Using genome-scale bioinformatics platforms to investigate the role of single nucleotide polymorphisms in the BRCA1 gene in key molecular pathways of disease

Rebecca Wall

Submitted to Swansea University in fulfilment of the requirements for the Degree of Msc Medical and Healthcare Studies by Research in 2021

DECLARATION

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed ...Rebecca Wall .... (candidate)
Date ....27/09/2021...

STATEMENT 1

This thesis is the result of my own investigations, except where otherwise stated. Where correction services have been used, the extent and nature of the correction is clearly marked in a footnote(s).
Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed ...Rebecca Wall .... (candidate)
Date ....27/09/2021...

STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed ...Rebecca .... (candidate)
Date ....27/09/2021...

Abstract

Single nucleotide polymorphisms (SNPs) are often associated with conferring risk for disease, and are associated with many complex diseases such as breast and ovarian cancer. The BRCA1 gene is known to carry mutations that can predispose an individual to such diseases. Currently, the clinical significance of most SNPs remains unknown due to the lack of successful and reliable classification tools, leading to the possibility that many pathogenic SNPs are not considered during genetic screening. In order to investigate the role of SNPs within crucial pathways andthe structural effects of SNPs, a database and data collection pipeline was constructed that sourced information from Reactome, ClinVar, and UniProt. A second pipeline was created that allowed for the modelling of variant proteins. Through querying the database, direct pathway associations with BRCA1 were identified. Protein variant modelling revealed a novelapproach to structural analysis of SNPs, allowing for differences in heuristic structural functions to be measured between pathogenic and benign variants. Of particular interest, theheuristic functions that showed the most significant differences were the van der Waals contacts and strict hydrogen bonds. Identification of SNPs within genes linked to complex diseases, such as BRCA1, can inform better targets of genetic screening and potentially provide new drug targets.

1. Introduction

1.1The Genetic Code

All genetic information is stored in long, double-helix strands called deoxyribonucleic acid (DNA).The genetic instructions contained within the DNA inform all processes of life, such as growthand repair, development, and metabolic processes. DNA consists of two polynucleotide chains that coil around each other to form a helix. DNA is formed of four main monomeric building blocks or ‘bases’ called nucleotides. These four nucleotides can be separated into two subgroups: purines and pyrimidines. These four bases are adenine and guanine (the purines) and thymine and cytosine (the pyrimidines), each often abbreviated to the first letter of the name (A, T, C, G). Each nucleotide contains nitrogen, allowing the two opposite bases to form hydrogen bonds holding the two strands together according to the base pairing rule. The base pairing rule dictates which two bases can be found opposite each other within the DNA where two of the same subgroup cannot be found opposite; pyrimidines are always found opposite purines. The rule also states there must be a 1:1 protein stereochemistry; thismeans the amount of guanine must equal the amount of cytosine, and the amount of thyminemust equal the amount of adenine.

DNA is catatonically known to exist in right-handed DNA helices where the DNA can adopt one of at least two known naturally occurring structures: A-DNA and B-DNA. The B-form is predominantly observed in cells and is most stable in high humidity, but there is a shift to the A-form upon a decrease of water activity. The shift from B-type to A-type affects the dynamics of transcription by removing direct access to the DNA, and also alters the physicochemical properties of the polymer.

DNA is transcribed in groups of three nucleotides, known as triplets or a codon. DNA is transcribed onto ribonucleic acid (RNA), where the opposite base to the parent DNA strand is represented on the RNA strand. Unlike DNA, the four bases of RNA are adenine, cytosine, guanine and, uracil (A, C, G, U). Uracil is seen in the place of thymine. This means that were an A is seen in the parent DNA, an U is added to the RNA strand. Each codon represents a specific amino acid or signal; these signal the start of the protein and the end of the protein, and are named start and stop codons. Redundancy is seen in the genetic code within these codons, where each amino acid can be represented by a number of different codons. Methionine is the only amino acid is coded by a singular codon, AUG, which also represents the start codon. The stop of transcription is coded by three codons, UAA, UAG, and UGA, and these do not encode any amino acids unlike the start codon.

DNA methylation is the epigenetic modification of the C5 position of cytosine through the covalent transfer of a methyl group. DNA methylation is generally important in the correct development; playing an important role in genomic imprinting, X-chromosome inactivation, and suppression of transcription and gene regulation (Jin, Li, & Robertson, 2011). The role of DNA methylation in carcinogenesis has been of increasing interest lately, and there have been multiple links between alterations in DNA methylation and cancer (Das, & Singal, 2016). It is thought that in particular, hypermethylation, a process that represses transcription of tumour suppressor genes leading to gene silencing, has been recognised as a cause of carcinogenesis.

1.2 Protein Synthesis and Folding

Protein synthesis occurs through a number of steps: transcription, translation, and posttranslational modification and folding.

1.2.1 DNA Transcription

There are three main steps to the transcription process: initiation, elongation, and termination.

During the initiation step, RNA polymerase binds to the promoter region of DNA, found near the beginning of genes. Each gene has its own promoter. RNA polymerase is the main enzyme in the transcription step; it uses a single-stranded DNA template to generate a complementary RNA molecule (Bailey, 2020). DNA is read from the 3' to the 5' end, while RNA is built in the 5' to 3' direction (each new nucleotide is added to the 3' end).

The second step, elongation, involves the formation and synthesis of the RNA strand by RNA polymerase. RNA polymerase builds the RNA strand with complimentary nucleotide bases to that of the DNA antisense (template) strand, meaning that the RNA strand carries the same information as the sense (non-template) strand of DNA (Khan Academy, 2021). The process can be seen visually in figure 1 below.

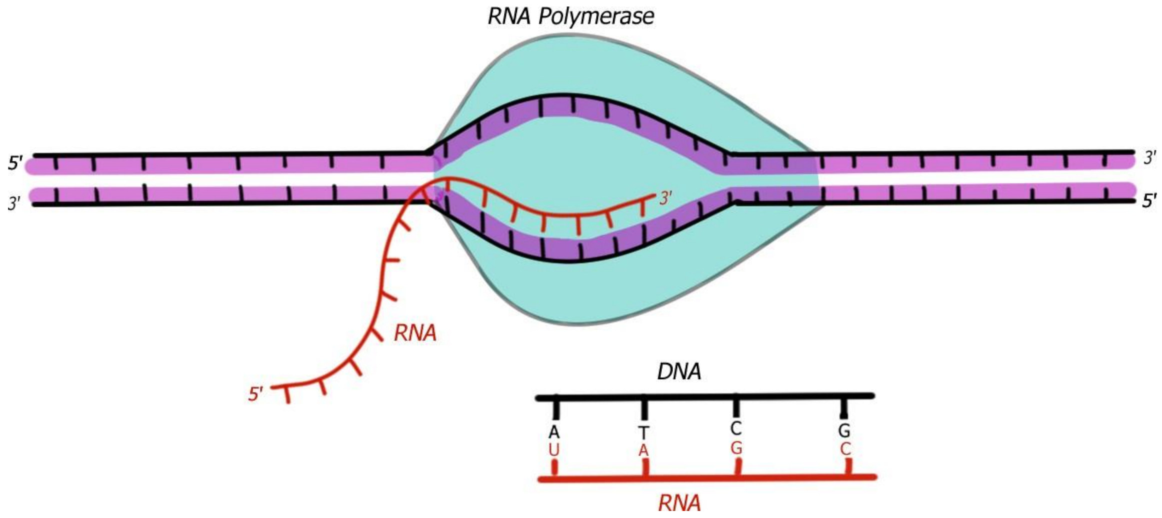

Figure 1 – The elongation step of DNA transcription. Transcription takes place in the 3' to 5' direction, while RNA is built in the 5' to 3' direction. RNA polymerase builds the new RNA strand with complimentary nucleotide bases to the template DNA strand. The complimentary base pairs are shown, including the base uracil in the place of thymine in the RNA strand.

The final step is termination of transcription, which can be achieved through a variety of mechanisms in eukaryotes depending on the RNA polymerase in use. Where there are protein-encoding genes, the signal for the termination of transcription occurs between a GC rich sequence and a U-rich tract (Roberts, 2019). These are separated by approximately 40to 60 bp on the RNA strand and so the termination signal is contained within the RNA itself. This termination method is known as rho-independent (Jun, S., Warner, & Murakami, 2013).The termination of transcription of this method is intrinsic to the RNA strand (Zenkin, 2014).

1.2.2 DNA Translation

During DNA translation, the protein itself is produced is the cell's cytoplasm. Before this can occur, the RNA strand must leave the cell nucleus. The RNA strand is modified to become a messenger RNA (mRNA) strand, and it often referred to being a pre-mRNA strand before this modification takes place (Gao, & Wang, 2020). Additionally, many eukaryotic pre-mRNA sequences need to undergo splicing to remove non-coding regions of DNA (introns) (Bhagavan, & Ha, 2015). If these introns are not removed, they will be translated along with the exons; producing a faulty polypeptide. Modified mature mRNAs are identified and are exported from the nucleus through the nuclear pore. As with transcription, translation occurs in three steps: initiation, elongation, and termination.

Initiation of translation begins once the start AUG codon is recognised, which is specific to the amino acid methionine and is nearly always the start of the polypeptide chain (for proteins that do not begin with methionine, this residue is removed post-translation) (Warren, 2020).

During translation elongation, transfer RNAs (tRNAs) carry amino acid residues to the mRNA molecule for the ribosome mechanism to add to the polypeptide chain. Complementation of the mRNA codons and the tRNA anticodon results in protein synthesis dependent on the mRNA nucleotide code (Pollard, Earnshaw, Lippincott-Schwartz, & Johnson, 2017). The polypeptide chain is extended via translocation of the ribosome along the mRNA. A visual version of the elongation step of translation can be seen below in Figure 2.

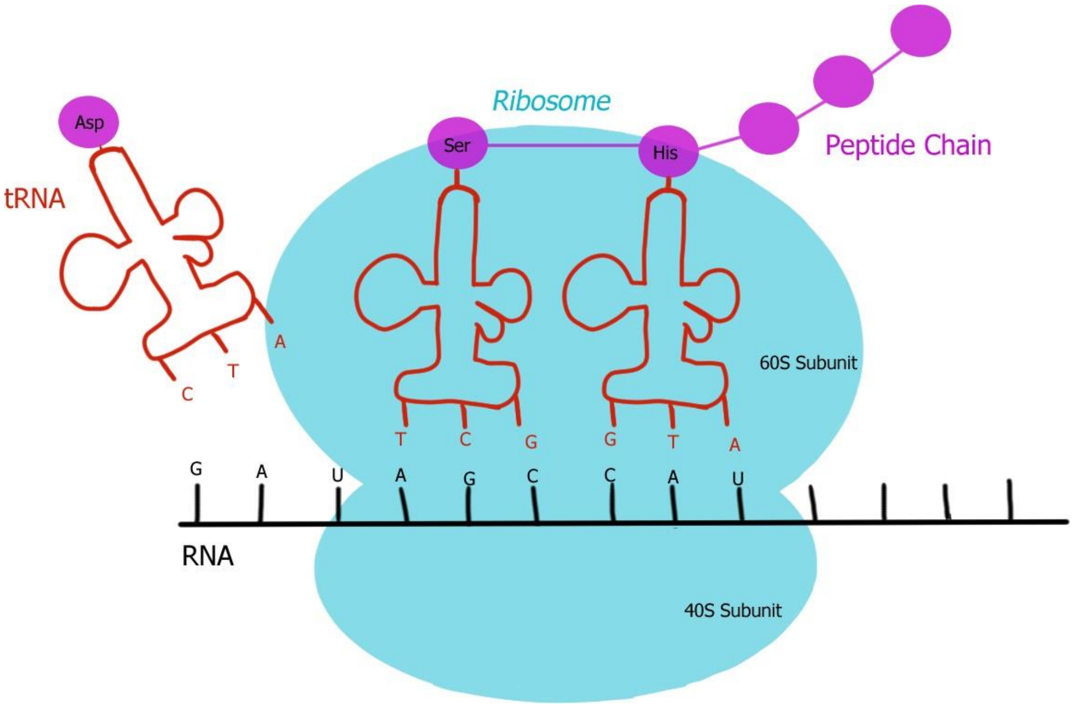

Figure 2 – the elongation step of DNA translation. The ribosome is a protein comprised of two subunits: the smaller 40S and the larger 60S subunits. A complimentary transfer RNA (tRNA) molecule binds to the respective codon on the RNA strand, and carries the matching peptide down into the ribosome. This peptide is added to the existing peptide chain, and the ribosome continues down the RNA chain until the next tRNA molecule.

Termination of translation occurs when one of the three stop codons passes into the ribosome in the A site. No tRNA can bind to these codons, so the growing polypeptide chain is released following the hydrolysation of the ribosome(Clancy, & Brown, 2008).

1.2.3 Post translational Modifications and Folding

Many proteins must undergo enzymatic posttranslational modifications (PTMs) in order to achieve the mature protein product via protein folding. Protein folding is the process by which polypeptide chains are converted into their native 3D structure, and often here the protein becomes biologically active (Cheriyedath, 2019). This process occurs in the cell's endoplasmic reticulum (the ER) and the Golgi apparatus, located within the cytoplasm. Conformational changes via methods such as ubiquitination and phosphorylation are introduced via PTMs, and increase protein stability and control localisation (Leach, & Brown, 2012).

PTMs can be reversible and irreversible chemical changes to the polypeptide chain following transcription and translation (Uversky, 2013). These modifications can occur on the amino acid side chains, or directly onto the protein backbone via the protein's C or N termini (Voet, 2006). Modifications can range from the addition of groups to the cleaving of peptide bonds, and often the sites targeted by PTMs often contain a functional group that can act as a nucleophile during the reaction. Functional groups such as phosphates can also be added through PTMs. The most common PTM is phosphorylation, and is an important mechanism in the regulation of enzymatic activity (Khoury, Baliban, & Floudas, 2011). Additionally, another common modification is glycosylation; the addition of sugar or glycan to helppromote protein folding and alter stability and function, and to act as protein destination signalling (Eichler, 2019).

Proteins have layers of structure that form the 3D structure beginning with the first structure: the amino acid sequence. This sequence determines the types of interactions between various atoms and regions within a protein as it folds. The second layer to a protein is the secondary structure; architectural structures that branch out into one dimension. α-Helices and β-sheets are included structures within the secondary structure layer, with α-helices being the most common structure within proteins. Both structures are held together by hydrogen bonds (Parker, Schneegurt, Thi Tu, Lister, & Forster, 2017) They form the backbone of the protein, and provide support to the folding process. Following the secondary structure, the tertiary structure of the protein is formed where the α-Helixes and β-sheets are further folded in a three dimensional structure. These helixes and sheets can contain amphipathic (hydrophilic and lipophilic), hydrophilic, or hydrophobic portions, which aids in the tertiary structure formation as the hydrophilic sides will fold to face the aqueous environment. The hydrophobic proteins will fold to either be in the centre of the protein, or be facing towards the centre. Once tertiary structure has formed and stabilised by hydrophobic interactions, disulphide bridges and covalent bonds may form also (Haim, Neubacher, & Grossmann, 2021). The final layer of protein structure isn't observed in all proteins and is termed the quaternary structure, and involves the assembly of multiple tertiary structures (subunits) to form a protein. Through this, many polypeptide chains can be folded around each other and interact to form the quaternary structure.

Protein folding is controlled by many molecular interactions such as the thermodynamic stability of the protein, the hydrophobic interaction, and the disulphide bridges formed (LibreTexts, 2021). The largest factor in dictating if a protein is able to fold is the thermodynamic properties. Since protein folding is a spontaneous process, the Gibbs free energy of the folding must be negative. In order for this value to be negative, then either the enthalpy or entropy (or both) of protein folding must be favourable due to the direct linkage of the Gibbs free energy to these properties (Voet, Voet, & Pratt, 2014). Minimising the hydrophobic interactions on the protein side chains helps to reduce the energy of folding. Hydrophobic regions of the protein orientate themselves towards the centre of the protein, away from the aqueous external environment. Water molecules from this environment tend to aggregate around the side chains or hydrophobic regions of the proteins generating a water shell (Cui, Ou, & Patel, 2016). This shell reduces the entropy of the system by forming an orderly layout around these regions, and it is this interaction that causes the hydrophobic collapse. This collapse is the inward folding of the hydrophobic regions into the protein, releasing the shell water of ordered water molecules, thus reintroducing entropy back into the system. The interaction of the hydrophobic core greatly increases the stability of the protein after folding, largely through van der Waals forces (in particular, London dispersion forces). Finally, disulphide bridges also play a role in the folding of the protein. These are sulphur to sulphur bonds that link non-adjacent cysteine residues, that are a stable part of a protein's final structure (Fu, Gao, Liang, & Yang, 2021). These bonds commonly help a protein fold back and link onto itself, and bonds between cysteine are very stable once created.

Splice variants arise from the alternative splicing of the introns and exons within the gene. The splicing process is catalysed by the spliceosome, a protein-RNA complex containing over 100 proteins and five small ribonucleoproteins (snRNAs) (Abramowicz, & Gos, 2018). The nature of these snRNA allows for the formation of RNA-RNA complexes and identification of the splicing sites. Errors within the splicing process can lead to improper removal of introns, altering the open reading frame of the gene. The *cis* elements within the process are crucial to identifying the splicing sites, and are known as the consensus splice sites. In general, splice variants arise when point mutations occur at the consensus site leading to incorrect identification of exons and introns (Sterne-Weiler, & Sanford, 2014). There has been clear links to suggest that splicing variants may be the cause of genetic disorders through alteration of the splicing pattern, and that there is potentially misclassification of mutations that are in fact splicing disorders.

1.3 When Biology goes Wrong: Mutations

Variation in the genetic code via changes in amino acid sequences through mutations have been known to influence disease for many years. Mutations affect proteins through many different changes, ranging from altering protein folding and stability to changing protein expression and localisation. Changes to protein observed through mutations can result in loss of protein function and protein-protein interaction sites, or sometimes may even be beneficial to the organism (Clark, Pazdernik, & McGhee, 2019).

There are many different types of mutations, most of which have minor effects or even no noticeable effects, due to humans carrying two copies of a gene. Unless the mutant is dominant, the second allele counteracts the mutation (Clancy, 2008). DNA sequence can be altered through a number of different mutations, and can have a number of different outcomes and presentation. Mutations that affect a single base are otherwise known as point mutations. These mutations are listed below:

- Base Substitution
- Deletion (can also affect many bases)
- Insertion (can also affect many bases)

1.3.1 Base Substitutions

Base substitutions occur when one nucleic base is replaced by another, thus altering the sequence, shown below in figure 3. There are two different types of substitutions: transitions and transversions. During a transition substitution, a pyrimidine is replaced by a pyrimidine, and a purine is replaced by a purine. During a transversion substitution, the base is replaced by the opposite type (eg, purine to pyrimidine).

