan, C. T. Porter, J. Allen, A. Fakorede, and A. C. Martin. The SAAPdb web resource: a large-scale structural analysis of mutant proteins. *Human mutation*, 30(4):616–624, apr 2009. 36, 42

[102] Y. C. Hwang, C. F. Lin, O. Valladares, J. Malamon, P. P. Kuksa, Q. Zheng, B. D. Gregory, and L. S. Wang. HIPPIE: A high-throughput identification pipeline for promoter interacting enhancer elements. *Bioinformatics*, 31(8):1290–1292, 2015. 141

[103] L. M. Iakoucheva, C. J. Brown, J. D. Lawson, Z. Obradović, and a. K. Dunker. Intrinsic disorder in cell-signaling and cancer-associated proteins. *Journal of Molecular Biology*, 323(3):573–584, 2002. 33

[104] J. Izarzugaza, M. Vazquez, A. D. Pozo, and A. Valencia. wKinMut: An integrated tool for the analysis and interpretation of mutations in human protein kinases. *BMC Bioinformatics*, 14(1):345+, nov 2013. 38

[105] J. M. G. Izarzugaza, O. C. Redfern, C. A. Orengo, and A. Valencia. Cancer-associated mutations are preferentially distributed in protein kinase functional sites. *Proteins*, 77(4):892–903, dec 2009. 56, 112

[106] G. Jego, A. Hazoumé, R. Seigneuric, and C. Garrido. Targeting heat shock proteins in cancer. *Cancer Letters*, 332(2):275–285, 2013. 174

[107] D. M. Jordan, V. E. Ramensky, and S. R. Sunyaev. Human allelic variation: perspective from protein function, structure, and evolution. *Current opinion in structural biology*, 20(3):342–350, jun 2010. 29, 41

[108] P. R. Jose Lugo-Martinez, Vikas Pejaver, Kymberleigh A. Pagel, Shantanu Jain, Matthew Mort, David N. Cooper, Sean D. Mooney. The Loss and Gain of Functional Amino Acid Residues Is a Common Mechanism Causing Human Inherited Disease. *PLoS Comput Biol.*, 12(8), 2016. 61, 86, 95, 119

[109] Y. S. Ju, I. Martincorena, M. Gerstung, M. Petljak, L. B. Alexandrov, R. Rahbari, D. C. Wedge, H. R. Davies, M. Ramakrishna, A. Fullam, S. Martin, C. Alder, N. Patel, S. Gamble, S. O'Meara, D. D. Giri, T. Sauer, S. E. Pinder, C. A. Purdie, Å. Borg, H. Stunnenberg, M. van de Vijver, B. K. T. Tan, C. Caldas, A. Tutt, N. T. Ueno, L. J. van 't Veer, J. W. M. Martens, C. Sotiriou, S. Knappskog, P. N. Span, S. R. Lakhani, J. E. Eyfjörd, A.-L. Børresen-Dale, A. Richardson, A. M. Thompson, A. Viari, M. E. Hurles, S. Nik-Zainal, P. J. Campbell, and M. R. Stratton. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature*, 543(7647):714–718, 2017. 27

[110] A. Kamburov, M. S. Lawrence, P. Polak, I. Leshchiner, K. Lage, T. R. Golub, E. S. Lander, and G. Getz. Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proceedings of the National Academy of Sciences of the United States of America*, 112(40):E5486–95, sep 2015. 17, 77, 78, 125

[111] J. S. Kaminker, Y. Zhang, C. Watanabe, and Z. Zhang. CanPredict: a computational tool for predicting cancer-associated missense mutations. *Nucleic acids research*, 35(Web Server issue):W595—-W598, jul 2007. 43

[112] W. I. Karain and N. I. Qaraeen. Weighted protein residue networks based on joint recurrences between residues. *BMC bioinformatics*, 16(1):173, 2015. 88

[113] R. Karchin, M. Diekhans, L. Kelly, D. J. Thomas, U. Pieper, N. Eswar, D. Haussler, and A. Sali. LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics (Oxford, England)*, 21(12):2814–2820, jun 2005. 42

[114] S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, M. Duesbury, M. Dumousseau, M. Feuermann, U. Hinz, C. Jandrasits, R. C. Jimenez, J. Khadake, U. Mahadevan, P. Masson, I. Pedruzzi, E. Pfeiffenberger, P. Porras, A. Raghunath, B. Roechert, S. Orchard, and H. Hermjakob. The IntAct molecular interaction database in 2012. *Nucleic Acids Research*, 40(D1), 2012. 141

[115] K. Khafizov, M. V. Ivanov, O. V. Glazova, and S. P. Kovalenko. Computational approaches to study the effects of small genomic variations. *Journal of molecular modeling*, 21(10):251, oct 2015. 43

[116] C. C. Khor, F. O. Vannberg, S. J. Chapman, H. Guo, S. H. Wong, A. J. Walley, D. Vukcevic, A. Rautanen, T. C. Mills, K.-C. C. Chang, K.-M. M. Kam, A. C. Crampin, B. Ngwira, C.-C. C. Leung, C.-M. M. Tam, C.-Y. Y. Chan, J. J. Sung, W.-W. W. Yew, K.-Y. Y. Toh, S. K. Tay, D. Kwiatkowski, C. Lienhardt, T.-T. T. Hien, N. P. Day, N. Peshu, K. Marsh, K. Maitland, J. A. Scott, T. N. Williams, J. A. Berkley, S. Floyd, N. L. Tang, P. E. Fine, D. L. Goh, and A. V. Hill. CISH and susceptibility

to infectious diseases. *The New England journal of medicine*, 362(22):2092–2101, jun 2010. 28

[117] A. P. Kornev, N. M. Haste, S. S. Taylor, and L. T. F. Eyck. Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism. *Proceedings of the National Academy of Sciences of the United States of America*, 103(47):17783–17788, nov 2006. 112, 113, 180, 182, 206, 207, 208

[118] A. Kumar and R. Purohit. Use of long term molecular dynamics simulation in predicting cancer associated SNPs. *PLoS computational biology*, 10(4):e1003318+, apr 2014. 44

[119] N. Kumar and J. Skolnick. EFICAz2.5: Application of a high-precision enzyme function predictor to 396 proteomes. *Bioinformatics*, 28(20):2687–2688, 2012. 59

[120] I. Kuperstein, E. Bonnet, H.-A. Nguyen, D. Cohen, E. Viara, L. Grieco, S. Fourquet, L. Calzone, C. Russo, M. Kondratova, M. Dutreix, E. Barillot, and A. Zinovyev. Atlas of Cancer Signalling Network: a systems biology resource for integrative analysis of cancer data with Google Maps. *Oncogenesis*, 4(7):e160, 2015. 53, 127, 144, 167

[121] P. Lahiry, A. Torkamani, N. J. Schork, and R. A. Hegele. Kinase mutations in human disease: Interpreting genotype-phenotype relationships, 2010. 56

[122] M. J. Landrum, J. M. Lee, M. Benson, G. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, J. Hoover, W. Jang, K. Katz, M. Ovetsky, G. Riley, A. Sethi, R. Tully, R. Villamarin-Salomon, W. Rubinstein, and D. R. Maglott. ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44(D1):D862–D868, 2016. 137

[123] R. A. Laskowski. Enhancing the functional annotation of PDB structures in PDBsum using key figures extracted from the literature. *Bioinformatics*, 23(14):1824–1827, 2007. 78

[124] R. a. Laskowski, F. Gerick, and J. M. Thornton. The structural basis of allosteric regulation in proteins. *FEBS Letters*, 583(11):1692–1698, 2009. 34

[125] K. Laukens, S. Naulaerts, and W. V. Berghe. Bioinformatics approaches for the functional interpretation of protein lists: From ontology term enrichment to network analysis, 2015. 19, 130, 132

[126] M. S. Lawrence, P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. a. Roberts, A. Kiezun, P. S. Hammerman, A. McKenna, Y. Drier, L. Zou, A. H. Ramos, T. J. Pugh, N. Stransky, E. Helman, J. Kim, C. Sougnez, L. Ambrogio, E. Nickerson, E. Shefler, M. L. Cortés, D. Auclair, G. Saksena, D. Voet, M. Noble, D. DiCara, P. Lin, L. Lichtenstein, D. I. Heiman, T. Fennell, M. Imielinski, B. Hernandez, E. Hodis, S. Baca, A. M. Dulak, J. Lohr, D.-A. Landau, C. J. Wu, J. Melendez-Zajgla, A. Hidalgo-Miranda, A. Koren, S. a. McCarroll, J. Mora, R. S. Lee, B. Crompton, R. Onofrio, M. Parkin, W. Winckler, K. Ardlie, S. B. Gabriel, C. W. M. Roberts, J. a. Biegel, K. Stegmaier, A. J. Bass, L. a. Garraway, M. Meyerson, T. R. Golub, D. a. Gordenin, S. Sunyaev, E. S. Lander, and G. Getz. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–8, 2013. 66

[127] B. Lee and F. M. Richards. The interpretation of protein structures: estimation of static accessibility. *Journal of molecular biology*, 55(3):379–400, 1971. 42

[128] D. A. Lee, R. Rentzsch, and C. Oreng. GeMMA: Functional subfamily classification within superfamilies of predicted protein structural domains. *Nucleic Acids Research*, 38(3):720–737, 2009. 52

[129] J. C. Lee, I. Vivanco, R. Beroukhim, J. H. Y. Huang, W. L. Feng, R. M. DeBiasi, K. Yoshimoto, J. C. King, P. Nghiemphu, Y. Yuza, Q. Xu, H. Greulich, R. K. Thomas, J. G. Paez, T. C. Peck, D. J. Linhart, K. A. Glatt, G. Getz, R. Onofrio, L. Ziaugra, R. L. Levine, S. Gabriel, T. Kawaguchi, K. O'Neill, H. Khan, L. M. Liau, S. F. Nelson, P. N. Rao, P. Mischel, R. O. Pieper, T. Cloughesy, D. J. Leahy, W. R. Sellers, C. L. Sawyers, M. Meyerson, and I. K. Mellinghoff. Epidermal growth factor receptor

activation in glioblastoma through novel missense mutations in the extracellular domain. *PLoS Medicine*, 3(12):2264–2273, 2006. 117, 118

[130] S. Lee and T. L. Blundell. Ulla: A program for calculating environment-specific amino acid substitution tables. *Bioinformatics*, 25(15):1976–1977, 2009. 184

[131] T.-Y. Lee. dbPTM: an information repository of protein post-translational modification. *Nucleic Acids Research*, 34(90001):D622–D627, 2006. 61

[132] G. Lettre, V. G. Sankaran, M. A. A. Bezerra, A. S. Araújo, M. Uda, S. Sanna, A. Cao, D. Schlessinger, F. F. Costa, J. N. Hirschhorn, and S. H. Orkin. DNA polymorphisms at the BCL11A, HBS1L-MYB, and beta-globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease. *Proceedings of the National Academy of Sciences of the United States of America*, 105(33):11869–11874, aug 2008. 28

[133] A. Liberzon, C. Birger, H. Thorvaldsd??ttir, M. Ghandi, J. P. Mesirov, and P. Tamayo. The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems*, 1(6):417–425, 2015. 140

[134] M. Liu, L. T. Watson, and L. Zhang. HMMvar-func: a new method for predicting the functional outcome of genetic variants. *BMC Bioinformatics*, 16(1):351, oct 2015. 40

[135] H.-C. Lu, S. S. Chung, A. Fornili, and F. Fraternali. Anatomy of protein disorder, flexibility and disease-related mutations. *Frontiers in molecular biosciences*, 2:47, jan 2015. 30, 33, 34, 48, 74, 109

[136] H.-C. Lu, A. Fornili, and F. Fraternali. Protein–protein interaction networks studies and importance of 3D structure knowledge. *Expert Review of Proteomics*, 10(6):511–520, 2013. 48