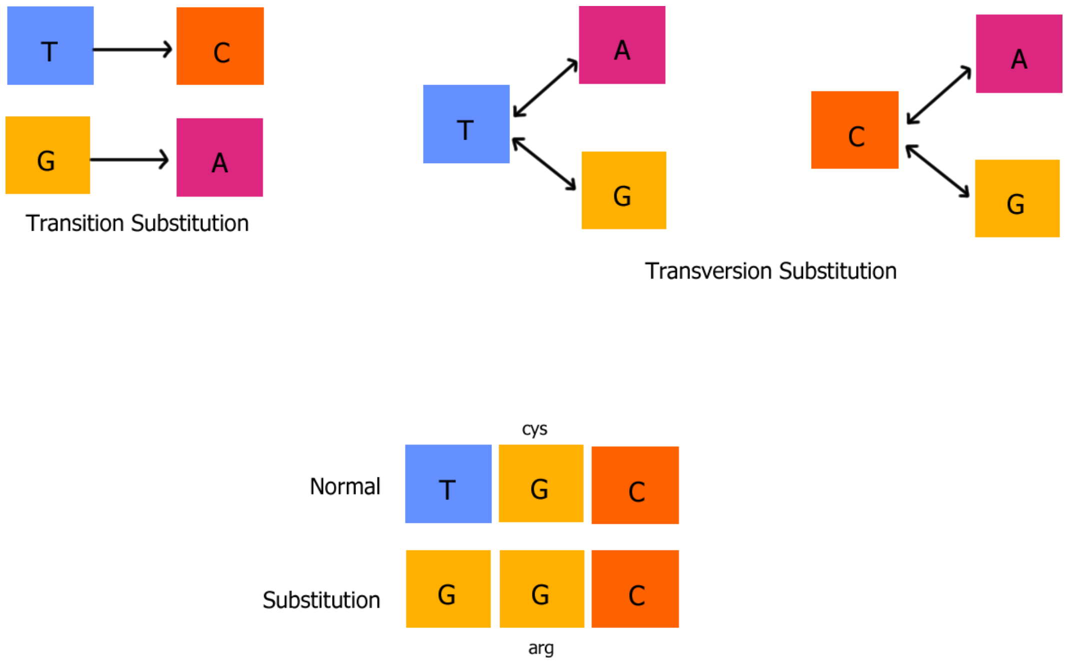

Figure 3 – DNA substitution mutations. The top section shows the difference between a transition substitution (a change from one nucleotide to another of the same type), and a transversion substitution (a change from one nucleotide to another of a different type). Below is an example of a substitution mutation The example is affecting a single base position and is therefore a point mutation. Here, the amino acid cysteine is changed to arginine.

1.3.2 Deletions

Deletion mutations denote the removal of a base, or bases, from the genetic sequence, shown in figure 4. The effect of the deletion largely depends on the number of bases removed from the sequence. Point deletions (affecting a single base) can still have a large impact on pathogenic presentation if they fall within a reading frame for a coding gene. Despite this,effects are mostly seen from gross deletions of a number of bases.

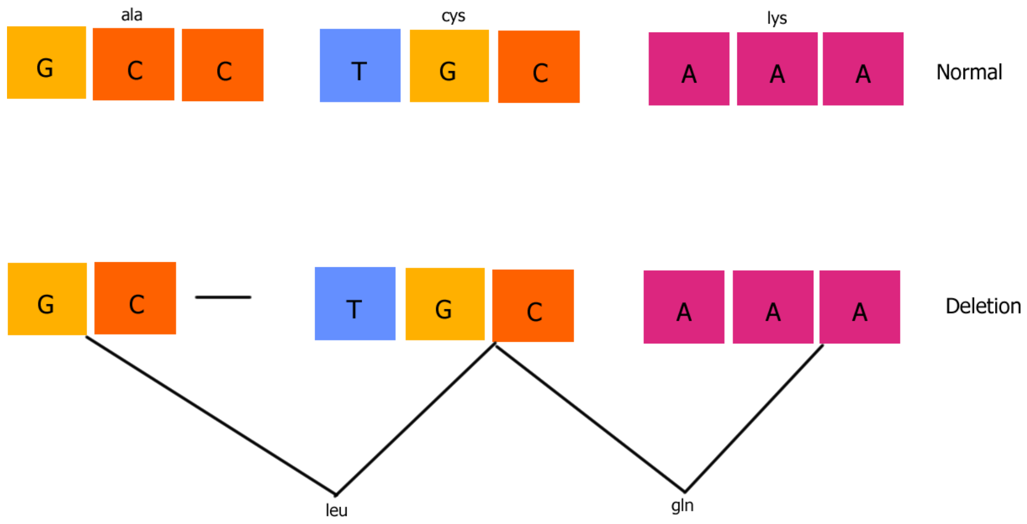

Figure 4 - An example of a deletion mutation The example is affecting a single base position and is therefore a point mutation. Here, a frameshift mutation is observed; the deletion of a single base alters the reading frame so different codons are used.

1.3.3 Insertions

Insertion mutations are the opposite event to deletion mutations; they denote the addition of bases to the genetic sequence, shown in figure 5. Genes can be inactivated through the addition of extra bases. When a foreign segment is added to a gene, the gene is said to be disrupted and is usually completely inactivated, though the effect varies depending on the amount of sequence added and the location.

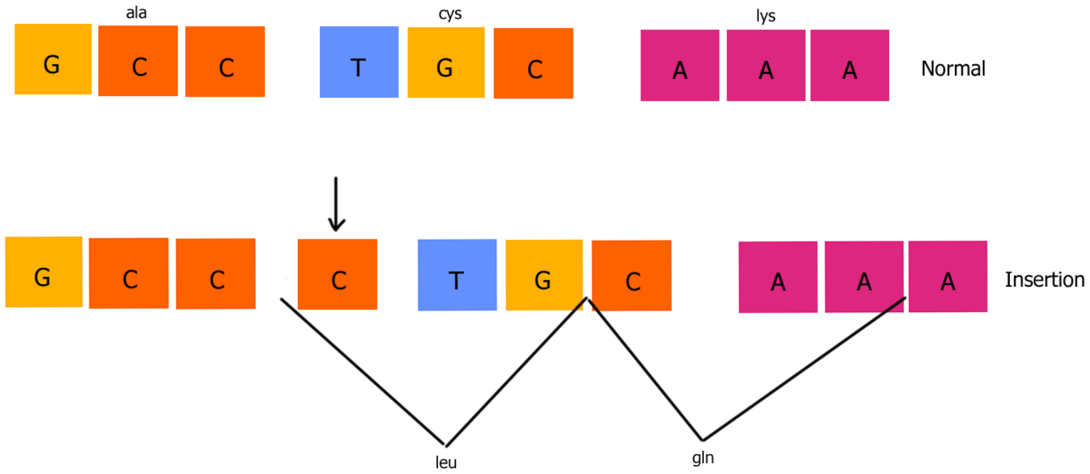

Figure 5 - An example of an insertion mutation The example is affecting a single base position and is therefore a point mutation. Here, a frameshift mutation is observed; the insertion of a single base alters the reading frame so different codons are used.

1.3.4 Causes of Mutations

There are a range of ways mutations can be introduced into the genome including errors from mispairing and recombination, spontaneous mutations from DNA polymerase errors, chemical mutagens, radiation, and tautomerisation.

Damage to DNA caused by chemical mutagens is known as induced mutations. Toxic chemicals that react with DNA bases and alter their properties are the most commonly observed form of mutagens. Nitrile, for example, converts the amino acid group on cytosine to a hydroxyl group causing a change to uracil (Alsøe et al., 2017). The altered bases are misread by DNA polymerase during replication, and incorrect bases are placed in place of these in the new strand. Other chemical mutagens include base analogues and intercalating agents. Base analogues mimic the natural bases, such as bromouracil which mimics the shape of thymine. Within the cell, bromouracil is converted into bromouridine triphosphate and is inserted into the strand in place of thymine. Bromouracil can also be observed in an alternate form, where it resembles cytosine and is paired with guanine (Holroyd, & van Mourik, 2015). Bromouracil can alternate between these isoforms. When the second isoform is observed within a template DNA sequence, a guanine base is inserted into the new strand instead of adenine. During the process of intercalation, a foreign mutagen such as acridine orange is inserted between two bases on a single DNA strand, interrupting the base pairing (Husain, Ishqi, Sarwar, Rehman, & Tabish, 2017). Where an intercalating agent is observed during replication, an additional base is inserted in the place of the mutagen, leading to an insertion mutation.

Recombination occurs between closely-related DNA sequences on the same strand, and can occur either intra-strand and between the corresponding chromosomes (Guirouilh-Barbat, Lambert, Bertrand, & Lopez, 2014). When two gene copies exist in the same orientation, intra-strand recombination forms a loop between the two copies. Through this, a circular closed fragment is made with the sequence between the first and second gene copy. This fragment is discarded, and the strand remains with one copy of the gene, potentially resulting in deletion of other genes. When two copies are seen, there can also be recombination between the two corresponding chromosomes. Mispairing of the copies and crossing over results in a duplicate copy on one chromosome, and a deletion on the other chromosome leaving one copy and nothing downstream between the two.

Errors can occur during DNA replication, and while human DNA polymerase has high fidelity, over time these errors can result in a serious mutation within a coding gene (Pandey, 2020). Errors in replication are often corrected via the proof-reading system of the DNA polymerases, however there exist cells where this mechanism has been destroyed or disabled due to mutations in the polymerase gene itself leading to a large increase in the number of spontaneous errors. There also exist cases when the sequence is a large number of repeating bases, or short tandem repeats, where DNA can slip and become misaligned between the template and the new strand. Cases of this exist within humans involving trinucleotide tandem repeats, such as the CAG trinucleotide repeat within Huntington's Disease (Nakamori et al., 2020). DNA slippage can lead to deletion of the repeat or insertion of further repeats into the new strand.

Rare cases of tautomerisation of the bases can occur and lead to mis-paired bases. Each nucleotide base exists in two forms: the main stable isoform, and the secondary less stable tautomers. If the second tautomer is within the cell when replication occurs, it can be incorporated into the new strand. Both thymine and guanine have keto and enol tautomers, with the keto form appearing most commonly (Li et al., 2014). When the enol form is observed, thymine instead base pairs to guanine while the enol form of guanine base pairs to thymine. Meanwhile, adenine and cytosine exist in the stable amino form and the inimo form. Adenine while in the inimo base pairs with cytosine, but both forms of cytosine pair with guanine. While these mutations are rare, the probability of the wrong tautomer being incorporated increases with the temperature (Gheorghiu, Coveney, & Arabi, 2020).

1.4 Single Nucleotide Polymorphisms

A SNP is a substitution of a single base (a point mutation) within the genome. They must be largely present within the population to be classified as a SNP. A large number of studies have shown that SNPs can have biological effects, with particular importance placed on those that generate pathogenic effects. These involvementsrange from association in complex diseases, to differing reactions to treatments and medication. Genome wide association studies (GWAS) show that there is often an association with SNPs that increase susceptibility to certain complex diseases such as cancer and heart disease (Nguyen, Huang, Wu, Nguyen, & Li, 2015). GWAS is the rapid observational study of genome-wide genetic markers to find associations between genetic variants and a trait, such as a complex disease (NIH, 2020). Often these studies include both individuals with the trait, and control individuals without the trait in order to completely assess the association of a trait with genetic variants. There has also been association studies for SNPs that reduce susceptibility to complex disease, such as SNPs within the apolipoprotein E (APOE) gene reducing risk of Alzheimer’s disease. These SNPs remain in the genome for the duration of a person’s life, making them powerful diagnostic tools over other current techniques, such as microarray gene expression assays taken from specific tissues (Batnyam N. et. al., 2013).

SNPs can occur anywhere in the genome, and their location directly results in the biological impact they may confer. They can be found in the coding regions of genes, non-coding regions of genes, and in the intergenic regions. SNPs within the coding region can be classified into two main categories: synonymous and non-synonymous. Synonymous SNPs are those that do not change the protein sequence, and non-synonymous SNPs are those that change the amino acid sequence of the protein. These non-synonymous SNPs are broken down into three further categories of mutation: missense and nonsessense (Bethesda, 2005). Often a direct change in the coding region of a gene does not impact the protein product produced due to the degeneracy of the genetic code. SNPs in thenon-coding region can still have biological effects as they can affect the splicing of genes, transcription factor binding, the sequence of non-coding RNA, and mRNA degradation. The altered expression seen by these types of SNPs is referred to as expression SNPs, or eSNPs. They can be upstream or downstream of a gene.

To date, there are 1,071,975,857 validated human SNPs in the NCBI dbSNP database, with 98,202 confirmed to be pathogenic, 170,131 confirmed to be benign (NCBI dbSNP, 2021), figures shown in table 1. The remaining SNPs are likely pathogenic, likely benign, a drug response, or of other clinical significance. On ClinVar, there are currently 895,843 single nucleotide variants. On here, 59,273 are confirmed to be pathogenic, and 119,392 are confirmed to be benign. ClinVar also shows SNPs of unknown or uncertain clinical significance, and currently there are 373,831 SNPs of unknown significance (NCBI ClinVar, 2021). The remaining are either likely to be pathogenic, or likely to be benign. Due to the large scale discovery of SNPs, there has been a growth in the interest in SNPs, particularly within the field of disease biomarkers and their role in complex diseases.

Table 1 – the number of human SNPs recorded in the dbSNP database and the ClinVar database (filtered for "single nucleotide mutation"). The number of characterised SNPs in disease highly varies between the two, and the data shows the number of SNPs that are unknown pathology. Data modified from the dbSNP and ClinVar databases from searches taken place in September 2021.

Database	Total SNPs	Pathogenic	Benign	Unknown	Other
dbSNP	1,071,975,857	98,202	170,131	-	1,071,707,524
ClinVar	894,650	59,238	119,364	373,464	342,584

1.4.1 Application of SNPs

As interest has grown in SNPs, further applications of SNPs have been discovered. Such applications have been identified and researched in areas such as biomarkers for complex disease, forensic evidence, and to study population genetics.

Population genetics: As next generation sequencing (NGS) improvements continue to advance the ease of high through-out genotyping and sequencing, population genetics has seen a shift from the use of microsatellites to direct analyses of sequence variation, such as SNPs (Grover, & Sharma, 2013). SNPs make for good genetic markers due to the availability of annotated markers, low error scores, the ease of calibration in the laboratory setting (as opposed to length based markers), and the ability to combine data sets from multiple laboratories. Not only this, but SNPs offer a simple mutation model, and the ability to study neutral variation, and variation under selected regions. Microsatellites have far larger allelic diversity over SNPs, but despite this SNPs show high promise as informative markers, even outperforming microsatellites in a small number of studies for population structure analysis. The advantage of SNPs over microsatellites lies in how strongly individual SNPs can segregate in a given population (Heylar et al., 2011).

Forensics: When there is too little template DNA, or the DNA is too degraded, alternate markers are needed. The applications of NGS expand into the area of forensic work through the use of short tandem repeats (STRs), mRNA, and SNPs (Phillips, 2012). SNPs in particular offer advantages for use in forensic identification due to the abundance of markers present in SNPs throughout the genome, the fact they are easily adapted into automated processes, and that the read length can be very short (60-80bp). SNP samples can provide an insight into kinship analyses, missing persons family reconstruction, identifying human remains, and possibly even provide evidence for lead cases with no suspect. There are four categories of SNPs used in forensic analyses depending on what is needed in a case by case situation. These are identity-testing SNPs; lineage informative SNPs; ancestry informative SNPs; and phenotype informative SNPs (Budowle, & van Daal, 2018). Identity-testing SNPs offer genetic information needed to differentiate populations and people, and so exclude individuals based on if they can be the progenitor of the evidentiary sample, or if they can be a potential family member. They require high heterozygosity, and low population heterogeneity (low inbreeding). Lineage SNPs are used to identify missing persons through the use of kinship analyses to find relations. They are mostly found on the mtDNA genome and on the Y chromosome. There is a lack of recombination in these regions and a low mutation rate, making them ideal to study samples that are separated by several generations. Ancestry informative SNPs are used to provide insight into the biogeographical information of a person. From this ancestry, potential phenotypic characteristics can be indirectly inferred and applied to a case where the suspect's appearance is known. Phenotype informative SNPs allow for high probability analyses that an individual has certain phenotypic features that match the description for an investigative lead. These features include hair colour, eye colour, and skin colour.

Complex disease biomarkers: SNPs can be used as biomarkers themselves for disease, revealing an individual's predisposition to a particular disease. This gives prewarning for treatment and preventative measures. Association studies are often carried out to identify relations between genetic variation and disease. SNPs have been used as biomarkers for complex disease such as Alzheimer's disease, where early onset is caused by mutations in the amyloid precursor protein (APP), and the presenilin genes (PSEN-1 and PSEN-2) (Erdoğan, & Son, 2014). SNPs can also be used to measure levels of disease biomarkers using machine learning methods. This method is a very powerful diagnostic tool and provides a useful complementary tool to standard diagnostic methods. New polygenic scores (PGS) are trained for biomarker prediction, aiding in predictive biomarker volume from a SNP. Fluctuation of the levels of a biomarker limit the reliability of the prediction, so therefore a more stable phenotype without fluctuating levels would therefore be more suitable for use of SNP only prediction (Widen, Raben, Lello, & Hsu, 2021).

1.5 Bioinformatics

Bioinformatics is a complex and diverse field within the biological sciences that combines computer science, mathematics, statistics, and biology to develop software tools and methods to understand biological data. Bioinformatics is used *in silico,* with the development of pipelines for repeated use of analyses and as well as the integration of computational methods to standard biological studies. This discipline is an important part of many other biological fields including genetics, molecular biology, and proteomics. The driving goal of bioinformatics is to ultimately increase the understanding of biological processes through applying computationally intensive techniques, setting it apart from other biological disciplines.

1.5.1 Structural Bioinformatics

The structural information of proteins is one of the most crucial data points to understanding biological function. Structure can be separated into four different levels: primary, secondary, tertiary, and quaternary (see section 1.4.3 for more information). The primary structure of a protein can easily be derived from the sequence of the coding gene, and in the majority of cases thisprimary structure can directly determine a unique protein structure within the native environment. Structural bioinformatics takes into account the interaction between these layers of structure, and the space coordinates of atoms and bonds.

Homology is an idea that runs throughout the entire discipline of bioinformatics; if one known biological product (eg. a gene, protein, genome) is similar to an unknown, then it is likely that the unknown product has many shared properties with the known product. This is shared within the branch of structural bioinformatics, where homologous proteins are likely to share either similar structure, function, or both. Homology is used to infer which regions of an unknown protein are significant in structure formation and protein interaction. Additionally, through a method known as homology modelling (detailed in section 1.2), uses homology between protein sequences and structures to predict structure of an unknown protein. A key idea within structural bioinformatics is that the structure of a protein is directly linked to its function. From this, homology between proteins can also be used to predict the function from the structure of similar proteins. Comparisons between the structure of an unknown protein and a collection of known proteins can derive multiple plausible protein functions.

1.5.2 Large Data Storage

Protein structure data that has been experimentally derived from methods such as crystal analysis is stored into large online databases such as the Protein Data Bank (PDB), making computational access easy to obtain. The use of bioinformatics spans across multiple fields, allowing for access to genetic information, information on genetic variants, and disease informatics. Often these databases are kept separate on individual hosting sites, such as Reactome, UniProt, and ClinVar. UniProt is the most expansive and accessible database for proteins, allowing you to follow protein structure and function from sequence. UniProt incorporates data from a number of sister databases, while the UniProt Knowledgebase itself consists of two main sections: Swiss-Prot and TrEMBL. Swiss-Prot contains manually annotated and reviewed entries from literature and computational analysis carried out by curators, while TrEMBL contains automatically annotated and unreviewed entries that are queued for full manual annotation (UniProt Consortium, 2021). The advancements to high throughput sequencing led to data being produced and submitted faster than Swiss-Prot was able to annotate and review, driving the need for a second database to store these unannotated or computationally annotated entries. Table 2 details the differences in number of entries (UniProtKB/Swiss-Prot Consortium, 2021) (UniProtKB/TrEMBL Consortium, 2021) between the two different databases, highlighting the need for TrEMBL to store unreviewed entries asSwiss-Prot couldn't keep with number being inputted for review.

Table 2 – The entries within each UniProt Knowledge Database as of 2nd June 2021. Swiss-Prot contains manually annotated and reviewed entries. TrEMBL contains automatically annotated entries that are not reviewed (adapted from the UniProtKB/TrEMBL and UniProtKB/Swiss-Prot Consortium statistics data, 2021).

	Swiss-Prot	**TrEMBL**
Sequence Entries	565,254	219,174,961
Fragments	9,261	23,817,622
Additional Sequences (via splicing, initiator or promoter usage, or ribosomal frameshift)	40,563	N/A

Where medical data is concerned, ClinVar is an open accessible source for relationships between human variants and phenotypes. ClinVar can be used and queries for multiple purposes, such as searching for a particular disease or phenotype, a particular gene, or position on a chromosome or assembly. The scope of ClinVar is large, and covers variants found within any part of the human genome, including the mitochondria. The variants covered within ClinVar can be of any length; single nucleotide polymorphisms (SNPs), small insertions or deletions, or to full copy number changes and cytogenetic rearrangements (Landrum et al., 2015). Within ClinVar, the variants recording have been observed both within a clinical environment and a research setting. Those discovered or have had further research studies add depth to the clinical significance through experimental evidence.

Molecular pathway data is stored within the Reactome database which is a bioinformatics database forthe visualisation, interpretation, and analysis of pathway knowledge. The data inputted into the database is peer-reviewed and curated by the Reactome team (Reactome Organisation, 2021). Reactome is a novel platform using a relational database of signalling and metabolic pathways, and their relationships sorted into biological pathways and various processes. Thecore of the Reactome system is the reactions, while entities such as nucleic acids, proteins, complexes, vaccines, etc. that are part of the reaction form the network of interactions are grouped within the pathways.

Connecting these databases above can yield potentially crucial new information about the links between disease, pathway, and variants. Linking through data through all steps of the biological process from gene transcription and translation, to protein generation, will allow for more in depth analysis of variants and localising where pathogenic elements happen within a biological system.

1.6 Homology Modelling for Structural Bioinformatics

Assessing the impact that individual variants had on structure and interacting regions relied on the use of protein modelling methods to generate visual aids and representations. Homology modelling is considered the most reliable, and easiest, form of structural prediction modelling and is frequently used for many biological applications. Homology modelling makes use of the primary (1D structure) amino acid sequence, and builds up from this to generate secondary structures (2D structures) and finally tertiary structures (3D structure). The amino acid sequence of a given protein often contains enough information to obtain a reliable 3D structure prediction. It is often assumed that protein function is closely related tosequence composition -and that the structural resemblance of proteins implies structural similarity (Krissinel, 2007). Homology modelling usually consists of four steps: sequence alignment**,** multiple sequence alignment, model construction, and finally model refinement (Grumezescu, 2018).

1.6.1 Sequence Alignment

The target sequence that is being queried is named the model sequence, while any sequences with known structures are named templates.
If the sequence similarity alignment of two proteins is above a certain threshold, then it is highly likely that they fold into the approximate same structure. Likewise, if this falls below the threshold (known as the 'protein twilight zone' as the threshold limit is highly debated though commonly taken as 30% similarity (Khor, Tye, Lim, & Choong, 2015)), then it is inconclusive if these proteins will fold approximately the same or very different. Similarity tools such as BLAST (for local alignments) and FASTA (for global alignments) are used to determine the sequence identity between the model and the template proteins (Makigaki, & Ishida, 2020). Both programmes utilise a scoring matrix (BLOSUM) to compare the model sequence with all sequences with a known structure in the Protein Data Bank (PDB) repository. A list of template sequences is returned, often ordered from highest similarity downwards. When doing homology modelling, it is possible to take the singular highest template, or use a combination of templates above a set threshold. The latter approach uses multiple sequence alignment (MSA) and is used by tools such as Swiss-Model (Padmanabhan, 2014).

1.6.2 Multiple Sequence Alignment

MSA is the process in which three or more biological sequences (generally DNA, RNA, or protein) are aligned. In many cases these sequences are thought to have an evolutionary relationship; where they share a common ancestor or share a linkage. From anMSA, sequence homology can be obtained. Often, these results are used for phylogenetic analysis and ancestry studies. MSA results can be used across nearly all of bioinformatics. Within structural bioinformatics, these can also be used to assess sequence conservation across secondary structure, tertiary structure, protein domains, individual amino acids, or nucleotides (Thompson, Linard, Lecompte, & Poch, 2011). Some alignments methods, such as 3D-Coffee and PROMALS3D use tertiary 3D protein structures to improve sequence alignment, but these methods rely on the tertiary structure of proteins and is not suitable for all proteins as a result (Deng, & Cheng, 2011).

Once an initial template list has been generated using the sequence alignment step detailed above, it can be further refined and corrected using MSA. Such regions that benefit from correction include those where the percentage similarity is very low and so it is difficult to align the template structures to the model structure. Using inherent structural information from this alignment better informs specific features in regions, such as hydrophobic residues, and increases the final model performance (Chatzou, Magis, Chang, Kemena, Bussotti, Erb, & Notredame, 2015). There are tools that use these position-specific scoring matrices to better improve models, such as T-Coffee.

1.6.3 Model Construction

Once alignment has been carried out, construction of the model can begin. The model construction process begins with the protein backbone generation. Where there are multiple templates, an average structure is taken, where the template structure distribution is weighted by the local similarity identity. In cases where a single model is used, the backbone coordinates of aligned residues from the template are transferred into the model (Saxena, Sangwan, & Mishra, 2013). The last step of model construction involves generating the residue side chains. If the percentage similarity of the template sequences are over 40% at any residue, then the side chain conformation is transferred directly into the model, conserving the orientation. Where low levels of similarity are seen, another knowledge-based approach is applied. Rotamer libraries are scanned, and each conformation within the library is scored with a number of energy functions and the highest scoring conformation is used in the model (vlab.amrita.edu., 2012).

1.6.4 Loop Modelling

Loops often represent regions of high disorder in protein, which in turn are often linked to regions of unaligned sequence from the sequence alignment. The more amino acids there are in a loop, the more inaccurate the loop becomes (Adhikari, Peng, Wilde, Xu, Freed, & Sosnick, 2012). There are two main approaches to loop modelling; energy based and knowledge based. Energy based loop modelling determines the quality of a loop by using an energy function, and so finds the best conformation using a statistical model. The loop is then subjected to energy minimisation using molecular dynamics techniques or a molecular dynamics simulator (Tang, Zhang, & Laing, 2014). MODELLER uses this non-template based technique. Knowledge based loop modelling (or template based) utilises the PDB by scanning for loops of similar length with relative end-point geometry from proteins with known structure (Soto, Fasnacht, Zhu, Forrest, & Honig, 2008). The identified protein (or proteins) is then aligned to the target protein gap, and the loop coordinates are transferred. The quality of the loop depends on thequality of the alignment; since loops are the least conserved regions, a known template cannot always be found that aligns with the gap in the target protein. SuperLooper is a specialised protein loop structure predictor that uses this template based method (Health Sciences Library System, 2014).

1.6.5 Model Refinement

To accurately solve the model structure, the main backbone chain of the protein must be energy minimised to find the best confirmation for the side chains. The model can be optimised by running a MD simulation. ModRefiner is a commonlyused tool to solve this and provide protein structure refinement using MD simulations.
ModRefiner constructs and refines protein structures using C-alpha traces based on a two-step, atomic level energy minimisation process (Dong & Zhang, 2011). Main-chain structures are first constructed from the C-alpha traces, and then the side chains are refined alongside the backbone atoms using composite physics and knowledge-based force field solutions. In chemistry, a force field is the functional form (the potential energy in bonded forms that describe electrostatic or van der Waals forces) and various parameters set to calculate the potential energy of a system of atoms, and estimates forces between the atoms within a molecule. These parameters are typically set through previous physical experiments (physics based), or through calculation by quantum mechanics (knowledge based) (Frenkel & Smit, 2007). All-atom force field systems provide parameters for all atoms in the system individually, including hydrogen atoms (Raval, Piana, Eastwood, Dror & Shaw, 2012). The two steps that Modrefiner uses includes a low-resolution generation of the backbone, followed by a high-resolution all-atom refinement guided by these force fields with two different parameters (knowledge based and physics based).

Model refinement can also be done through iterative threading, as used in I-TASSER (Roy, Kucukural, & Zhang, 2010). The process can be seen below in figure 6. This platform generates and refines protein models through creating a 3D-structure from multiple threading alignments and iterative structural function assembly. Refinement by iterative threading involves an additional fragment assembly step based on the clusters selected in the second step (structure assembly). This structure re-assembly has the same I-TASSER potential, but additional restraints are placed upon the construction by pooling previous threading alignments and PDB structures closest to the cluster. This second iteration is designed to remove steric clashes and refine the protein topology. I-TASSER also suggests a protein function for the model by searching for structurally matching the 3D model with known proteins in the PDB (Zhang, Freddolino, & Zhang, 2017).

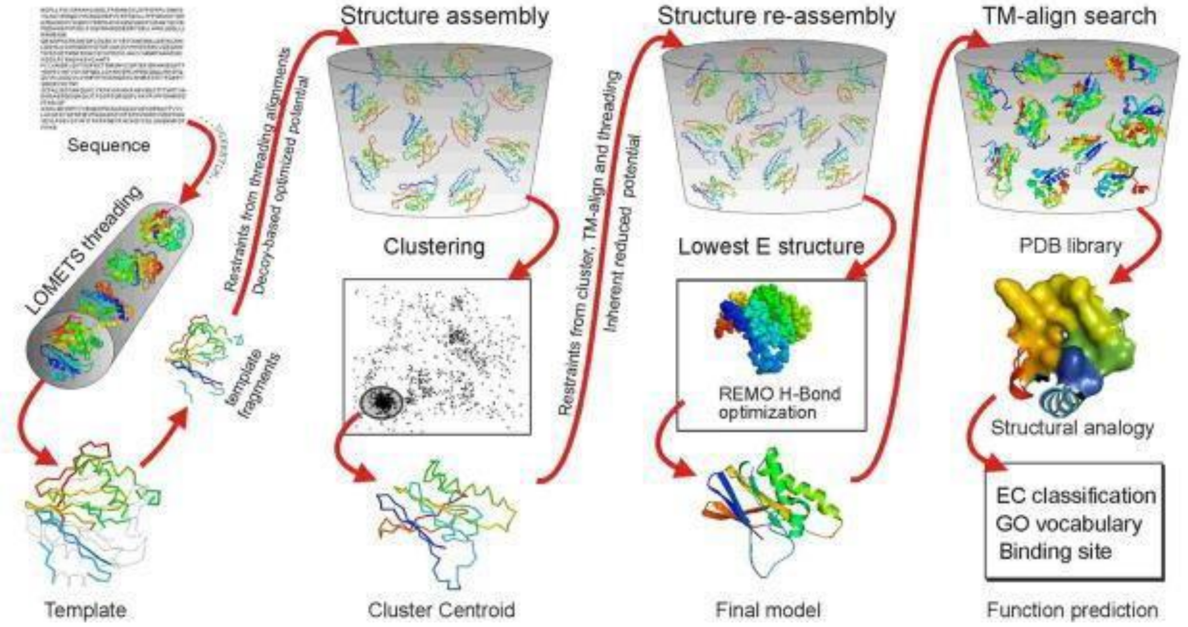

Figure 6 – The schematic representation of the I-TASSER protocol. Threading can be seen in the first step, where the query protein is matched against non-redundant proteins in the PDB, then threaded through a representative PDB library. The structural assembly within I-TASSER is conducted ab-initio, outlined in section 3.4.3. Iterative threading occurs in the third refinement step, where fragment assembly occurs a second time dependent on the cluster selected in the step before. The second iteration removes steric clashes and refine the topology of the protein. In the final step, the protein function is inferred by structurally matching the model with the PDB library (Roy, Kucukural, & Zhang, 2010) .

1.7 The BRCA1 Gene and Protein

Even before the discovery of the BRCA genes, it was thought that there were genetic origins for familial breast cancer and later ovarian cancer. Family studies conducted by Henry Lynch were the first to characterise this (Murthy & Muggia, 2019). It was later discovered through cloning andidentification that the genes for these diseases were located on chromosome 17, and that two genes were responsible. These genes were termed *breast cancer susceptibility gene 1 (BRCA1)*, and *breast cancer susceptibility gene 2 (BRCA2)*. BRCA1 has been found to be expressed within a large number of different tissues, ranging from cervix, liver, uterus, prostate, pancreas, lung, kidney, bone, brain, to the lymph nodes, skin, and bladder. As such, BRCA1 has been associated to have a role within other cancers beyond that of familial breast and ovarian cancer.

The *BRCA1* gene encode for proteins that are involved in the DNA repair mechanism, and are involved in a number of different cellular pathways vital to genomic stability such as chromatin remodelling, protein ubiquitination, transcriptional regulation, and apoptosis (Liu et al., 2021). The gene product of *BRCA1* is heavily involved in tumour suppression, and loss of the BRCA1 protein is associated with failure of homologous recombination (HR), a system of DNA repair. BRCA1 translocates to the DNA damage site, where it coordinates both the DNA damage repair and the DNA damage signalling, where through multiple steps BRCA1 promotes the use of HR to repair double-strand breaks (DSBs) in DNA (Liu, & Lu, 2020). Exogenous or endogenous DNA damage via harmful agents or mechanism failure, can result in a variety of changes including DSBs, single-strand breaks (SSBs), base damages, intrastrand cross-links, and interstrand cross-links (Zhang, 2013).

BRCA1 contains multiple functional domains, including a highly conserved zinc finger at the

N terminal contributing to E3 ligase activity. It is also noted that BRCA1 both directly and indirectly interacts with a number of other molecules. The earliest, and most direct, indication that BRCA1 was involved within HR repair was its association with RAD51, a homolog of the yeast protein involved with HR (Tavares, Wright, Heyer, Le Cam, & Dupaigne, 2019). BRCA1 has also been identified to interact with another repair protein, RAD50, forming a distinct nuclear foci from that of RAD51. The loss of BRCA1 functions leads to defects within the S phase and the G2/M phase of the cell cycle, and spindle checkpoints (Mylavarapu, Das, & Roy, 2018). Defects within these phases causes an increase in genetic instability, consequently leading to an increase in DNA damage response; increasing the risk of tumour formation.

Glycosylation was one of the first biomarkers for cancer due to the extensive role of glycosylation in the process of the cell cycle, and ultimately the ability to evade cell cycle checkpoints. Specific glycosylation motifs can control certain cell signalling pathways such as abnormal growth factor signalling, a critical feature of cancer tumour growth. Changes of glycosylation within cancer cells typically involve an increase of certain patterns, such as an increase of sialyl Lewis structures and *N*-glycan branching, and the exposure of the mucin-type *O*-glycan (Reily, Stewart, Renfrow, & Novak, 2019). *N*-glycans refer to the attachment of N-acetylglucosamine (GlcNAc) to the nitrogen atom of specific side chains. *O*-glycans refer to the glycosylation of amino acids with a functional hydroxyl group (eg. serine and threonine), and are abundant on many extracellular and secreted glycoproteins including mucins. Sialyl Lewis structures are generated through extrinsic glycosylation events, where soluble glycan-modifying enzymes circulate in the blood and conjugate a monosaccharide extracellularly onto an existing sugar structure (Mulloy, Dell, Stanley, & Prestegard, 2017). As a cancer cell evolves, the changes in glycosylation pattern can be tracked in parallel with the metabolic changes. It is known that receptor tyrosine-protein kinase erbB2 (*HER2*) is overexpressed within many cancers, including familial breast cancer. *HER2* is a protooncogene encoding for epidermal growth factor, and in breast cancer *HER2* is overexpressed by 15% - 20% and is linked to protein overexpression (Ahn, Woo, Lee, & Park, 2019). *HER2* screening is possible for breast cancer prognosis, and is associated with a high rate of recurrence and morality. It currently remains the only marker for *HER2*-targetting agents, such as trastuzumab. Identification of other markers for these targeting agents would greatly increase the prognosis for breast cancers by increasing optimisation of screening.

It has been reported that SNPs within cancer causing genes have a relationship with a variety of different forms of cancer, including familial breast and ovarian cancer from *BRCA1.* It is known that SNPs can occur in gene promoters, exons, introns, and effect gene expression through a number of different mechanisms. Also, SNPs can cause alterations in epigenetic regulation of genes, increasing the complexity of SNP susceptibility to cancer (Deng, Zhou, Fan, & Yuan, 2017).