[137] T. Madden. The BLAST sequence analysis tool. *The BLAST Sequence Analysis Tool*, pages 1–17, 2013. 186

[138] T. Madej, K. J. Addess, J. H. Fong, L. Y. Geer, R. C. Geer, C. J. Lanczycki, C. Liu, S. Lu, A. Marchler-Bauer, A. R. Panchenko, J. Chen, P. A. Thiessen, Y. Wang, D. Zhang, and S. H. Bryant. MMDB: 3D structures and macromolecular interactions. *Nucleic Acids Research*, 40(D1), 2012. 49

[139] P. Maietta, G. Lopez, A. Carro, B. J. Pingilley, L. G. Leon, A. Valencia, and M. L. Tress. FireDB: a compendium of biological and pharmacologically relevant ligands. *Nucleic acids research*, 42(Database issue):D267–72, 2014. 56

[140] U. ManChon, E. Talevich, S. Katiyar, K. Rasheed, and N. Kannan. Prediction and prioritization of rare oncogenic mutations in the cancer Kinome using novel features and multiple classifiers. *PLoS computational biology*, 10(4):e1003545, 2014. 39, 100

[141] L. Manzella, S. Stella, M. S. Pennisi, E. Tirrò, M. Massimino, C. Romano, A. Puma, M. Tavarelli, and P. Vigneri. New insights in thyroid cancer and p53 family proteins, 2017. 146

[142] A. Marchler-Bauer, M. K. Derbyshire, N. R. Gonzales, S. Lu, F. Chitsaz, L. Y. Geer, R. C. Geer, J. He, M. Gwadz, D. I. Hurwitz, C. J. Lanczycki, F. Lu, G. H. Marchler, J. S. Song, N. Thanki, Z. Wang, R. A. Yamashita, D. Zhang, C. Zheng, and S. H. Bryant. CDD: NCBI's conserved domain database. *Nucleic Acids Research*, 43(D1):D222–D226, 2015. 61

[143] K. a. Marino, L. Sutto, and F. L. Gervasio. The Effect of a Widespread Cancer-Causing Mutation on the Inactive to Active Dynamics of the B-Raf Kinase. *Journal of the American Chemical Society*, page 150417093156009, 2015. 16, 32

[144] A. C. Martin, A. M. Facchiano, A. L. Cuff, T. Hernandez-Boussard, M. Olivier, P. Hainaut, and J. M. Thornton. Integrating mutation data and structural analysis of the TP53 tumor-suppressor protein. *Human mutation*, 19(2):149–164, feb 2002. 28, 30, 31, 114

[145] C. A. Mather, S. D. Mooney, S. J. Salipante, S. Scroggins, D. Wu, C. C. Pritchard, and B. H. Shirts. CADD score has limited clinical validity for the identification of pathogenic variants in noncoding regions in a hereditary cancer panel. *Genetics in medicine : official journal of the American College of Medical Genetics*, (October 2015):1–7, 2016. 139

[146] C. C. Matthew H. Ung, Chun-Chi Liu. integrative analysis of cancer genes in a functional interactome. *Sci Rep*, 6(29118), 2016. 20, 136

[147] S. Maurer-Stroh, M. Debulpaep, N. Kuemmerer, M. Lopez de la Paz, I. C. Martins, J. Reumers, K. L. Morris, A. Copland, L. C. Serpell, L. Serrano, J. W. H. Schymkowitz, and F. Rousseau. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nature methods*, 7(3):237–42, 2010. 41

[148] N. McGranahan and C. Swanton. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer Cell*, 27(1):15–26, 2015. 19, 127, 128

[149] A. D. McLachlan. Rapid comparison of protein structures. *Acta Crystallographica Section A*, 38(6):871–873, 1982. 186

[150] R. D. Melamed, K. J. Emmett, C. Madubata, A. Rzhetsky, and R. Rabadan. Genetic similarity between cancers and comorbid Mendelian diseases identifies candidate driver genes. *Nature Communications*, 6:7033, 2015. 29

[151] D. Merico, R. Isserlin, O. Stueker, A. Emili, and G. D. Bader. Enrichment map: A network-based method for gene-set enrichment visualization and interpretation. *PLoS ONE*, 5(11), 2010. 139

[152] B. Mészáros, A. Zeke, A. Reményi, I. Simon, and Z. Dosztányi. Systematic analysis of somatic mutations driving cancer: uncovering functional protein regions in disease development. *Biology Direct*, 11(1):23, 2016. 66

[153] H. Mi, A. Muruganujan, J. T. Casagrande, and P. D. Thomas. Large-scale gene function analysis with the PANTHER classification system. *Nature protocols*, 8(8):1551–66, 2013. 72, 133, 144

[154] M. L. Miller, E. Reznik, N. P. Gauthier, B. A. Aksoy, A. Korkut, J. Gao, G. Ciriello, N. Schultz, and C. Sander. Pan-Cancer Analysis of Mutation Hotspots in Protein Domains. *Cell Systems*, 1(3):197–209, 2015. 17, 68, 69, 70, 83, 84, 126, 127, 141, 142, 167

[155] M. P. Miller and S. Kumar. Understanding human disease mutations through the use of interspecific genetic variation. *Human Molecular Genetics*, 10(21):2319–2328, oct 2001. 29, 38

[156] M. a. Molina-Vila, N. Nabau-Moretó, C. Tornador, A. J. Sabnis, R. Rosell, X. Estivill, T. G. Bivona, and C. Marino-Buslje. Activating mutations cluster in the "molecular brake" regions of protein kinases and do not associate with conserved or catalytic residues. *Human mutation*, 35(3):318–28, 2014. 112, 206

[157] S. Mooney. Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. *Briefings in bioinformatics*, 6(1):44–56, mar 2005. 42

[158] S. Mooney. Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. *Briefings in Bioinformatics*, 6(1):44–56, 2005. 44