In 2018, a meta-analysis was conducted by Xu et al. into SNPs within the *BRCA1* gene that were associated with disease. Within this analysis, four gene polymorphic variants were selected using specific criteria. The selected variants were rs799917, rs1799950, rs1799966, and rs16941. It was discovered that rs799917 was able to decrease the risk of various cancers within Asian populations, rs1799950 could decrease risk of breast cancer within Caucasian populations, and rs16941 could increase the overall risk for any cancers. Despite these findings, it was noted that the sample size and the number of cancer types within this study were limited. Therefore, the overall conclusion was that more research needs to be conducted into SNPs in *BRCA1*, and their impact on cancer risk.

Furthering this work, in early 2021 a novel study was conducted by Coignard et al. was undertaken to expand the knowledge of SNPs within breast cancer and the *BRCA1* gene.

The study was based off of GWAS data from breast cancer cases within the Breast Cancer Association Consortium (BCAC) and mutation carriers from the Consortium of Investigators of Modifiers of *BRCA1/2* (CIMBA). This novel study was the first analysis of genetic modifiers of breast cancer that examined the difference between common genetic variants with breast cancer risk within the general population and in women with *BRCA1* (or 2) mutations. This novel study was the first analysis of genetic modifiers of breast cancer that examined the difference between common genetic variants with breast cancer risk within the general population and in women with BRCA1 (or 2) mutations. This study was able to identify eight novel SNPs within the *BRCA* genes associated with breast cancer risk; 4 within *BRCA1* and 4 within *BRCA2,* which had not been reported in previous association studies. Research into the precise effects of SNPs in both these genes in mutation carrier could provide insights into the biological mechanism of cancer development. The study concluded that more detailed mapping and functional analysis is required to not only elucidate the role of the newly discovered variants, but also other variants.

There is a large number of BRCA1 variants with unknown clinical significance; as of early September 2021 there were 2,868 unknown significance single nucleotide variants on ClinVar (NCBI ClinVar, 2021). Table 3 below details the number of variants within each category on ClinVar. The data on ClinVar highlights a growing and pressing need for the quick determination of clinical significance – with current sequencing technologies, the number of novel SNPs will continue to grow and be discovered. *In silico* methods of analysis of known variants could be undertaken to prioritise variants that are potentially pathogenic, accelerating the clinical study of these variants. Identifying these uncertain variants aids with therapeutic treatments and management, as well as early diagnosis and prevention of further cancers (Kim et at., 2021).

Table 3 – Search results of BRCA1 on ClinVar as of September 2021. Results were filtered by single nucleotide variants and the total variants by clinical significance are shown. Data adapted from ClinVar, 2021.

Clinical Significance	Number of Variants
Pathogenic	756
Benign	633
Likely Pathogenic	147
Likely Benign	1,586
Uncertain Significance	2,868
Conflicting Interpretations	394
Risk Factor	0

A recent reclassification of clinical variants with uncertain significance of both BRCA1 and 2 was performed in a Korean Hereditary Breast Cancer (KOHBRA) study by Kim et al. in 2021. Variants were reclassified using ClinVar data and data from the Korean Reference Genome Database (KRGDB). The odds ratio for each SNP was calculated using Korean population data from the KRGDB by the Wald Chi-Squared Test. The confidence interval was calculated via the same method. The OR was calculated based on the occurrence of the variant in the 2403 patient cases and from the KRGDB. Roughly two thirds of the variants in the KOHBRA study were reclassified as benign or likely benign, which included a total of 69.42% of all patients with a BRCA1 mutation. In this study, six unclassified variants were reclassified as pathogenic or likely pathogenic, four of which occurred as BRCA1 mutations. All six variants were supported by evidence by past studies and were not reported in the general Korean population database. Additionally, the mutations were all found to be deleterious, which account for nearly all pathogenic variants. As the study suggested, patients who are found to have these mutations can be referred for counselling or management, such as genetic familial testing and risk-reducing medication and surgery.

1.8 Project Aims

The main aim of this project is to establish the role of BRCA1 SNPs within disease, and apply this information to SNPs with unknown clinical significance in order to classify them. The large wealth of data stored in various data sources can provide links between the function of BRCA1 in pathways, the structural impacts of SNPs, and potentially reveal the role of SNPs in disease. This project hopes to be able to not only classify unknown SNPs, but also reveal where in pathways these SNPs have an effect. Additionally, identifying interactions that are interrupted by pathogenic SNPs could provide new drug targets. Classifying unknown SNPs aids in the diagnosis of BRCA1 related diseases, where patients with thesevariants can be treated early before the spread of the worst effects of disease. It is hoped that these reclassifications can ultimately improve prognosis of cancer patients, and those with other BRCA1 related diseases.

A number of target questions for this project have been generated and can be seen below:

- What are the direct pathway associations with BRCA1?
- Where do most variants occur in BRCA1?
- Can a structural model be generated for BRCA1?
- What are the structural impacts of SNPs in BRCA1?
- Are there any trends within the structural impacts of SNPs that could be used for profiling SNPs as pathogenic or benign?

2. Methods

2.1 Preliminary Study

A preliminary study was carried out into the nature of the BRCA1 gene, and its variants using Human3DProteome and ESP. The gene was mapped from chromosome, to gene, to protein, to metabolic pathway, and then to disease. The number of variants was recorded in table 4.

Table 4 - Preliminary research findings. The first column after the headings include the basic information, and the second column includes detailed information about the research area highlighted in the heading and specific examples.

Protein	Human_BRCA1 Human Breast Cancer type 1 susceptibility protein	Uniprot P38398-2
Gene	BRCA1	
Chromosome	17	17:41199671
SNP Variants	385 entries - pathological link often unknown	VAR_007769 selected (pathogenic link known) Missense variant rs28897696 G>A Protein change A1729V
Metabolic/Signalling Pathways	Central role in DNA damage repair by facilitating response to DNA damage. Has E3 ubiquitin-protein ligase activity - needed for tumour suppressing function. Required for cell cycle arrest, and progression from G2 to mitosis.	acetyl-CoA-carboxylase alpha (ACACA) catalyses irreversible carboxylation of acetyl-CoA to malonyl-CoA. Important first step in fatty acid synthesis.
Potential Disease	PolyPhen2 rating: Probably damaging (1.0)	Prevents ACACA binding. BRCA1 reduces ACCA activity through its phospho-dependent binding to ACCA. Control of lipogenesis.
Issues?	Many variants.	https://thebiogrid.org/interaction/697233/brca1-acaca.html

2.2 Database Creation

The preliminary research highlighted the need to generate an external database to store all the information regarding BRCA1, its variants, and pathways with which it is associated. For this, SQLite was used via a virtual machine connection to a server. Before it was possible to create the database and insert data, the database was planned to eliminate irrelevant associations and highlight any errors or additional information that was found to be required following the preliminary research. It became clear that additional information was required from the preliminary data, as well as a need to reorganise the data and remove information that didn't address to the target questions.

SQLite is a widely used relational database scripting and management tool contained within a C library (a collection of non-volatile resources written in the C language). SQLite is not a client-server model, meaning that it is only accessible via a file system rather than via a server (Schenker, 2020). SQLite saves and stores the entire database as a singular cross-platform file onto the local host machine. Due to being without a server, a SQLite database requires little to no configuration compared to client-server databases and is termed *zero-conf* (SQLite Organisation, 2021). There is no setup procedure; there are no server processes that must be stopped, started, or paused in order to use SQLite. There is also no need for access control within SQLite (granting access permissions to users) as the access is handled by the file-system permissions given directly to the database file itself. Furthermore, there are no troubleshooting processes needed to use SQLite. These advantages and ease of use made SQLite the ideal tool to craft a database.

The entity relationship diagram (ERD) for the database can be seen below in figure 7 and lays out the order and flow the database will follow. Each entity (blue rectangle) represents a table within the database, and each of these tables contain attributes dependent on what is being stored within that table. The entities identified as necessary in this database were 'gene', 'pathway', 'sequence', 'variants', and 'disease' data. By each entity, there are a number of smaller ovals that represent the attributes of the tables; this is what is stored within the table as an entry. Taking the gene entity as an example, the geneID acts as the primary key for the entity; the primary key is the unique identity point for each data set and is underlined in attributes (IBM Cloud Education, 2019). Each entry into the table is given an unique geneID assigned from 1 onwards. Each entry also includes the name of the gene in this instance. The unique geneID prevents the same gene from being entered into the database multiple times. Entities within the database are connected via relationships, either one to one (1 - 1), or many to many (N - N) (Levene, 2005). Where the link is a one to one relationship, the tables can be directly linked to each other, such cases include where one gene can have one sequence. There are often more complex relationships than one to one, such as where one variant can cause many diseases, and a disease can be caused by many variants (N - N). Where this is seen, the entities must be linked via an associative entity (pink diamond). The associative entity solves many to many relationships and is implemented into the database structure via associative tables. Associative tables contain two or more primary keys from the tables that they map together. These primary keys that are taken from other linking tables are known as foreign keys. As such, these tables contain a number of foreign keys, each from a one to many relationship from the junctioning associative table to the individual data tables (Moes & Sheldon, 2005). The primary keys of the associative table are these foreign keys.

The design of the database takes into account the relationship between gene, variant, pathway, and disease. The nature of their relationship means that it was possible to directly link these four entities together and build the additional sections of the database around these. For instance, a pathway can have a number of genes within it, and these genes can have an individual variant, and these variants can be involved in a number of different diseases. Through this, it is possible to trace the relationship through the entire database, aiding to eliminate the need for manual linking.

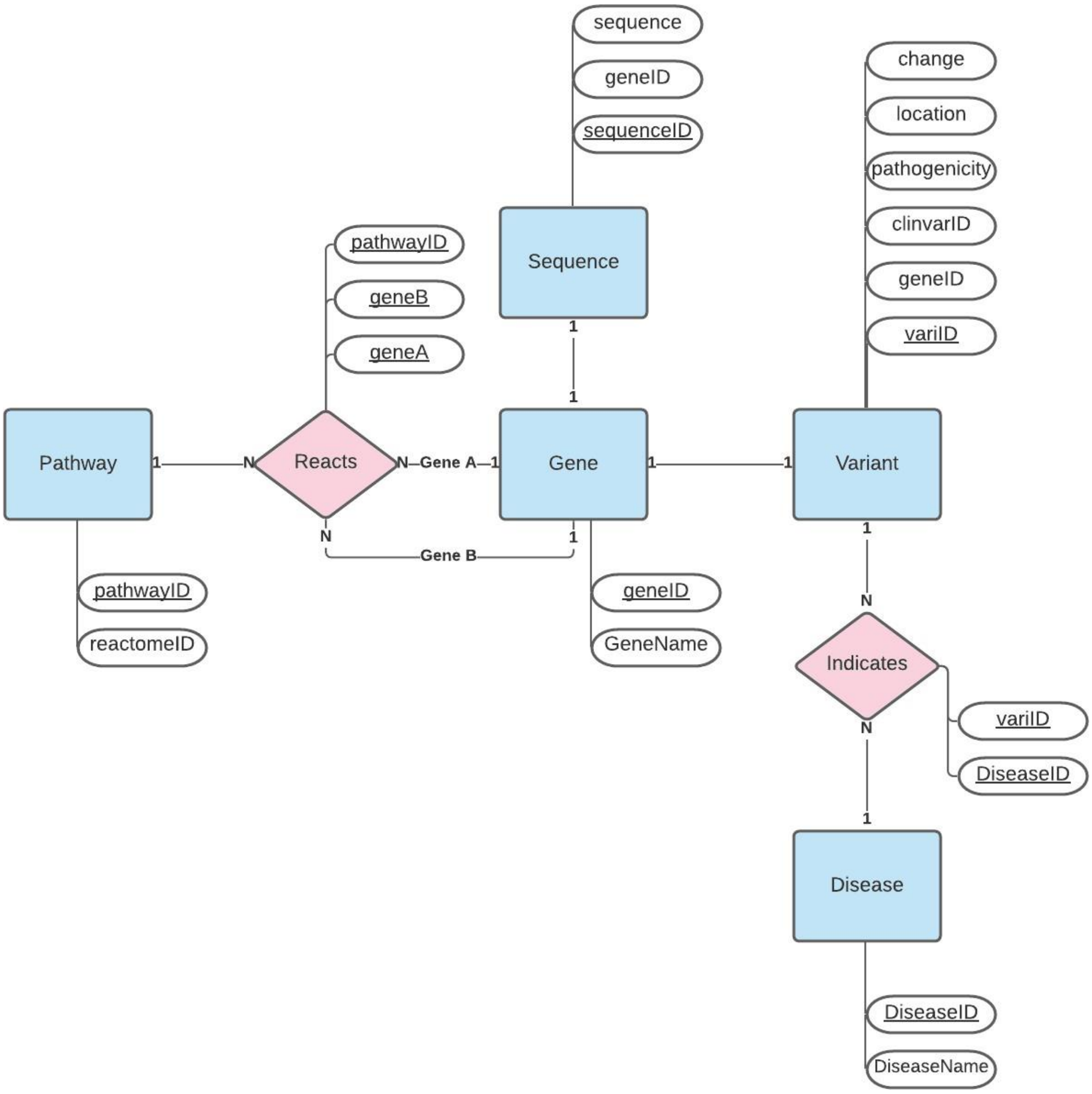

Figure 7 – The entity relationship diagram for SQLite3. Entities can be seen within the blue rectangles and are: Gene, Variant, Disease, Pathway, and Sequence. Pathway and Gene are linked via the associative entity (seen in the pink diamonds) Reacts. Variant and Disease are also linked via the associative entity Indicates. An entity functions as a table which contains a number of attributes (seen in the ovals) which is the information being saved into each table. The unique primary key for each entity is underlined. Relationships between the entities are denoted as either one to one (1 – 1) or many to many (N – N). Many to many relationships are split via associative entities.

2.3 Automating the Data Collection

2.3.1 Data Collection Pipeline

To assist in automating the process of data retrieval and reduce human error when inputting data into a database, a web scraping pipeline was developed by integrating a number of smaller scripts used to parse data from individual websites. Data was largely collected from two main sources: ClinVar and Reactome, as well as UniProt for a small amount of data. ClinVar was selected for it’s widely applicable use in human diseases, with most entries containing information on the pathological effects (or, are given unknown clinical significanceto highlight cases where this information isn’t yet known). Reactome was used due to the wide scope of pathways included in the database; while alternative pathway data sources exist containing more detailed metabolic pathway information (eg. KEGG), Reactome was selected as pathways outside of metabolic reactions, for example, cellular and signalling pathways, are also included. Seeing how BRCA1 is identified to be key during the cell cycleand mitosis/meiosis, it was crucial that a wide variety of pathways could be considered.

A visual flow chart of the data collection pipeline can been seen below in figure 8, highlighting the scripts, sources, and data collected and used during the process, as well as how this relates to the information inputted into the database. The majority of scripts were written using Python, with a few smaller simpler scripts written in BASH (such as the "search_uniprot" script). The overall pipeline was written to string together the individual data mining scripts and reduce the time taken to complete the process; by automatically continuing onto the next step without human input.

This pipeline takes the input as a gene name, and inputs this gene name into a series of scripts that returns and enters the data into the database in the required format. There are essentially two sides to the pipeline: one side searches ClinVar for variants, while the other uses UniProt IDs to search Reactome for pathway information. The ClinVar search can be seen down the left side of figure 8 and the UniProt/Reactome search can be seen down the right side. In the pipeline, these process continuously run one after each other with the left occurring before the right, but for the visual representation they are treated as separate processes that run parallel.

For the left hand process, the gene name is taken and inputted into Entrez to construct a complex search query. Entrez comprises of 39 molecular and literature databases, and is constantly growing as medical science develops and new data types are created (NCBI Bookshelf, 2016). All search boxes on NCBI websites connect to Entrez to generate a search query using Boolean operators, indexed fields, query translation, and automatic term mapping. This query building function is able to be utilised to create a query ID with the given input gene and pre-defined search conditions. The pre-defined search conditions for this study are: single nucleotide variants, single gene, missense, and nonsense mutations, pathogenic, likely pathogenic, benign, and likely benign variants. These search parameters were chosen to restrict the search to only SNP variants with the most frequent types of mutation. It was also chosen to restrict the search to only single genes to avoid confusion around interacting genes. The Entrez query ID is then taken, within the same script, and used to connect to ClinVar and generate list of search results found within the given search parameters. This list is then inputted into the final script before the data can be inserted into the database. This last script processing the search results list from ClinVar, and inserts the input gene into the database if it is not already found within the database. The discovered variants for the gene are also inserted, as well as their clinical significance and the disease associated with the gene. The disease and gene are linked via a separate "indicates" table, which shows which gene indicates which disease(s).

For the process shown on the right of figure 8, the gene name is taken and is used for a search on Uniprot. This, and the following three steps, are contained within a smaller pipeline known as "pipeline_uniprot_reactome", which takes the gene name and returns reactome information in XML files through Uniprot IDs. As it is more effect to search via UniProt codes for Reactome pathways to reduce irrelevant search results, the first step of this smaller pipeline is to convert the gene name into a UniProt ID code. This UniProt ID is then resubmitted back to the UniProt website, and all Reactome pathway IDs associated with the the UniProt ID are returned. Each associated Reactome ID is then submitted to the Reactome website, and the XML file for that pathway is downloaded and saved onto the server. The data within the XML file is stored in such a way that additional data parsing was required in order to extract the desired information. A final script is used within the smaller pipeline in order to achieve this. This parsing script searches the XML files for directly interacting proteins with the query gene, and outputs a log of these proteins. Direct protein-protein interactions were selected in order to minimise the ambiguity of the involvement of the query gene within a reaction, since this makes structural analysis of mutations on interaction sites more apparent – the exact location of interaction is known. Finally, a list of all pathway protein neighbour associations is outputted from the smaller pipeline, and is used as the input for the final step of the overall pipeline: processing the Reactome results in order to insert them into the database. Here, the script searches for all new genes not already present in the database, and inserts these new genes in as entries. Additionally, the pathways are also inserted into the database, and are connected to the genes via a "reacts" table.