[159] A. Mora and I. M. Donaldson. IRefR: An R package to manipulate the iRefIndex consolidated protein interaction database. *BMC Bioinformatics*, 12(1), 2011. 141

[160] K. L. Morley and R. J. Kazlauskas. Improving enzyme properties: When are closer mutations better? *Trends in Biotechnology*, 23(5):231–237, 2005. 31

[161] P. a. J. Muller and K. H. Vousden. P53 Mutations in Cancer. *Nature cell biology*, 15(1):2–8, 2013. 31, 35

[162] M. E. Murphy. The HSP70 family and cancer. *Carcinogenesis*, 34(6):1181–1188, 2013. 174

[163] P. Nasarre, V. Potiron, H. Drabkin, and J. Roche. Guidance molecules in lung cancer, 2010. 160

[164] P. C. Ng and S. Henikoff. SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13):3812–3814, jul 2003. 38, 39

[165] P. C. Ng and S. Henikoff. Predicting the effects of amino acid substitutions on protein function. *Annual review of genomics and human genetics*, 7(1):61–80, aug 2006. 28, 31, 38

[166] Nikos G. Gavalas, Michalis Liontos, Sofia-Paraskevi Trachana, Tina Bagratuni, Calliope Arapinis, Christine Liacos, Meletios A. Dimopoulos and A. Bamias. Angiogenesis-Related Pathways in the Pathogenesis of Ovarian Cancer. *Int J Mol Sci*, 14(8):15885–15909, 2013. 172

[167] A. Niroula, S. Urolagin, and M. Vihinen. PON-P2: Prediction method for fast and reliable identification of harmful variants. *PLoS ONE*, 10(2), 2015. 137

[168] A. Niroula and M. Vihinen. Harmful somatic amino acid substitutions affect key pathways in cancers. *BMC medical genomics*, 8(1), 2015. 137

[169] H. Nishi, M. Tyagi, S. Teng, B. a. Shoemaker, K. Hashimoto, E. Alexov, S. Wuchty, and A. R. Panchenko. Cancer missense mutations alter binding properties of proteins and their interaction networks. *PloS one*, 8(6):e66273, 2013. 32, 61, 62, 123, 133, 203

[170] R. Nussinov and C.-J. Tsai. 'Latent drivers' expand the cancer mutational landscape. *Current opinion in structural biology*, 32C:25–32, 2015. 29

[171] Z. Obradovic, K. Peng, S. Vucetic, P. Radivojac, and a. K. Dunker. Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins: Structure, Function and Genetics*, 61(SUPPL. 7):176–182, 2005. 33

[172] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. CATH–a hierarchic classification of protein domain structures. *Structure (London, England : 1993)*, 5(8):1093–1108, aug 1997. 51

[173] C. a. Orengo and W. R. Taylor. SSAP: sequential structure alignment program for protein structure comparison. *Methods in enzymology*, 266:617–635, 1996. 84, 186

[174] K. Paňková, D. Rösel, M. Novotný, and J. Brábek. The molecular mechanisms of transition between mesenchymal and amoeboid invasiveness in tumor cells, 2010. 162

[175] Y. Pan, K. Karagiannis, H. Zhang, H. Dingerdissen, A. Shamsaddini, Q. Wan, V. Simonyan, and R. Mazumder. Human germline and pan-cancer variomes and their distinct functional profiles. *Nucleic Acids Research*, 42(18):11570–11588, 2014. 19, 60, 98, 133, 134, 144

[176] R. G. Parra, N. P. Schafer, L. G. Radusky, M. Y. Tsai, A. B. Guzovsky, P. G. Wolynes, and D. U. Ferreiro. Protein Frustratometer 2: a tool to localize energetic frustration in protein molecules, now with electrostatics. *Nucleic acids research*, 44(W1):W356–W360, 2016. 65

[177] M. Patani, H, Bunney, TD, Thiyagarajan, N, Norman, RA, Ogg, D, Breed, J, Ashford, P, Potterton, A, Edwards, M, Williams, SV, Thomson, GS, Pang, CSM, Knowles, MA, Breeze, AL, Orengo, C, Phillips, C and Katan. Landscape of activating cancer mutations in FGFR kinases and their differential responses to inhibitors in clinical use. *Oncotarget*, pages 1949–2553, 2016. 31, 185, 208

[178] M. B. Paulina Kucharzewska, Helena C. Christianson. Global Profiling of Metabolic Adaptation to Hypoxic Stress in Human Glioblastoma Cells. *PLoS One.*, 10(1), 2015. 129

[179] J. Pei and N. V. Grishin. AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics (Oxford, England)*, 17(8):700–712, 2001. 56

[180] B. Petersen, T. Petersen, P. Andersen, M. Nielsen, and C. Lundegaard. A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Structural Biology*, 9(1):51, 2009. 42

[181] T. A. Peterson, I. I. M. Gauran, J. Park, D. H. Park, and M. G. Kann. Oncodomains: A protein domain-centric framework for analyzing rare variants in tumor samples. *PLoS Computational Biology*, 13(4), 2017. 70

[182] T. a. Peterson, N. L. Nehrt, D. Park, and M. G. Kann. Incorporating molecular and functional context into the analysis and prioritization of human variants associated with cancer. *Journal of the American Medical Informatics Association*, 19(2):275–283, 2012. 16, 66, 67, 70, 127

[183] A. Petitjean, S. Mathe, E.; Kato, C. Ishioka, S. V. Tavtigian, P. Hainaut, and M. Olivier. Impact of Mutant p53 Functional Properties on TP53 Mutation Patterns and Tumor Phenotype: Lessons from Recent Developments in the IARC TP53 Database. *Human Mutation*, 27(July):796–802, 2006. 40

[184] U. Pieper, B. M. Webb, D. T. Barkan, D. Schneidman-Duhovny, A. Schlessinger, H. Braberg, Z. Yang, E. C. Meng, E. F. Pettersen, C. C. Huang, R. S. Datta, P. Sampathkumar, M. S. Madhusudhan, K. Sjölander, T. E. Ferrin, S. K. Burley, and A. Sali. ModBase,a database of annotated comparative protein structure models,and associated resources. *Nucleic Acids Research*, 39(SUPPL. 1), 2011. 186

[185] D. E. Pires, D. B. Ascher, and T. L. Blundell. MCSM: Predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, 30(3):335–342, 2014. 36, 184

[186] D. E. Pires, T. L. Blundell, and D. B. Ascher. Platinum: A database of experimentally measured effects of mutations on structurally defined protein-ligand complexes. *Nucleic Acids Research*, 43(D1):D387–D391, 2015. 36

[187] D. E. V. Pires, D. B. Ascher, and T. L. Blundell. DUET: A server for predicting ef-

fects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Research*, 42(W1):314–319, 2014. 36, 184, 187, 209

[188] E. Porta-Pardo, L. Garcia-Alonso, T. Hrabe, J. Dopazo, and A. Godzik. A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces. *PLoS Computational Biology*, 11(10), 2015. 63, 72, 117

[189] E. Porta-Pardo and A. Godzik. e-Driver: a novel method to identify protein regions driving cancer. *Bioinformatics (Oxford, England)*, 30(21):3109–3114, nov 2014. 37, 63, 70, 73, 117, 125

[190] E. Porta-Pardo, T. Hrabe, and A. Godzik. Cancer3D: understanding cancer mutations through protein structures. *Nucleic Acids Research*, 43(D1):D968—-D973, jan 2015. 37

[191] E. Porta-Pardo, A. Kamburov, D. Tamborero, T. Pons, D. Grases, A. Valencia, N. Lopez-Bigas, G. Getz, and A. Godzik. Comparison of algorithms for the detection of cancer drivers at subgene resolution. *Nature Methods*, 14(8):782–788, 2017. 38

[192] C. T. Porter, G. J. Bartlett, and J. M. Thornton. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic acids research*, 32(Database issue):D129—-D133, jan 2004. 41, 49, 56, 61, 85

[193] D. D. Rafael A Jordan, Feihong Wu and V. Honavar. ProtinDb: A data base of protein-protein interface residues. 70

[194] V. Rajendran, R. Purohit, and R. Sethumadhavan. In silico investigation of molecular mechanism of laminopathy caused by a point mutation (R482W) in lamin A/C protein. *Amino acids*, 43(2):603–615, aug 2012. 45

[195] J. Reumers, S. Maurer-Stroh, J. Schymkowitz, and F. Rousseau. SNPeffect v2.0: A new step in investigating the molecular phenotypic effects of human non-synonymous SNPs. *Bioinformatics*, 22(17):2183–2185, 2006. 41

[196] J. Reumers, J. Schymkowitz, J. Ferkinghoff-Borg, F. Stricher, L. Serrano, and F. Rousseau. SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs. *Nucleic acids research*, 33(Database issue):D527+, jan 2005. 32, 41

[197] B. Reva, Y. Antipin, and C. Sander. Determinants of protein function revealed by combinatorial entropy optimization. *Genome biology*, 8(11):R232, 2007. 40

[198] B. Reva, Y. Antipin, and C. Sander. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research*, 39(17):gkr407—-e118, sep 2011. 29, 38, 40

[199] C. J. Richardson, Q. Gao, C. Mitsopoulous, M. Zvelebil, L. H. Pearl, and F. M. Pearl. MoKCa database–mutations of kinases in cancer. *Nucleic acids research*, 37(Database issue):gkn832+, jan 2009. 38, 72

[200] G. a. Ryslik, Y. Cheng, K.-H. Cheung, R. D. Bjornson, D. Zelterman, Y. Modis, and H. Zhao. A spatial simulation approach to account for protein structure when identifying non-random somatic mutations. *BMC bioinformatics*, 15(1):231, 2014. 76

[201] G. a. Ryslik, Y. Cheng, K.-H. Cheung, Y. Modis, and H. Zhao. A graph theoretic approach to utilizing protein structure to identify non-random somatic mutations. *BMC bioinformatics*, 15(1):86, 2014. 17, 76, 77

[202] N. Sahni, S. Yi, M. Taipale, J. I. Fuxman Bass, J. Coulombe-Huntington, F. Yang, J. Peng, J. Weile, G. I. Karras, Y. Wang, I. A. Kovács, A. Kamburov, I. Krykbaeva, M. H. Lam, G. Tucker, V. Khurana, A. Sharma, Y.-Y. Y. Liu, N. Yachie, Q. Zhong, Y. Shen, A. Palagi, A. San-Miguel, C. Fan, D. Balcha, A. Dricot, D. M. Jordan, J. M. Walsh, A. A. Shah, X. Yang, A. K. Stoyanova, A. Leighton, M. A. Calderwood, Y. Jacob, M. E. Cusick, K. Salehi-Ashtiani, L. J. Whitesell, S. Sunyaev, B. Berger, A.-L. L. Barabási, B. Charloteaux, D. E. Hill, T. Hao, F. P. Roth, Y. Xia, A. J. Walhout,

S. Lindquist, and M. Vidal. Widespread macromolecular interaction perturbations in human genetic disorders. *Cell*, 161(3):647–660, apr 2015. 115

[203] V. Sanz-Moreno and C. J. Marshall. The plasticity of cytoskeletal dynamics underlying neoplastic cell migration, 2010. 162, 165

[204] C. T. Saunders and D. Baker. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *Journal of molecular biology*, 322(4):891–901, sep 2002. 41

[205] E. Scholzová, R. Malík, J. Sevcík, and Z. Kleibl. RNA regulation and cancer development. *Cancer letters*, 246(1-2):12–23, 2007. 166

[206] B. Schuster-BÃ\Pckler and A. Bateman. Protein interactions in human genetic diseases. *Genome Biology*, 9(1):R9+, jan 2008. 95

[207] A. W. R. Serohijos and E. I. Shakhnovich. Merging molecular mechanism and evolution: Theory and computation at the interface of biophysics and evolutionary population genetics. *Current Opinion in Structural Biology*, 26(1):84–91, 2014. 202