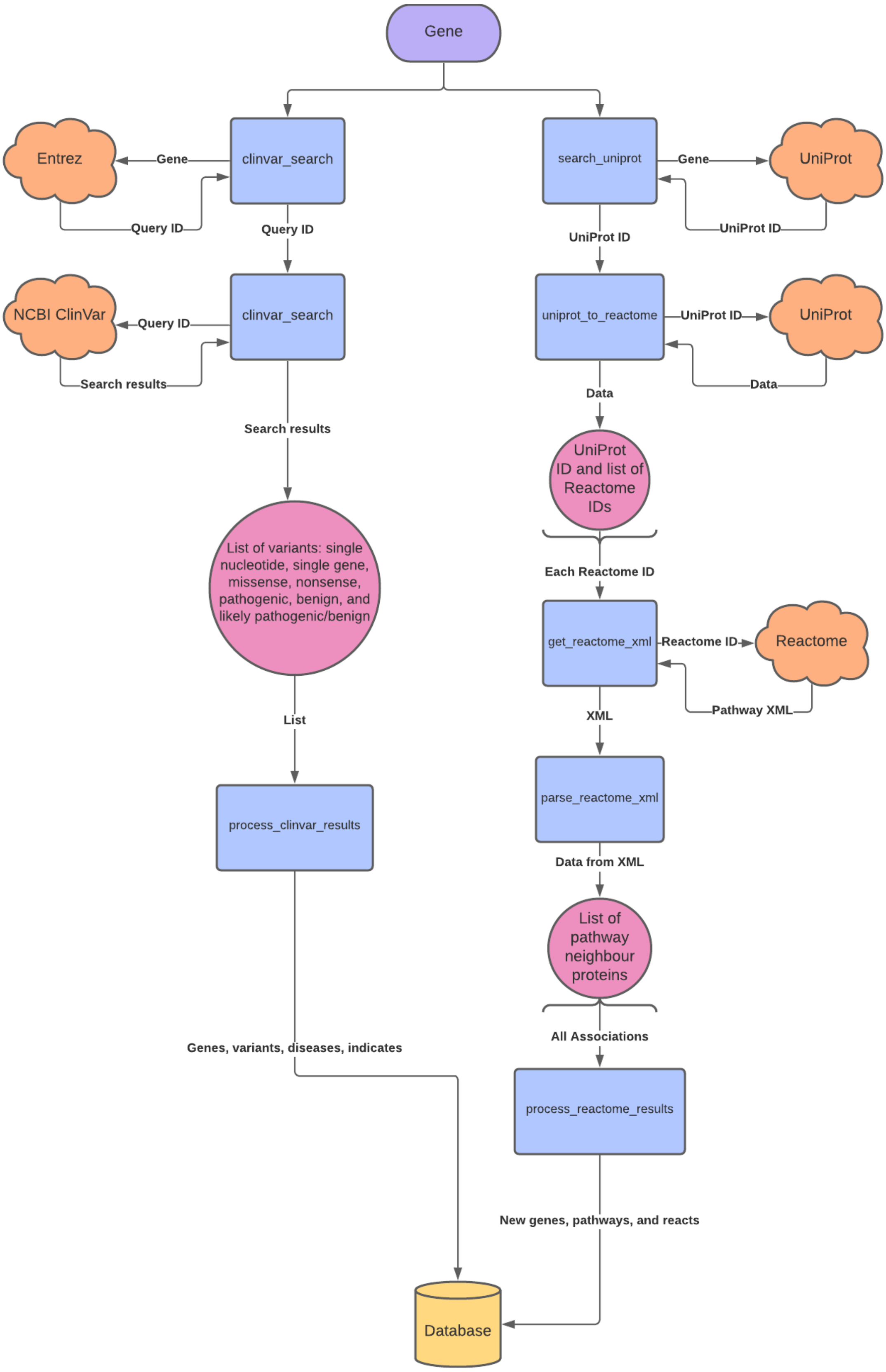

Figure 8– Flowchart of the data pipeline used to collect and store information within the database. Each shape represents a type of information or resource: the purple oval is the input, the blue rectangles are the induvial scripts within the pipeline, the orange clouds are the internet sources, the pink circle is data types, and the yellow cylinder is the database. The input is

simply a gene name, which is inserted into the pipeline which runs two processes to extract pathway, disease, and variant information from online data sources. First, the gene is inputted into a Entrez query link outlining the ClinVar filters for the search: single nucleotide mutation, single gene, missense, and nonsense mutations during the ClinVar search script. This Entrez link is used to connect to the ClinVar database, returning the search results for the filtered query as a list. This list is then run through another command to process the results, which scans the list for new genes, variants, diseases, and the indication of gene to a disease, and inputs these into the specified database. The pipeline then enters the gene name into a search on UniProt, and returns a UniProt ID; which is then used to gather information on the pathways associated with the UniProt ID on Reactome. These UniProt IDs and Reactome IDs are inputted into a script to download the Reactome pathway XML files. These files need to be parsed before they are readable, and this is achieved in the process Reactome results script. The same script inputs the data it has extracted from the XML file into the specified database.

2.3.2 Data Pipeline Iteration

The data collection pipeline was iterated to increase the number of genes within the database and give a wider view of the impact of SNPs. Genes that were identified as neighbouring pathway members are included within the iteration. Figure 9 below summarises the process of the iteration of the pipeline. The original gene, in this study BRCA1, is the first entry into the pipeline and is therefore iteration zero. Once the complete data pipeline has been run, the database is searched for new genes that were added from the pathways identified to contain BRCA1. These new genes are then submitted back into the pipeline individually, and the process is repeated. The process is currently limited to three iterations to prevent collection of data that isn’t detailed or relevant to the original target gene.

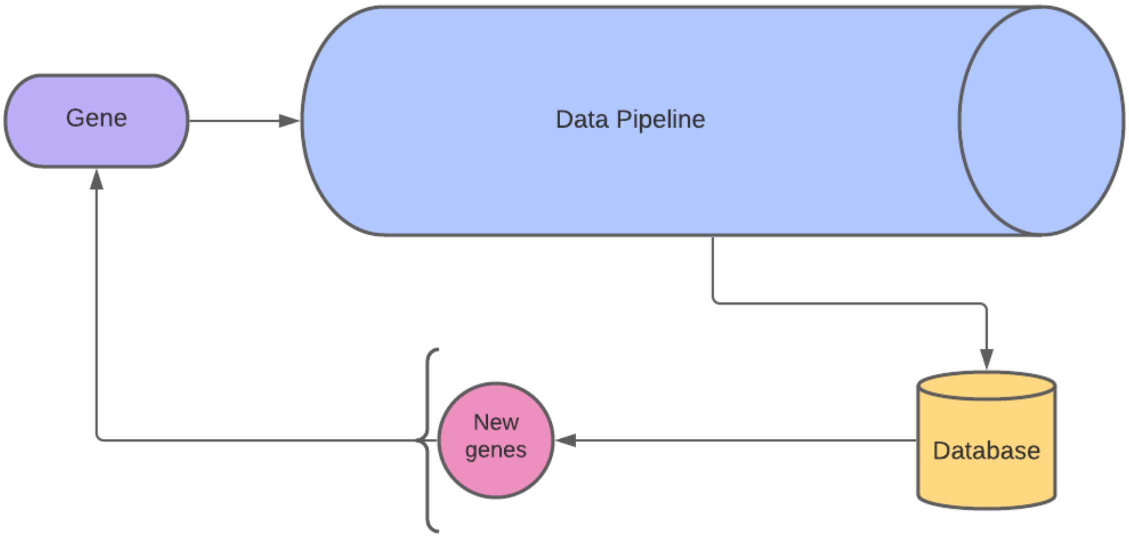

Figure 9 – Flowchart of the simplified process of the data pipeline iteration. Each shape represents a type of information or resource: the purple oval is the input, the blue cylinder is the data pipeline detailed above, the pink circle is data types, and the yellow cylinder is the database. The original query gene is inputted into the pipeline, and the resulting additions to the database are scanned for any new genes. If any new genes are identified, then these genes are then individually inputted back into the pipeline, and are similarly processed. The process is capped at three iterations currently.

2.4 BRCA1 Structural Modelling

The BRCA1 query sequence was input into an in-house modelling pipeline (written by Dr Karl Austin-Muttitt, Swansea University Medical School) in order to generate a homology model.The schematics of the pipeline is shown in figure 10. To begin, the query sequence is inputted into BLAST to identify a range of homologous proteins using Blocks Substitution Matrix (BLOSUM) scoring. Proteins with a low similarity to the query are discarded (often below 30%, otherwise

known as the ‘twilight zone proteins’), and proteins with similarity above 30% are kept. These are then run through MSA steps using both MAFFT (Rozewicki, Li, Amada, Standley, & Katoh, 2019) and 3D-COFFEE (Taly et a., 2011) to provide potential structural information on gaps in the protein alignment by using sequence conservation. 3D- COFFEE utilises given sequences and structure to perform a MSA – so for where there are 19 sequences and 1 query, 3D-COFFEE is given these sequences and the corresponding 19structures (O'Sullivan, Suhre, Abergel, Higgins, & Notredame, 2004). Due to the nature of thequery, it is important to use 3D-COFFEE instead of T-COFFEE, where the latter will search the PDB for a structure to assign the query also. 3D-COFFEE searches for the 3D overlap of each inputted structure and applies this to the query. MAFFT is additionally used to set the global alignment constraint. Homology modelling is then performed using MODELLER.
There are two unique spaces that MODELLER assesses: the sequence space and the structural space (Eswar, Webb, Marti-Renom, Madhusudhan, Eramian, Shen, Pieper, & Sali, 2016). The combined output of these results in a homology based protein model. Within the sequence space, the query sequence is compared to a singular homologous template sequence. Each time the sequence matches in a certain amino acid position, this is translated into the structural space. The protein structure of the homologous template sequence is cross referenced with the sequence space; so that where a match has been identified, the query sequence is assumed to have the same structural identity in this position as the template. This is performed for all homologous sequences. The protein model is then further refined through a number of methods to generate the high quality model. The following were considering in the refinement step: the fundamental stereochemical properties, statistical properties (angles within the protein), and energy minimisation. The output is given as a .pdb file and contains the model information suitable for viewing in Chimera (Pettersen, Goddard, Huang, Couch, Greenblatt, Meng, & Ferrin, 2004).

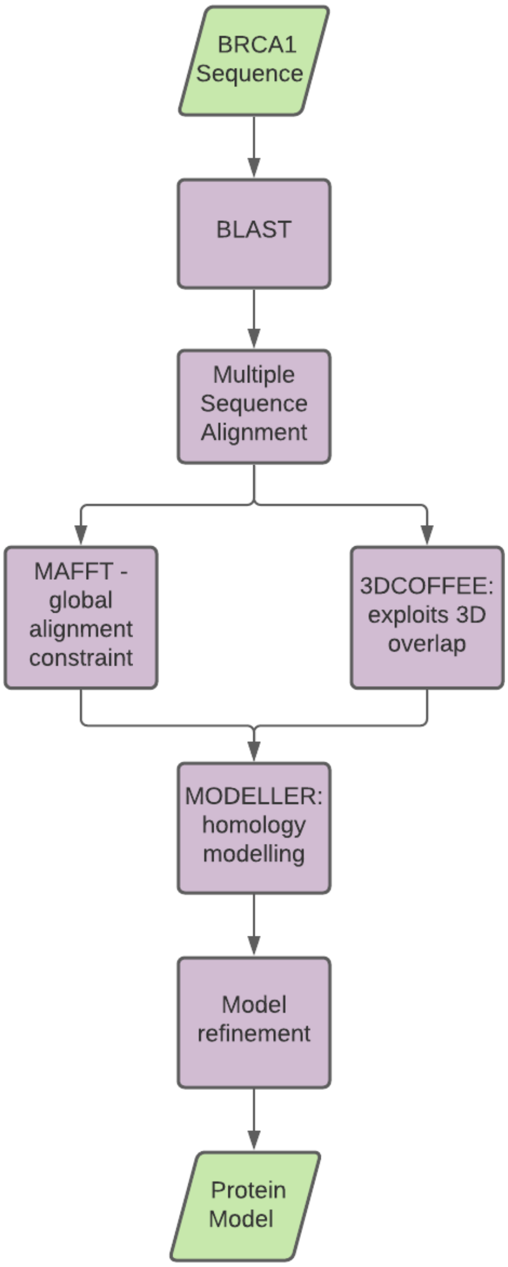
BRCA1 Sequence
BLAST
Multiple Sequence Alignment
MAFFT - global alignment constraint
3DCOFFEE: exploits 3D overlap
MODELLER: homology modelling
Model refinement
Protein Model

Figure 10 – Protein homology pipeline. The BRCA1 sequence is inputted as the query sequence. Homologous sequences are identified using a BLAST search, and any with similarity over 30% are carried forward into a MSA run with both MAFFT and 3DCOFFEE. MAFFT sets the global alignment constraint and 3DCOFFEE searches for 3D overlaps from the inputted sequences and structures to apply to the query sequence (which had no structure provided). MODELLER is used for the homology modelling steps, and considers both sequence space and structure space. Model refinement is run last and considers the following: fundamental stereochemical properties, statistical properties (angles within the protein), and energy minimisation. The output is the query protein model.

Alongside the above method, some online tools were also used to generate comparative models. I-TASSER (Yang & Zhang, 2015) and Phyre2 (Kelley, Mezulis, Yates, Wass, & Sternberg, 2015) were selected for their strong *ab initio* approach to structural modelling. Due to the length of the entire BRCA1 protein, there was a need to split the protein into domains in order to be able to submit the sequence to these tools. A search of BRCA1 on InterPro (EMBL-EBI, 2021) yielded information about 4 possible domains and a linking disordered region. The BRCA1 sequence was split into these 5 regions and submitted to each tool. When using Phyre2, the intensive mode was selected so that multiple templates and *ab initio* methods were used across the entire protein region. These 5 individual regions were then re-joined using a pipeline (written by Dr Karl Austin-Muttitt, Swansea University Medical School) to create a whole BRCA1 model.

2.5 Protein Mutation

Once a final model was constructed using the method above, the model was subjected to mutations identified within each of the 5 regions. An equal number of pathogenic mutations from each region were selected at random to be implemented into the model. For the benign mutations, some regions contained a low number of variants and therefore were all implemented; the number of mutations from regions with a higher number were kept as equal as possible.

In order to mutate the protein, a new set was scripted pipeline was written that utilised quick re-calculation ofmodel coordinates to generate a new model for the mutated protein with the best possible dynamics. This script required the complete wildtype model, and a new .fasta file of the mutated bases. The output of this script was the mutated .pdb file, suitable for viewing in Chimera.

Originally, each mutated protein was viewed in Chimera for visual differences to answer questions such as:

- What is the size difference between the new and wildtype side chain?
- Does the new side chain block any important regions?
- Does the new side chain alter the orientation of the amino acid?
- Is there a change in the hydrogen bonding of this residue?
- Is there a change in the clashes and contacts of this residue?

While effective, when the number of proteins was scaled up, it was evident that this approach would prove to be too time consuming and allow for too many human mistakes (such as errors in counting or setting the hydrogen bonds up incorrectly). This process was also automated by using a connection to Chimera through the server to run the processes that were originally completed by hand. This process took a specified residue location and the .pdb file of the protein in question, and produced the hydrogen bonds, clashes, and contacts for the neighbouring atoms and residues to the specified location. The hydrogen bonds were set to be either ‘strict’ or ‘lenient’, where lenient applies a tolerance of 0.4 A and 20 degrees to the standard length and angle definitions for a hydrogen bond. This generated numerical values for each of these conditions, and allowed for quantitative analysis between the mutated models and the wildtype model. The change between the wildtype and mutant was recorded as an absolute value of the difference between the two readings for each position specified. Absolute values were chosen to prevent negative changes from skewing the data result since the scale of change was the parameter, not the direction of the change. To assess the significance of the change between the pathogenic and benign datasets, statistical tests were applied to the data. Welch’s t-test was used to generate p-values for the difference between the two datasets, and to suggest significance.

3. Results

3.1 Database Findings

3.1.1 Variants in the Database

After completion of the data collection pipeline, the database was searched to identify all listed BRCA1 variants, this can be seen in table 5. This query was undertaken in order to see the differences between the total number of pathogenic and benign variants to assess the impact of SNPs.

Table 5 – the total number of variants in the database after running the pipeline to completion, including a breakdown of the pathogenic and benign mutations.

Total Number of Variants	Pathogenic Variants	Benign Variants
6,448	4,840	1,608

3.1.2 Genes Within the Database

Querying the database following one iteration of the pipeline revealed an additional 7 directly interacting genes were added into the system, excluding the original BRCA1. These genes are listed below in table 6, along with their role in pathways containing BRCA1. Each gene listed has a direct interaction with BRCA1 in the pathway specified, and were identified as potential targets for structural impact in SNPs. The majority of direct interactions are seento occur during the formation of the meiotic single-stranded DNA invasion complex.

Table 6 – the 7 associated genes with BRCA1, found through pathway analysis. The pathway identifier of the shared pathway is shown, as well as the role of the gene in the stated pathway. Data taken and adapted from Reactome, 2021.

Gene Name	Pathway Name	Biological Pathway Role
UBE21	PIAS1,4 SUMOylates BRCA1 with SUMO1, PIAS1,4 SUMOylates BRCA1 with SUMO2,3	PIAS1,4 SUMOylate BRCA1 with SUMO1:C93-UBE21, SUMO2:UBE21, and SUMO3:UBE21.
BARD1	BRCA1 forms a heterodimer with BARD1	BRCA1 and BARD1 form a heterodimer between sequences surrounding the N-terminal RING domains.
DMC1	Formation of meiotic single-stranded DNA invasion complex	A RecA homolog, coats the single-stranded 3’ DNA produced by resection of double stranded breaks.
CDK4	Formation of meiotic single-stranded DNA invasion complex	Regulate the cell-cycle during G1/S transition.
RAD51	Formation of meiotic single-stranded DNA invasion complex	A RecA homolog, coats the single-stranded 3’ DNA produced by resection of double stranded breaks.
BRCA2	Formation of meiotic single-stranded DNA invasion complex	Participates in the loading of DMC1 and RAD51 onto single strand DNA.
ATM	Formation of meiotic single-stranded DNA invasion complex	Localised to double-strand breaks to phosphorylates histone H2AX.

3.1.3 Variant Clusters

The pathogenic variants within the database were clustered by location to elucidate further informationabout where mis-sense mutations due to SNPs were commonly located within the BRCA1 protein. This information can be used further to target SNPs searches to regions of high levels of variants, and identify potentially crucial structural regions of BRCA1 by pathogenic effect. Figure 11 below shows the distribution of unique variants across the entire protein,grouped by 100 amino acid locus.