[208] L. Serrano, J. Schymkowitz, J. Borg, F. Stricher, R. Nys, and F. Rousseau. The FoldX web server: an online force field. *Nucl. Acids Res.*, 33(suppl_2):W382–388, 2005. 41, 57, 62, 183, 187

[209] N. Shah, Y. Claire Hou, H. Yu, R. Sainger, E. Dec, B. Perkins, C. Thomas Caskey, J. Craig Venter, and A. Telenti. Identification of misclassified ClinVar variants using disease population prevalence. *bioRxiv*, pages 1–23, 2016. 65

[210] H. M. S. Shahul and S. P. Sarma. The structure of the thioredoxin-triosephosphate isomerase complex provides insights into the reversible glutathione-mediated regulation of triosephosphate isomerase. *Biochemistry*, 51(1):533–544, 2012. 35

[211] Q. Shen, F. Cheng, H. Song, W. Lu, J. Zhao, X. An, M. Liu, G. Chen, Z. Zhao, and J. Zhang. Proteome-Scale Investigation of Protein Allosteric Regulation Perturbed

by Somatic Mutations in 7,000 Cancer Genomes. *American Journal of Human Genetics*, 100(1):5–20, 2017. 109

[212] R. R. Shen and W. C. Hahn. Emerging roles for the non-canonical IKKs in cancer, 2011. 174

[213] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, 29(1):308–311, 2001. 48, 56, 57, 61

[214] Z. Shi and J. Moult. Structural and functional impact of cancer-related missense somatic mutations. *Journal of molecular biology*, 413(2):495–512, oct 2011. 35, 63

[215] H. A. Shihab, J. Gough, D. N. Cooper, P. D. Stenson, G. L. Barker, K. J. Edwards, I. N. Day, and T. R. Gaunt. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Human mutation*, 34(1):57–65, jan 2013. 39

[216] B. A. Shoemaker, D. Zhang, M. Tyagi, R. R. Thangudu, J. H. Fong, A. Marchler-Bauer, S. H. Bryant, T. Madej, and A. R. Panchenko. IBIS (Inferred Biomolecular Interaction Server) reports, predicts and integrates multiple types of conserved interactions for proteins. *Nucleic acids research*, 40(Database issue):D834—-D840, jan 2012. 49, 85, 86, 90

[217] I. Sillitoe, T. E. Lewis, A. Cuff, S. Das, P. Ashford, N. L. Dawson, N. Furnham, R. A. Laskowski, D. Lee, J. G. Lees, S. Lehtinen, R. A. Studer, J. Thornton, and C. A. Orengo. CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Research*, 43(D1):D376—-D381, jan 2015. 52

[218] H. Singh, J. S. Chauhan, M. M. Gromiha, and G. P. S. Raghava. ccPDB: Compilation and creation of data sets from Protein Data Bank. *Nucleic Acids Research*, 40(D1), 2012. 49, 61, 86

[219] E. Sitbon and S. Pietrokovski. Occurrence of protein structure elements in conserved sequence regions. *BMC structural biology*, 7(1):3+, jan 2007. 29

[220] V. Sobolev, A. Sorokine, J. Prilusky, E. E. Abola, and M. Edelman. Automated analysis of interatomic contacts in proteins. *Bioinformatics*, 15(4):327–332, 1999. 59

[221] W. Song, S. A. Gardner, H. Hovhannisyan, A. Natalizio, K. S. Weymouth, W. Chen, I. Thibodeau, E. Bogdanova, S. Letovsky, A. Willis, and N. Nagan. Exploring the landscape of pathogenic genetic variation in the ExAC population database: Insights of relevance to variant classification. *Genetics in Medicine*, 18(8):850–854, 2016. 65

[222] E. L. L. Sonnhammer, S. R. Eddy, and R. Durbin. Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins: Structure, Function and Genetics*, 28(3):405–420, 1997. 39, 48, 51, 68

[223] M. Soundararajan, A. K. Roos, P. Savitsky, P. Filippakopoulos, A. N. Kettenbach, J. V. Olsen, S. A. Gerber, J. Eswaran, S. Knapp, and J. M. Elkins. Structures of down syndrome kinases, DYRKs, reveal mechanisms of kinase activation and substrate recognition. *Structure*, 21(6):986–996, 2013. 121

[224] C. Stark and B. Breitkreutz. The BioGRID interaction database: 2011 update. *Nucleic acids . . .* , 39(Database issue):D698—-704, 2011. 141

[225] S. Stefl, H. Nishi, M. Petukh, A. R. Panchenko, and E. Alexov. Molecular mechanisms of disease-causing missense mutations. *Journal of Molecular Biology*, 425(21):3919–3936, 2013. 27, 30, 31

[226] H. Stehr, S.-H. J. Jang, J. M. Duarte, C. Wierling, H. Lehrach, M. Lappe, and B. M. H. Lange. The structural impact of cancer-associated missense mutations in oncogenes and tumor suppressors. *Molecular cancer*, 10:54, 2011. 16, 35, 38, 43, 57, 58, 59, 63, 65, 70, 74, 76, 79, 82, 103, 182

[227] P. D. Stenson, E. V. Ball, M. Mort, A. D. Phillips, J. a. Shiel, N. S. T. Thomas, S. Abeysinghe, M. Krawczak, and D. N. Cooper. Human Gene Mutation Database (HGMD®): 2003 Update. *Human Mutation*, 21(6):577–581, 2003. 39, 61, 63, 64, 65

[228] R. E. Steward, M. W. MacArthur, R. A. Laskowski, and J. M. Thornton. Molecular basis of inherited diseases: a structural perspective. *Trends in genetics : TIG*, 19(9):505–513, sep 2003. 28

[229] N. O. Stitziel, Y. Y. Y. Tseng, D. Pervouchine, D. Goddeau, S. Kasif, and J. Liang. Structural location of disease-associated single-nucleotide polymorphisms. *Journal of molecular biology*, 327(5):1021–1030, apr 2003. 29

[230] R. A. Studer, B. H. Dessailly, and C. A. Orengo. Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. *The Biochemical journal*, 449(3):581–594, feb 2013. 31, 34, 35

[231] S. Sunyaev, V. Ramensky, and P. Bork. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends in genetics : TIG*, 16(5):198–200, may 2000. 29, 41

[232] S. Sunyaev, V. Ramensky, I. Koch, W. Lathe, A. S. Kondrashov, and P. Bork. Prediction of deleterious human alleles. *Human Molecular Genetics*, 10(6):591–597, mar 2001. 28

[233] S. R. Sunyaev, F. Eisenhaber, I. V. Rodchenkov, B. Eisenhaber, V. G. Tumanyan, and E. N. Kuznetsov. PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein engineering*, 12(5):387–394, 1999. 42

[234] D. C. Sushant Kumar and M. Gerstein. Localized structural frustration for evaluating the impact of sequence variants. *Nucleic Acids Res*, 44(21):10062–10073, 2016. 64

[235] P. Taillon-Miller, Z. Gu, Q. Li, L. Hillier, and P. Y. Kwok. Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms. *Genome research*, 8(7):748–754, jul 1998. 28

[236] H. Takane, E. Shikata, K. Otsubo, S. Higuchi, and I. Ieiri. Polymorphism in human organic cation transporters and metformin action. *Pharmacogenomics*, 9(4):415–422, apr 2008. 28

[237] D. Talavera, R. a. Laskowski, and J. M. Thornton. WSsas: a web service for the annotation of functional residues through structural homologues. *Bioinformatics (Oxford, England)*, 25(9):1192–4, 2009. 56

[238] D. Talavera, M. S. Taylor, and J. M. Thornton. The (non)malignancy of cancerous amino acidic substitutions. *Proteins*, 78(3):518–529, feb 2010. 28

[239] D. Talavera, M. S. Taylor, and J. M. Thornton. The (non)malignancy of cancerous amino acidic substitutions. *Proteins*, 78(3):518–29, 2010. 56, 106

[240] D. Tamborero, A. Gonzalez-Perez, and N. Lopez-Bigas. OncodriveCLUST: Exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics*, 29(18):2238–2244, 2013. 44, 66, 73

[241] M. Tanabe and M. Kanehisa. Using the KEGG database resource. *Current Protocols in Bioinformatics*, (SUPPL.38), 2012. 53

[242] S. S. Taylor and A. P. Kornev. Protein kinases: Evolution of dynamic regulatory proteins. *Trends in Biochemical Sciences*, 36(2):65–77, 2011. 21, 112, 113, 180, 181, 206, 208

[243] S. Teng, E. Michonova-Alexova, and E. Alexov. Approaches and resources for prediction of the effects of non-synonymous single nucleotide polymorphism on protein function and interactions. *Current pharmaceutical biotechnology*, 9(2):123–133, apr 2008. 28

[244] J. F. Tien, A. Mazloomian, S. W. Cheng, C. S. Hughes, C. C. Chow, L. T. Canapi, A. Oloumi, G. Trigo-Gonzalez, A. Bashashati, J. Xu, V. C. Chang, S. P. Shah, S. Aparicio, and G. B. Morin. CDK12 regulates alternative last exon mRNA splicing and promotes breast cancer cell invasion. *Nucleic Acids Research*, 45(11):6698–6716, 2017. 166

[245] Y. K. Tingting Jiang. Methods for detecting co-mutated pathways in cancer samples to inform treatment selection. *bioRxiv*, 2016. 20, 138, 139, 168

[246] A. E. Todd, C. A. Orengo, and J. M. Thornton. Plasticity of enzyme active sites. *Trends in biochemical sciences*, 27(8):419–426, aug 2002. 29

[247] C. Tokheim, R. Bhattacharya, N. Niknafs, D. M. Gygax, R. Kim, M. Ryan, D. L. Masica, and R. Karchin. Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure. *Cancer Research*, 76(13):3719–3731, 2016. 38, 63, 74, 75, 78, 125, 148

[248] N. Tokuriki, F. Stricher, L. Serrano, and D. S. Tawfik. How protein stability and new functions trade off. *PLoS Computational Biology*, 4(2):35–37, 2008. 30, 31

[249] K. Tomczak, P. Czerwińska, and M. Wiznerowicz. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge, 2015. 63, 68, 70, 83, 112, 185

[250] A. Torkamani, N. Kannan, S. S. Taylor, and N. J. Schork. Congenital disease SNPs target lineage specific structural elements in protein kinases. *Proceedings of the National Academy of Sciences of the United States of America*, 105(26):9011–9016, jul 2008. 34, 206

[251] E. Tüchsen, M. Ø. Jensen, and P. Westh. Solvent accessible surface area (ASA) of simulated phospholipid membranes. *Chemistry and Physics of Lipids*, 123(1):107–116, 2003. 62

[252] V. Vacic and L. M. Iakoucheva. Disease mutations in disordered regions–exception to the rule? *Molecular bioSystems*, 8(1):27–32, jan 2012. 32, 33, 34

[253] W. S. Valdar. Scoring residue conservation. *Proteins*, 48(2):227–241, aug 2002. 29, 52, 87

[254] S. Velankar, J. M. Dana, J. Jacobsen, G. van Ginkel, P. J. Gane, J. Luo, T. J. Oldfield, C. O'Donovan, M.-J. Martin, and G. J. Kleywegt. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic acids research*, 41(Database issue):D483–9, 2013. 82

[255] J. a. Veltman and H. G. Brunner. De novo mutations in human genetic disease. *Nature Reviews Genetics*, 13(8):565–575, 2012. 27

[256] D. Vitkup, C. Sander, and G. Church. The amino-acid mutational spectrum of human genetic disease. *Genome Biology*, 4(11):R72+, 2003. 29, 30

[257] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. a. Diaz, and K. W. Kinzler. Cancer genome landscapes. *Science (New York, N.Y.)*, 339(6127):1546–58, 2013. 43, 79, 81, 126, 144, 168, 171