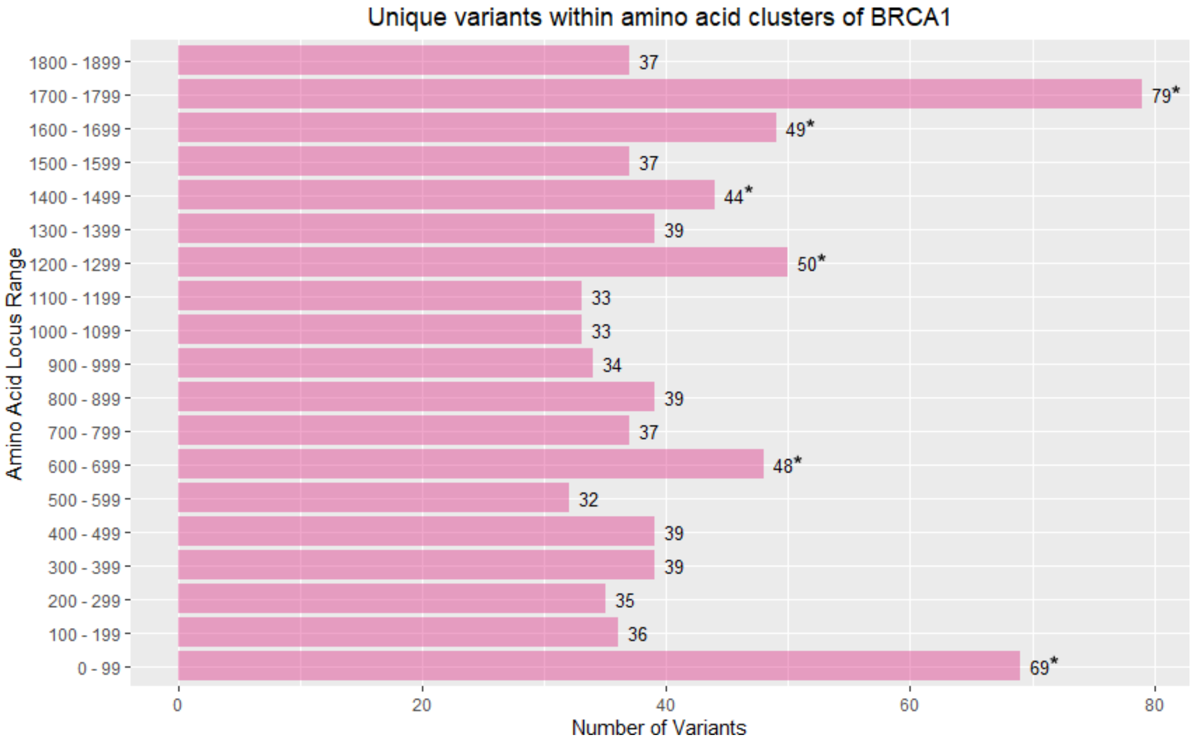

Figure 11 – The distribution of unique variants within amino acid clusters of 100 locus within the BRCA1 protein. The average number of variants across the regions was found to be 43 (to the nearest whole number). 6 regions were identified to be of interest as they contained higher than the average number of variants: 0 – 99, 600 – 699, 1200 – 1299, 1400 – 1499, 1600 – 1699, and 1700 – 1799.

Cluster analysis of the variants revealed 6 regions of interest with higher than the average number of variants, highlighted with an asterisk (*). 4 (1200 – 1299, 1400 – 1499, 1600 – 1699, and 1700 – 1799) of these over average clusters occurred in the latter half of the protein, compared to 2 in the first half of the protein (0 – 99, 600 – 699).These 6 identified regions were further investigated in slices of 10 amino acid loci in order to further discover where variants occurred most frequently across the protein and the results of this analysis can be seen in figure 12.

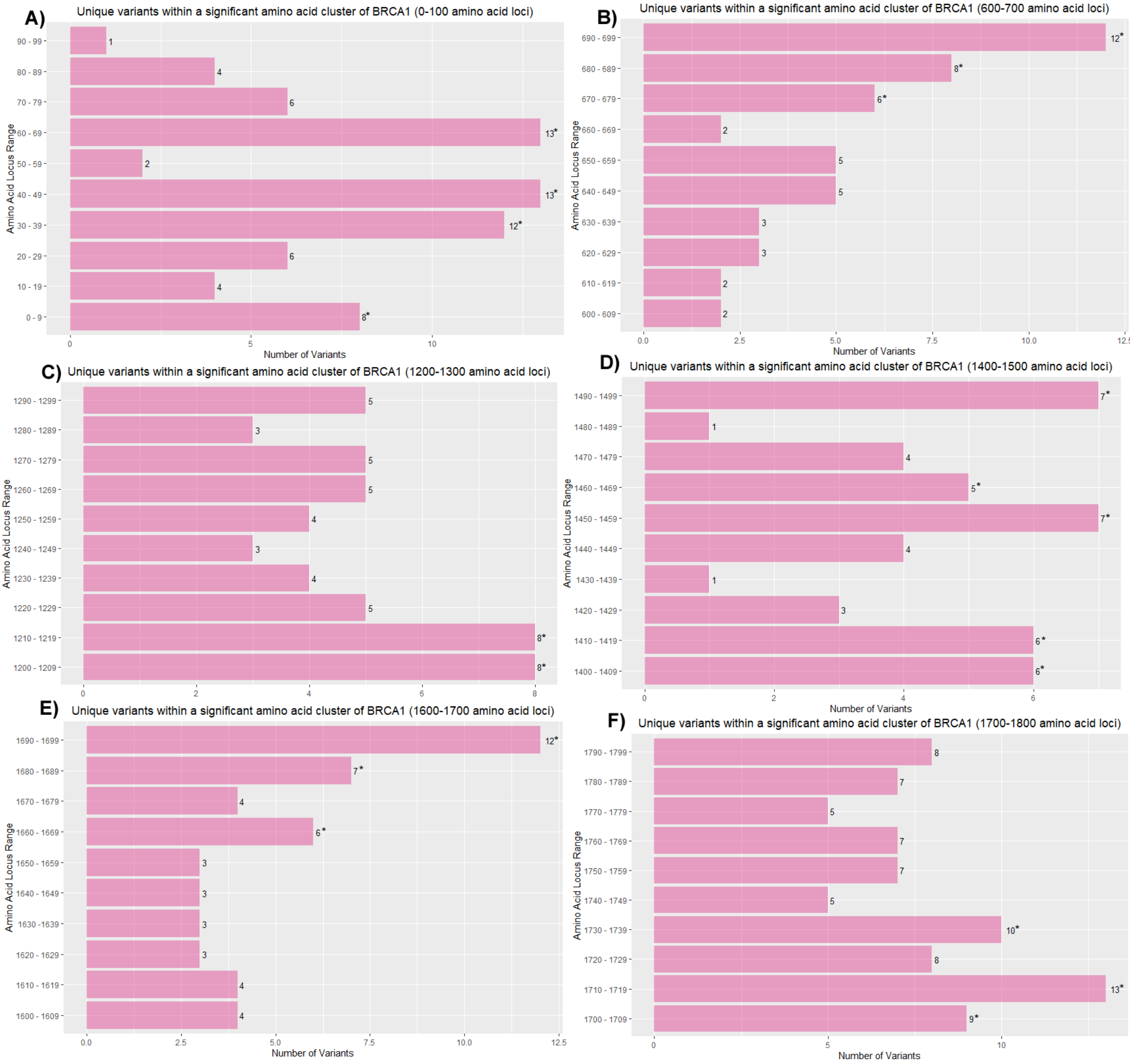

Figure 12 - The distribution of unique variants within amino acid clusters of 10 locus within clusters of interest in the BRCA1 protein. Within each of the loci, clusters that were above the average were identified. Graph A) 0 – 100 amino acid loci, contains 4 regions above average: 0 – 9, 30 – 39, 40 – 49, 60 – 69. Graph B) 600 – 700 amino acid loci, contains 3 regions above average: 670 – 679, 680 – 689, 690 – 699. Graph C) 1200 – 1300 amino acid loci, contains 2 regions above average: 1200 – 1209, 1210 – 1219. Graph D) 1400 – 1500 amino acid loci, contains 5 regions above average: 1400 – 1409, 1410 – 1419, 1450 – 1459, 1460 – 1469, 1490 – 1499. Graph E) 1600 – 1700 amino acid loci, contains 3 regions above average: 1660 – 1669, 1680 – 1689, 1690 – 1699. Graph F) 1700 – 1800 amino acid loci, contains 3 regions above average: 1700 – 1709, 1710 -1719, 1730 – 1739.

Analysis of the smaller clusters of interest revealed different levels of regions above the area for each cluster. Graph A in figure 12 contains 4 regions, Graph B contains 3 regions, graph C contains 2 regions, Graph D contains 5 regions, Graph E contains 3 regions, and Graph F contains 3 regions of above average numbers of unique variants. It can be seen within some clusters that variants occur more frequently at the start of the cluster (as in Graph C and

Graph F), or more frequently towards the end of the cluster (as in Graph B and Graph E), perhaps indicating functional regions of the protein.

3.2 BRCA1 Protein Modelling

3.2.1 Homology Modelling of BRCA1

Homology modelling as per the method outlined in section 1.2 was applied to the BRCA1 gene sequence. The model had low confidence across the middle regions of the BRCA1 protein, between the known start and end domains. The model contained these regions of lower quality because of low template homology across the wholeprotein; BRCA1 is a large protein, therefore not all regions had homology. Additionally, there has been a focus on identifying crystal structures for the main functional domains. Due to this model being of poor quality, additional tools to the original pipeline, such as iTasser and Phyre 2, were used to generate models.

3.2.2 Domains of BRCA1

Using InterPro, the BRCA1 sequence was broken down into 5 domains for modelling based on the domain predictions due to sequence length restrictions on the modelling tools. This breakdown can be seen below in table 7. Phyre2 and iTasser were both chosen to model these domains, and a comparison of the models produced was conducted. Models for domain 1 and 5 were not generated until the modelling tool was selected for the joined BRCA1 structure due to already having known structure.

Table 7 – the breakdown of the domain predictions of BRCA1 used for modelling, taken from InterPro. If the domain name/function is known, it has been listed.

Domain	Length	Domain name (if known)
1	1 – 107	Zinc Finger
2	108 – 570	Serine-rich domain associated with BRCT
3	571 – 1181	n/a
4	1182 – 1608	n/a
5	1609 – 1863	BRCT Domain

3.2.3 Modelling with Phyre2

Phyre2 predominantly produced globular protein structures of the individual domains, where these structures were not expected based off of the known domains. Models for domain 2 and 3 can be seen in figure 13 and 14. Domain 4, shown in figure 15, was the only one constructed to be structured and ordered, however, from previous searches on InterPro, this domain was found to be the intrinsically disordered region of BRCA1. Domain 2was found to be 62% disordered, with 0 residues modelled at above 90% confidence. Domain 3 was found to be 69% disordered, with 0 residues modelled at above 90% confidence. Finally, domain 4 was found to be 79% disordered, with 0 residues modelled atabove 90% confidence. Therefore, the models generated from Phyre2 were not used in thejoined BRCA1 model.

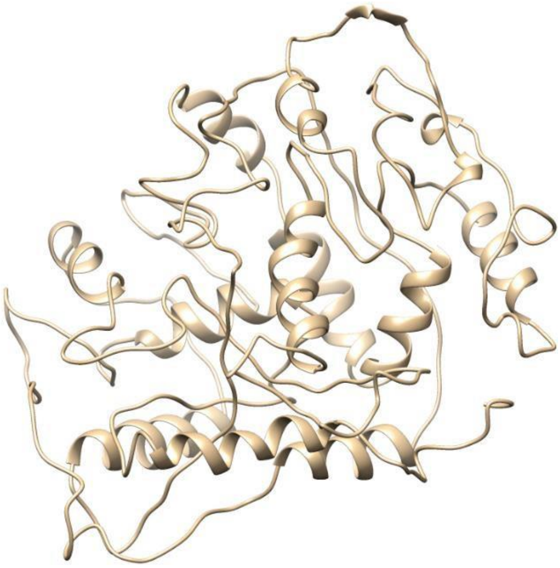

Figure 13 – domain 2 of BRCA1, model generated by Phyre2 and visualised in Chimera. The model produced had an unexpected globular structure.

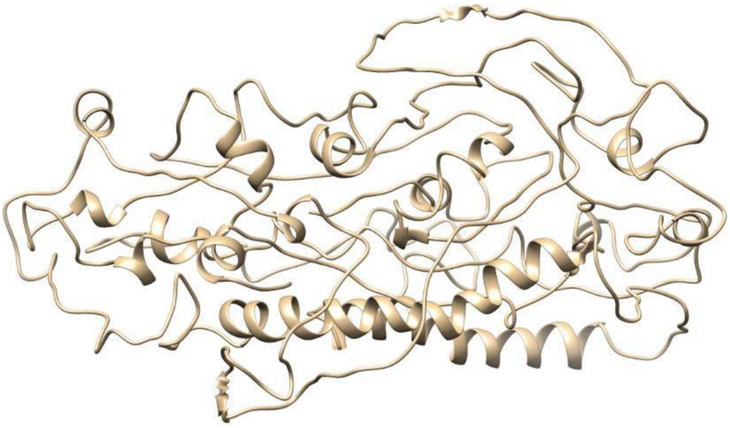

Figure 14 – domain 3 of BRCA1, model generated by Phyre2 and visualised in Chimera. The model produced had an unexpected globular structure.

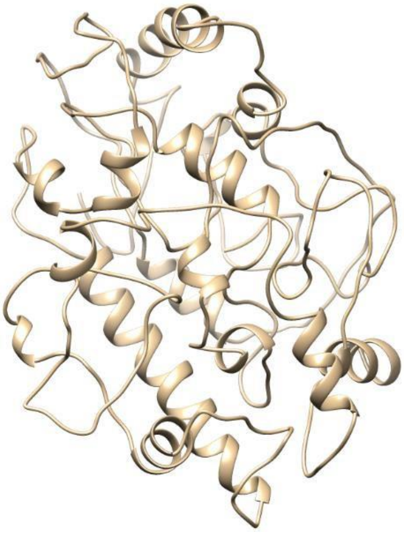

Figure 15 - domain 4 of BRCA1, model generated by Phyre2 and visualised in Chimera. The model produced was ordered, structured, and globular. The expected structure for domain 4 was disordered with little organised structure.

3.2.4 Modelling with iTasser

Models produced by iTasser appeared to have structures that were expected for the BRCA1 protein based off of the known domain structures. In the case of domain 1 and 5, only one model was produced, which matched the NMR structure in the PDB for both domains (PDB ID 1JM7 and PDB ID 3PXE respectively) . For the remaining 3 domains, 5 models were produced for each and were returned in order of confidence. Each model was assessed for the structural likelihood based on confidence score and knowledge of the function of BRCA1. For domain 2, the top model in the list was taken as there was no compelling reason to dispel it and can be seen in figure 16. PDB ID 5JCS was used as a template with a 0.913 alignment score. For domain 3, the top model was discarded due to containing a specific motif for trimerisation that was unexpected within the protein. The third model for domain 3 contained a large helix section that was unexpected. Therefore, the second model was taken as the overall shape appears to containsimilar folds and can be seen in figure 17. PDB ID 5A1U was used as a template with a 0.895 alignment score. Arguably the most difficult domain to model was domain 4 due to theintrinsic disorder. Model 1 contained no structural order, and so could not be discredited.

Model 2 was discarded due the structure appearing as a single long helical structure. The third model produced contained structured regions and was discarded for containing too much order. As such, model 1 was selected for use and can be seen in figure 18. PDB ID 2NBI was used as a template with a 0.854 alignment score. Interesting, this protein is a cell wall protein, further adding to the assumption that this domain is highly disordered.

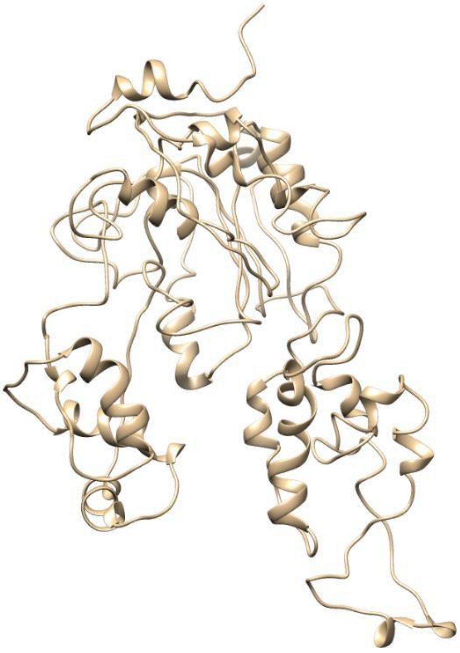

Figure 16 - domain 2 of BRCA1, model generated by iTasser and visualised in Chimera. The model produced shows a potential docking site, and follows an expected structure.

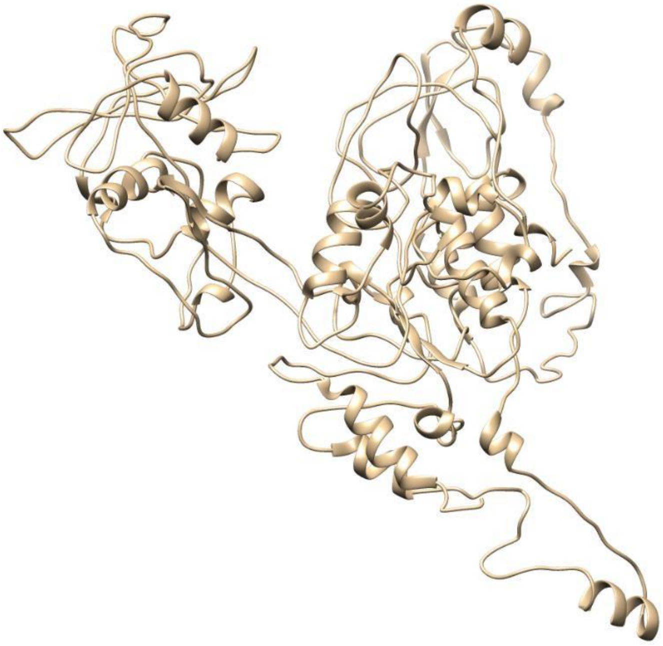

Figure 17 - domain 3 of BRCA1, model generated by iTasser and visualised in Chimera. The model produced shows a potential docking site, and follows an expected structure.