[258] H. Vuong, F. Cheng, C.-C. Lin, and Z. Zhao. Functional consequences of somatic mutations in cancer using protein pocket-based prioritization approach. *Genome medicine*, 6(10):81, 2014. 30

[259] M. S. Walid. Prognostic factors for long-term survival after glioblastoma. *The Permanente journal*, 12(4):45–8, 2008. 169, 174

[260] P. T. Wan, M. J. Garnett, S. M. Roe, S. Lee, D. Niculescu-Duvaz, V. M. Good, C. M. Jones, C. J. Marshall, C. J. Springer, D. Barford, R. Marais, and Cancer Genome Project. Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF. *Cell*, 116(6):855–867, mar 2004. 31, 32

[261] G. Wang and A. R. Fersht. Multisite aggregation of p53 and implications for drug rescue. *Proceedings of the National Academy of Sciences*, 114(13):E2634–E2643, 2017. 31

[262] X. Wang, X. Wei, B. Thijssen, J. Das, S. M. Lipkin, and H. Yu. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nature biotechnology*, 30(2):159–164, feb 2012. 41, 63

[263] Z. Wang and J. Moult. SNPs, protein structure, and disease. *Human mutation*, 17(4):263–270, apr 2001. 28, 30

[264] J. J. Ward, L. J. McGuffin, K. Bryson, B. F. Buxton, and D. T. Jones. The DISOPRED server for the prediction of protein disorder. *Bioinformatics*, 20(13):2138–2139, 2004. 33, 48, 74, 184

[265] B. Webb and A. Sali. Comparative Protein Structure Modeling Using MODELLER. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, 47:5.6.1–32, 2014. 185

[266] M. Wiederstein and M. J. Sippl. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic acids research*, 35(Web Server issue):W407–10, 2007. 186

[267] C. L. Worth, S. Gong, and T. L. Blundell. Structural and functional constraints in the evolution of protein families. *Nature reviews. Molecular cell biology*, 10(10):709–720, 2009. 35

[268] C. L. Worth, R. Preissner, and T. L. Blundell. SDM–a server for predicting effects of mutations on protein stability and malfunction. *Nucleic Acids Research*, 39(suppl):W215–W222, 2011. 183

[269] G. Wu, X. Feng, and L. Stein. A human functional protein interaction network and its application to cancer data analysis. *Genome Biol*, 11(5):R53, 2010. 19, 53, 134, 135, 137, 141, 145, 154, 170

[270] Y. Xu, D. Xu, H. N. Gabow, and H. Gabow. Protein domain decomposition using a graph-theoretic approach. *Bioinformatics (Oxford, England)*, 16(12):1091–104, 2000. 57, 76

[271] F. Yang, E. Petsalaki, T. Rolland, D. E. Hill, M. Vidal, and F. P. Roth. Protein Domain-Level Landscape of Cancer-Type-Specific Somatic Mutations. *PLoS computational biology*, 11(3):e1004147, 2015. 35, 63, 69, 74, 75, 79, 103, 123, 126, 141, 142, 167

[272] C. M. Yates, I. Filippis, L. A. Kelley, and M. J. Sternberg. SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. *Journal of molecular biology*, 426(14):2692–2701, jul 2014. 42

[273] C. M. Yates and M. J. Sternberg. Proteins and domains vary in their tolerance of non-synonymous single nucleotide polymorphisms (nsSNPs). *Journal of molecular biology*, 425(8):1274–1286, apr 2013. 28, 66

[274] C. M. Yates and M. J. Sternberg. The effects of non-synonymous single nucleotide polymorphisms (nsSNPs) on protein-protein interactions. *Journal of molecular biology*, 425(21):3949–3963, nov 2013. 30, 62

[275] S. Yellaboina, A. Tasneem, D. V. Zaykin, B. Raghavachari, and R. Jothi. DOMINE: a comprehensive collection of known and predicted domain-domain interactions. *Nucleic acids research*, 39(Database issue):D730–5, 2011. 42

[276] P. Yue, W. F. Forrest, J. S. Kaminker, S. Lohr, Z. Zhang, and G. Cavet. Inferring the functional effects of mutation through clusters of mutations in homologous proteins. *Human Mutation*, 31(3):264–271, 2010. 66

[277] P. Yue, Z. Li, and J. Moult. Loss of protein structure stability as a major causative factor in monogenic disease. *Journal of molecular biology*, 353(2):459–473, oct 2005. 30, 95

[278] P. Yue, E. Melamud, and J. Moult. SNPs3D: candidate gene and SNP selection for association studies. *BMC bioinformatics*, 7(1):166+, 2006. 43

[279] P. Yue and J. Moult. Identification and analysis of deleterious human SNPs. *Journal of molecular biology*, 356(5):1263–1274, mar 2006. 28

[280] J. Zhang, M. F. G. Stevens, and T. D. Bradshaw. Temozolomide: Mechanisms of Action, Repair and Resistance. *Current Molecular Pharmacology*, 5:102–114, 2012. 128

[281] Z. Zhang, J. Norris, C. Schwartz, and E. Alexov. In silico and in vitro investigations of the mutability of disease-causing missense mutation sites in spermine synthase. *PloS one*, 6(5):e20373, 2011. 70, 83, 112

[282] M. Zhao, J. Sun, and Z. Zhao. TSGene: a web resource for tumor suppressor genes. *Nucleic acids research*, 41(Database issue):D970–6, 2013. 81

[283] Q. Zhong, N. Simonis, Q.-R. R. Li, B. Charloteaux, F. Heuze, N. Klitgord, S. Tam, H. Yu, K. Venkatesan, D. Mou, V. Swearingen, M. A. Yildirim, H. Yan, A. Dricot, D. Szeto, C. Lin, T. Hao, C. Fan, S. Milstein, D. Dupuy, R. Brasseur, D. E. Hill, M. E. Cusick, and M. Vidal. Edgetic perturbation models of human inherited disorders. *Molecular systems biology*, 5(1), nov 2009. 33