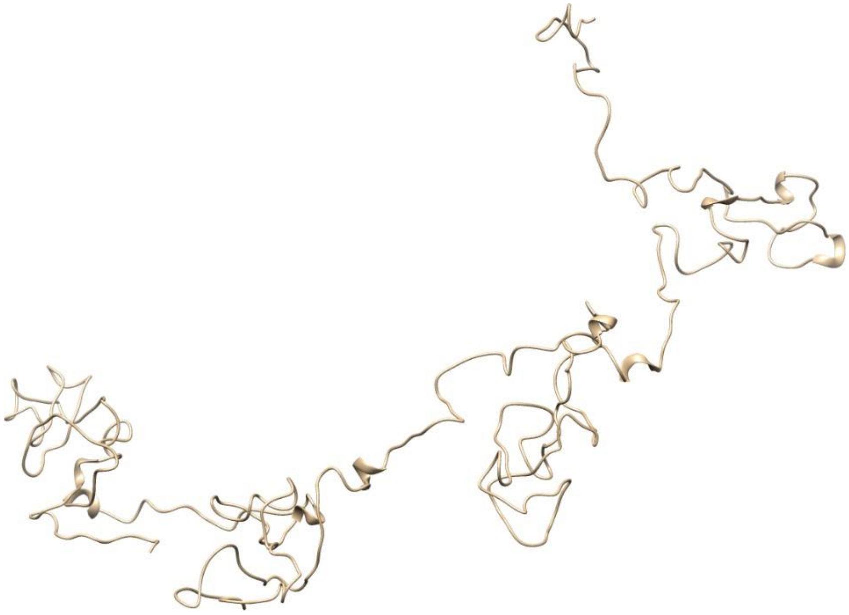

Figure 18 - domain 4 of BRCA1, model generated by iTasser and visualised in Chimera. The model produced is highly disordered and expected for this domain.

3.2.5 Joined BRCA1 Model

The combined model was generated by linking together the 5 individual domain models from iTasser using the method in section 2.4. Individual domains were not analysed due to the potential structural effect across entire protein structure. The combined model can be seen below in figure 19coloured by domain. The domains have been coloured according to this scheme: domain 1 red, domain 2 yellow, domain 3 green, domain 4 cyan, and domain 5 purple The model follows the rough structure of 2 main sections made of domains 1, 2, and 3 and then domain 5, linked by the disordered region domain 4 as expected.

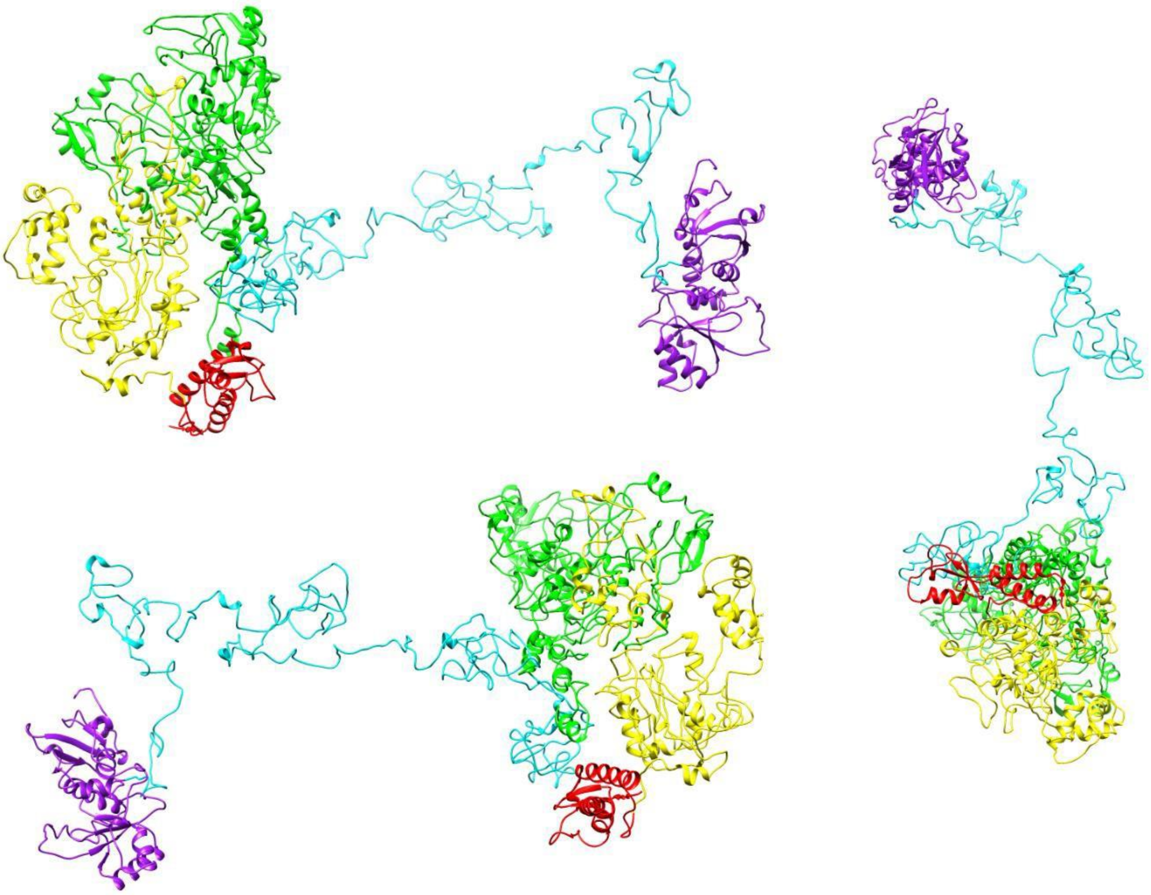

Figure 19 – the complete joined BRCA1 model, generated from individual iTasser models visualised in Chimera. The domains have been coloured according to this scheme: domain 1 red, domain 2 yellow, domain 3 green, domain 4 cyan, and domain 5 purple. The joined model had the expected structure of two distinct regions linked by the disordered domain.

3.3 Structural Analysis

All models generated from the mutation pipeline were analysed for heuristic hydrogen bonds and van der Waals clashes and contacts around a specific residue. The mutated BRCA1 proteins and the wildtype BRCA1 proteins were compared for differences in the values between these measurements. Any differences could indicate a change in structure – potentially effecting binding regions or overall stability of the protein. Whole protein structure was considered to maintain rotamer constraint, opposed to analysis by individual domains.

3.3.1 Pathogenic SNP Dataset

The pathogenic dataset consisted of 32 SNP mutations. Each was inputted into the mutation pipeline, generating a mutated protein model which was used to calculate values for the number of hydrogen bonds and van der Waals forces around the mutated residue. This data was used to assess structural impact of SNPs, with the idea that pathogenic SNPs would affect structure more than benign SNPs. The results of these are shown below in figure 20. The data is presented in the raw collection form, where no adjustments to measurements have been taken. The highest and lower pointis marked by the error bars and excludes any outliers to the data. The lower quartile and theupper quartile are represented in the box plots, and the median of the data is shown within the box. Notably, there is a large variation in number of stabilising van der Waals contacting residues and atoms. In some variants the number of disruptive van der Waals clashes are high, while the number of hydrogen bonds remains low across all variants.

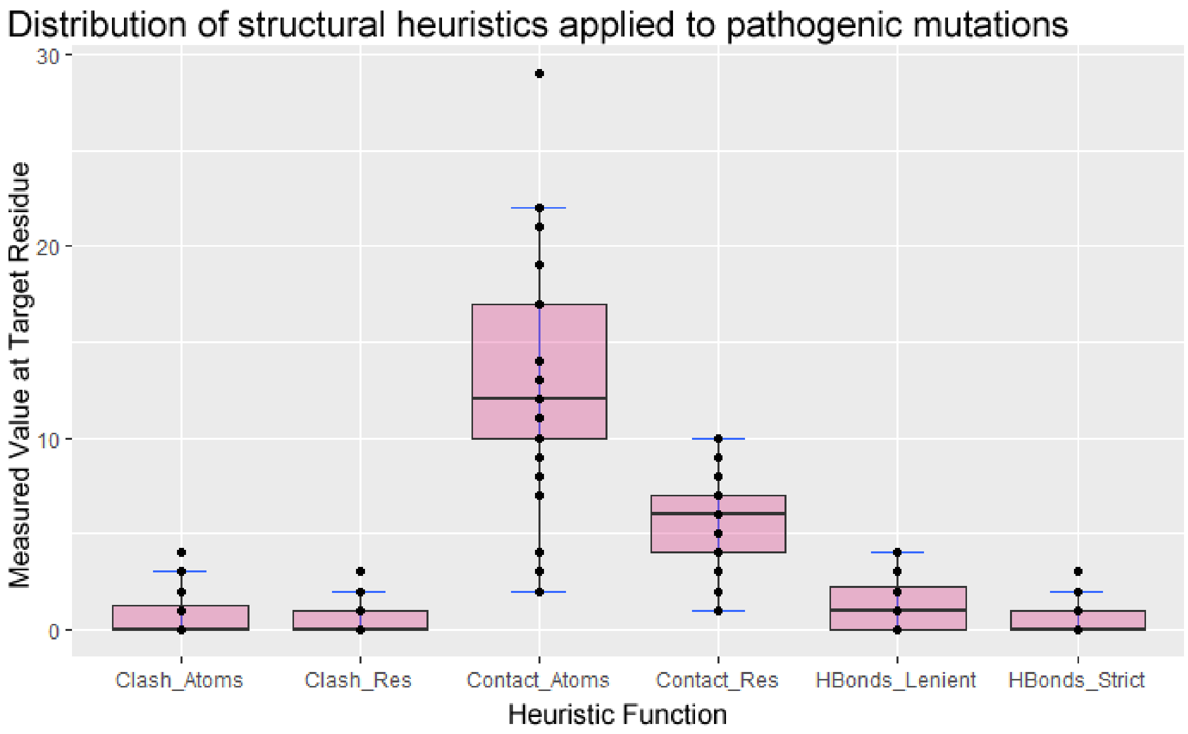

Figure 20 – The distribution of structural heuristic functions applied to the mutated residue in known pathogenic SNPs. The largest spread in data occurs in the stabilising van der Waals forces, while there exists a small number of disruptive van der Waals clashes.

3.3.2 Benign SNP Dataset

The benign dataset also consisted of 32 SNP mutations. Each was inputted into the mutation pipeline, generating a mutated protein model which was used to calculated values for the number of hydrogen bonds and van der Waals forces around the mutated residue to assess structural impact. The results of these are shown below in figure 21. The data is presented in the raw collection form, where no adjustments to measurements have been taken. The highest and lowest point is marked by the error bars and excludes any outliers to the data. The lower quartile and the upper quartile are represented in the box plots, and the median of the data is shown within the box. There is a large spread in the number of stabilising van der Waals contacts, however the number of disruptive van der Waals clashes are low. The number of lenient hydrogen bonds varies within the benign mutations.

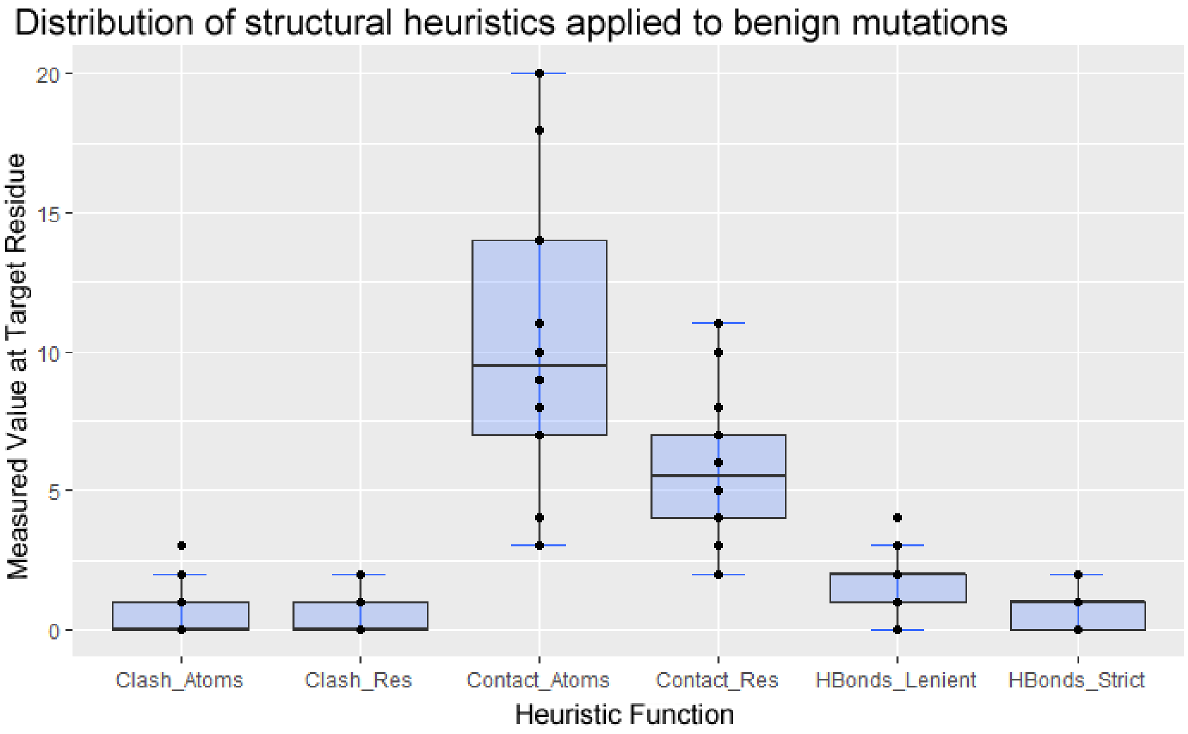

Figure 31 – The distribution of structural heuristic functions applied to the mutated residue in known benign SNPs. The largest spread in data occurs in the stabilising van der Waals forces, while there are only a small number of disruptive van der Waals clashes.

3.3.3 Pathogenic SNP Dataset versus Wildtype Residue Dataset

The wildtype dataset was collected in order to visualise and calculate the differences in these bonds and forces. The wildtype dataset was collected by specifying the same locations identified as SNPs, but using the wildtype model as the reference instead of a mutated protein. Here, the raw data collected from the wildtype sample and the pathogenic sample are compared to reveal the changes in heuristics and can be seen in figure 22. Throughout the measurements taken, there are clear differences between the number of the pathogenic and wildtype proteins. The comparison between the pathogenic mutations and the wildtype reveal that on average, there is a smaller number of stabilising van der Waals contacts, as well as fewer hydrogen bonds in the pathogenic mutants than the wildtype.

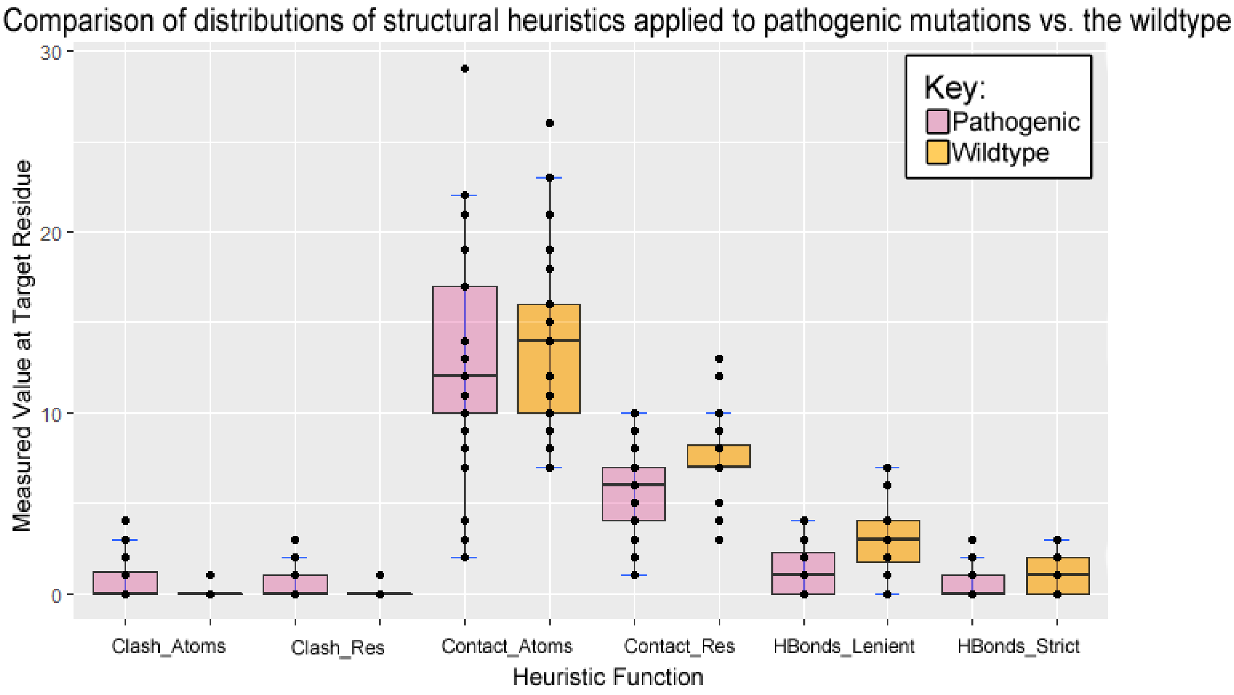

Figure 22 – A comparison showing the different distribution of structural heuristic functions applied to the mutated residue in known pathogenic SNPs (pink) and the corresponding unmutated residue in the wildtype (orange). The comparison shows that the pathogenic mutated residues have on average more van der Waals clashes, fewer stabilising van der Waals contacts, and fewer H-bonds than they would in the wildtype protein.

Welch's T-test was used to identify which heuristic functions showed significant difference (p<0.05) between the mutant pathogenic dataset and the wildtype protein dataset and the results can be seen in table 8. Significant values are indicated in the table by an asterisk (*). The P-value for nearly all functions was less than 0.05, showed significant difference. Interestingly, the only function identified as not significantly different was the van der Waals connecting atoms. Van der Waals clashes and contacting residues, as well as both types of hydrogen bonding, were identified as significantly different in the pathogenic proteins compared with the wildtype proteins.

Table 8 – Welch's T-test of the measured value at target residue of structural heuristic functions of known pathogenic SNPs and the corresponding unmutated residue in the wildtype. Van der Waals clashes, contacting residues, and both types of hydrogen bonds were identified as being statistically significantly different (p>0.05) between the wildtype and pathogenic proteins, and are marked with an asterisk (*).

Heuristic Function	p-Value
Clashing Atoms	0.001004 *
Clashing Contacts	0.001007 *
Contacting Atoms	0.692
Contacting Residues	0.003501 *
Lenient Hydrogen Bonds	0.004478 *
Strict Hydrogen Bonds	0.00116 *

3.3.4 Benign SNP Dataset versus Wildtype Residue Dataset

The wildtype dataset was collected in order to visualise and calculate the differences in these bonds and forces. The wildtype dataset was collected by specifying the same locations identified as SNPs, but using the wildtype model as the reference instead of a mutated protein. Here, the raw data collected from the wildtype sample and the benign sample are compared and can be seen in figure 23. This comparison shows that on average, the difference in heuristic functions between the benign mutations and the wildtype protein is minimal. There is a slight decrease in the stabilising van der Waals contacts between residues, but also slightly more disruptive van der Waals clashes between both atoms and residues in the benign mutations compared to the wildtype residues.

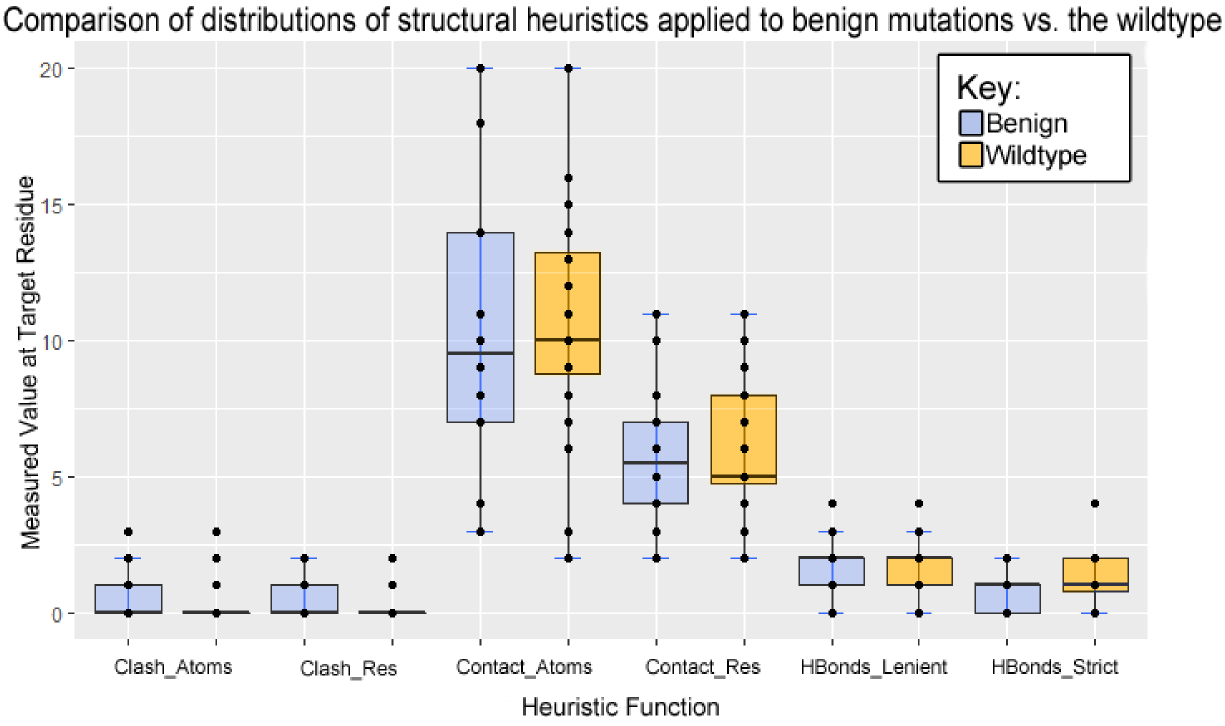

Figure 23 – A comparison showing the different distribution of structural heuristic functions applied to the mutated residue in known benign SNPs (blue) and the corresponding unmutated residue in the wildtype (orange). The comparison shows that the benign mutated residues have slightly fewer stabilising van der Waals contacts between residues, and more disruptive van der Waals clashes between both atoms and residues than they would in the wildtype protein.

Welch's T-test was used to identify which heuristic functions showed significant difference (p<0.05) between the mutant benign dataset and the wildtype protein dataset and the results can be seen in table 9. Significant values are indicated in the table by an asterisk (*). Two functions were identified as being statistically different between the wildtype and benign mutants with a p-value less than 0.05: the van der Waals clashing residues and the lenient hydrogen bonds. While these are statistically significant at this p-value, the differences are markedly less significantly different than for the pathogenic mutations for the same functions. The remaining functions are not statistically different between the wildtype residues and benign mutants.

Table 9 - Welch's T-test of the measured value at target residue of structural heuristic functions of known benign SNPs and the corresponding unmutated residue in the wildtype. Van der Waals clashing residues and lenient hydrogen bonds were shown to have a statistical significant difference (p>0.05) between the wildtype and benign proteins, and are marked with an asterisk (*).

Heuristic Function	p-Value
Clashing Atoms	0.06574
Clashing Residues	0.04488 *
Contacting Atoms	0.5722
Contacting Residues	0.5556
Lenient Hydrogen Bonds	0.02062 *
Strict Hydrogen Bonds	0.2016

3.3.5 Pathogenic SNP Dataset Absolute Deviation

The absolute deviation of the pathogenic data from the wildtype data was calculated by finding the difference in the values between the two datasets for each specified residue and can be seen in figure 24. These changes can be used to infer which changes potentially have a large impact on structural stability. An absolute value was taken to remove any negative changes between the two datasets. The data shows that there was a large change in the number of stabilising van der Waals contacting atoms, and a smaller but noticeable change in the van der Waals contacting residues. Smaller changes can be observed in the van der Waals clashes and hydrogen bonds, with changes in the hydrogen bonds being on average greaterthan that of the van der Waals clashes.

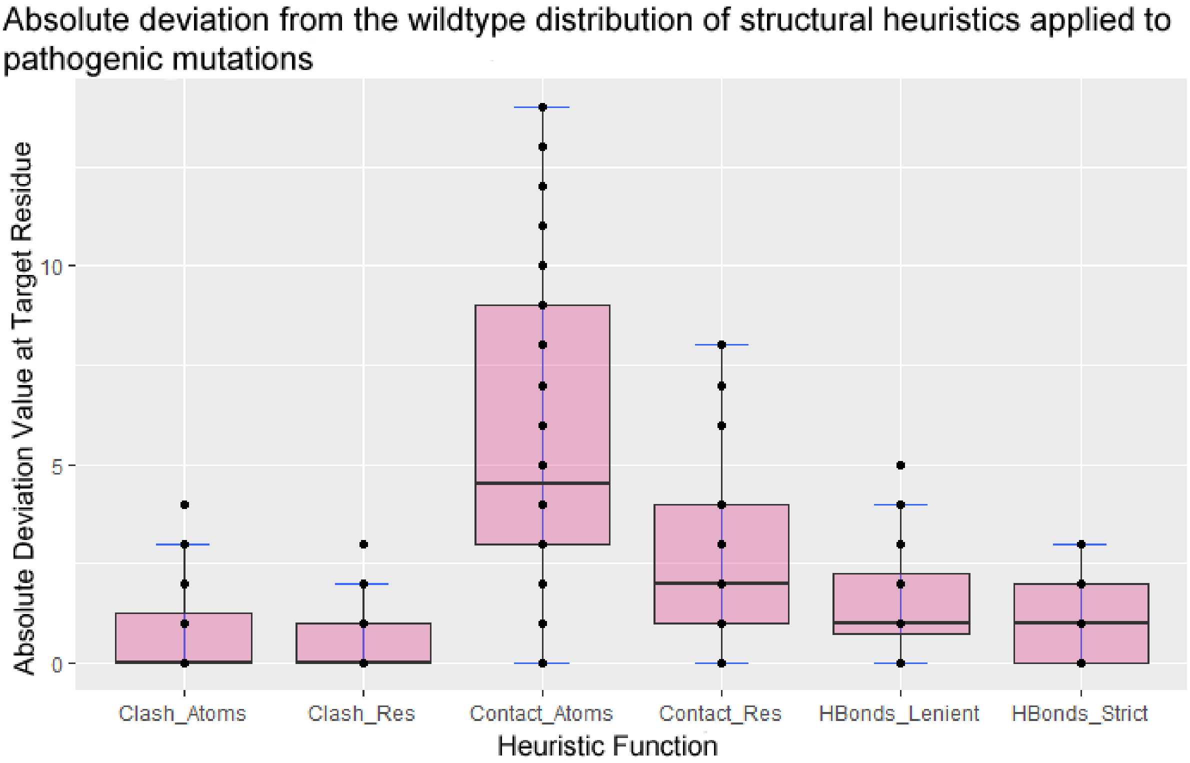

Figure 24 – The distribution of absolute deviation from the wildtype of structural heuristic functions applied to the mutated residue in known pathogenic SNPs. The largest change in data occurs in the stabilising van der Waals forces between both atoms and residues, while smaller changes are seen in the van der Waals cashes and hydrogen bonds.

3.3.6 Benign SNP Dataset Absolute Deviation

The absolute deviation of the pathogenic data from the benign data was calculated by finding the difference in the values between the two datasets for each specified residue and can be seen in figure 25. An absolute value was applied to remove any negative changes between the two datasets. The data showed a significantly small change in most measurements taken, with four measurements having the largest change of less than 5 from the wildtype of the same location. The van der Waals clashes and hydrogen bonds had on average an equally sized change. As with the pathogenic dataset, the largest change can be observed in the van der Waals contacting atoms, but also a smaller change in the van der Waals contacting residues can be seen.

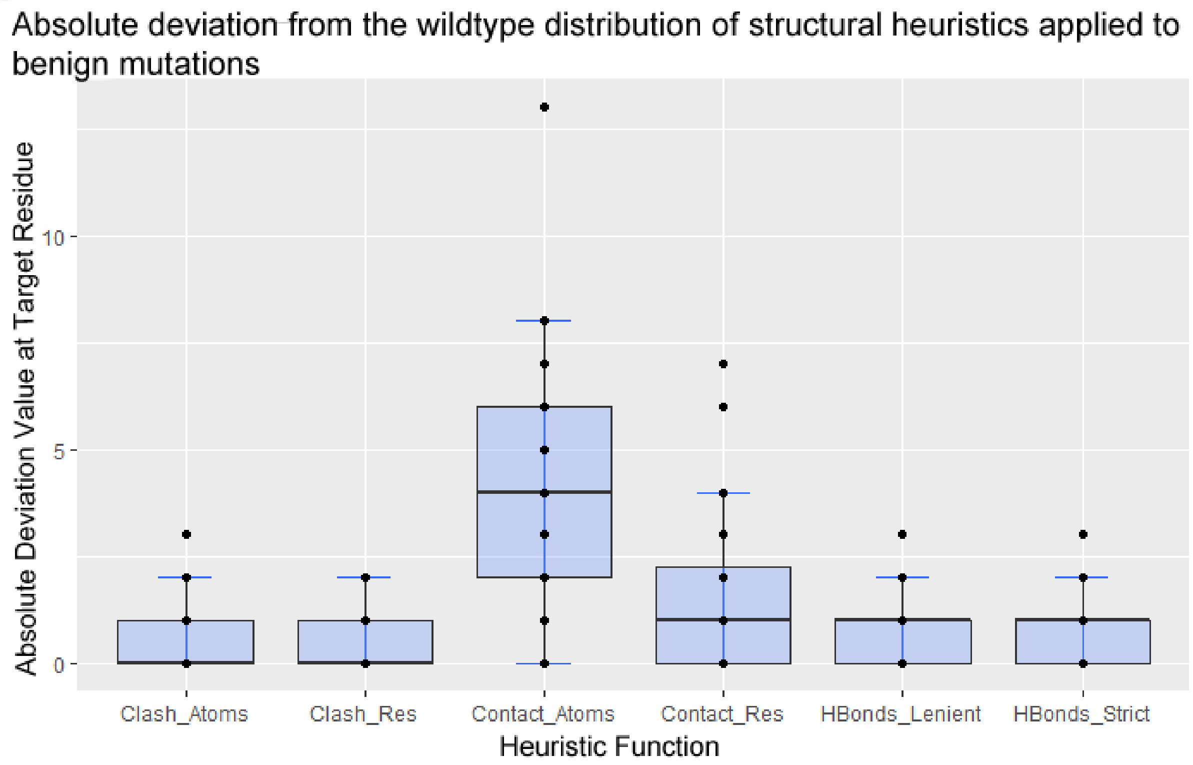

Figure 25 - The distribution of absolute deviation from the wildtype of structural heuristic functions applied to the mutated residue in known benign SNPs. The largest change in data occurs in the stabilising van der Waals forces between both atoms and residues, while equal changes are seen in the van der Waals cashes and hydrogen bonds.

3.3.7 Pathogenic SNP Dataset versus Benign SNP Dataset

The pathogenic and benign datasets were then compared to each other using the absolute deviation of the values from the wildtype and can be seen in figure 26. The comparison shows that there is a larger change in the stabilising van der Waals contacts in the pathogenic dataset compared to the benign dataset. However, the disruptive van der Waals clashes appear to have almost equal changes in both the pathogenic and benign dataset. It can also be seen that the strict hydrogen bonds have a similar level of change between the pathogenic and benign datasets, while there is a slightly larger change in the lenient hydrogen bonds in the pathogenic dataset compared to the benign dataset.

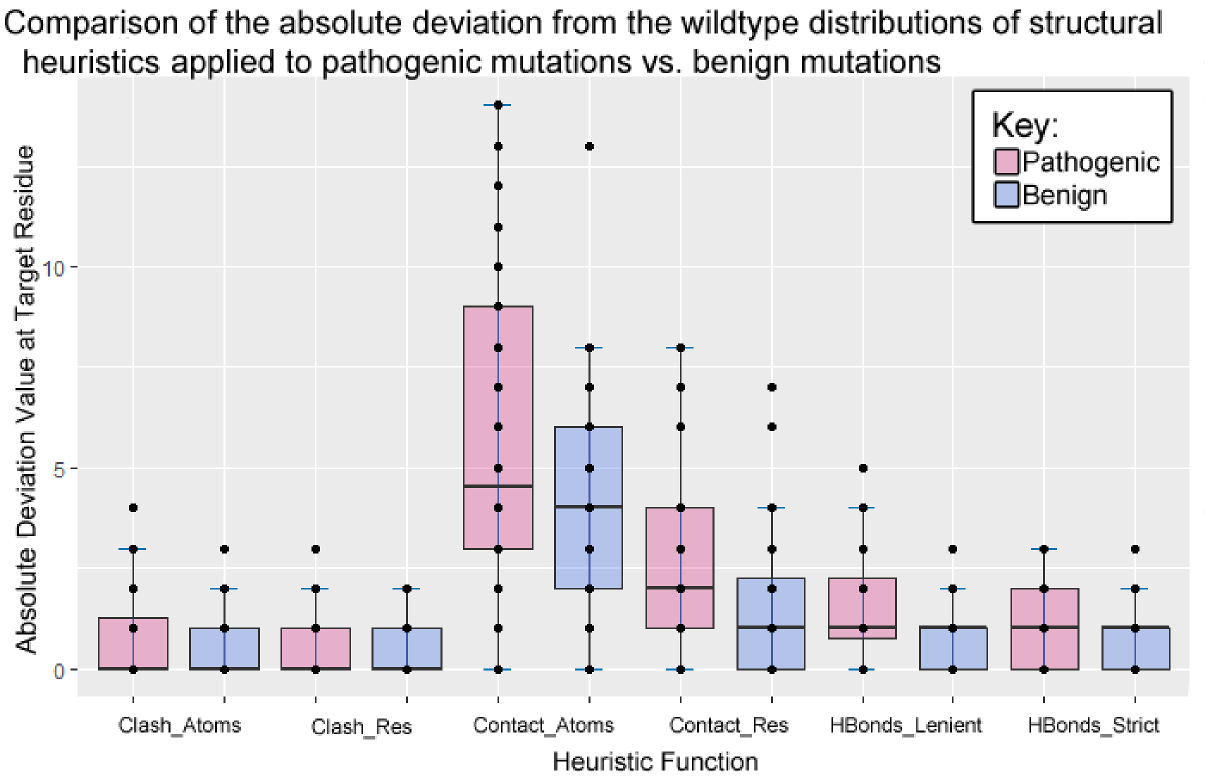

Figure 26 - A comparison of the different distributions of absolute deviation from the wildtype of structural heuristic functions applied to the mutated residue in known pathogenic SNPs (pink) and in known benign SNPs (blue). The comparison shows that there is a larger change in the stabilising van der Waals contacts in the pathogenic dataset compared to the benign dataset, while the disruptive van der Waals clashes appear to have almost equal changes in both the pathogenic and benign dataset.

Welch's T-test was used to identify which heuristic functions showed significant difference (p<0.05) between the mutant pathogenic dataset and the mutant benign absolute deviation dataset and the results can be seen in table 10. Significant values are indicated in the table by an asterisk (*). Three functions were identified as being statistically different between the pathogenic and benign mutants with a p-value less than 0.05: the van der Waals contacting atoms and residues, and the strict hydrogen bonds. The remaining functions are not statistically different between the pathogenic mutants and benign mutants.

Table 10 - Welch's T-test of the absolute deviation value at target residue of structural heuristic functions of known pathogenic SNPs and known benign SNPs. Van der Waals clashing residues and lenient hydrogen bonds were shown to have a statistical significant difference (p>0.05) between the wildtype and benign proteins, and are marked with an asterisk (*).

Heuristic Function	p-Value
Clashing Atoms	0.5275
Clashing Residues	0.7836
Contacting Atoms	0.01014 *
Contacting Residues	0.04702 *
Lenient Hydrogen Bonds	0.1642
Strict Hydrogen Bonds	0.01232 *

4. Discussion

4.1 Summary

A semi-manual run, requiring some human input, of the entire process from gene name to generating variant protein models, was completed with a small set of just over 60 SNPs spread between both pathogenic and benign variants. This process was aided by the use of computational tools such as databases to ensure consistency of data, and pipelines to reduce the need for human input and human error. Without these tools, the analysis done by this project wouldn’t be possibledue to the sheer scale of data and errors that arise when databases are created by hand.

Structural analysis of the proteins generated through homology modelling methods and mutation modelling pipelines shows promising significance in the change between selected heuristic functions – hydrogen bonding and van der Waals forces. The most significant difference between the benign and pathogenic changes lies within the contacting atoms and residues, suggesting that the alteration in these features can potentially be used to determine pathogenicity.

Areas of interesting further study were also identified through this project, such as the consideration of SNP location within the protein (i.e. does this lie within an important functional region?), the proteins that BRCA1 directly interacts with (i.e. is the interaction disrupted by a mutation?), and the thermodynamics of protein stability (i.e. does the increase or decrease of heuristic functions within a protein cause a physical, measurable, change to stability?).

Additional areas of future study and incorporation were identified, such as the inclusion of splice variants, that were not currently present in the pipeline. Splice variants were not the focus point of this study currently, but it would be possible to account for these without modification to the pipeline if there was a known UniProt ascension code for each splice variant. Additionally, information about PTMs and DNA methylation would be interesting to add within the pipeline. However, there is currently a shortage within the pool of known structures with this information. With both DNA methylation and PTMs, such as acetylation, there is a lack of faithful modifications, reducing the ability to generate accurate and reliable structural models.

4.2 BRCA1 Pathway Associations

The direct interactions with BRCA1 found from pathway analysis reveal specific target locations to study for SNP structural changes. Using existing knowledge of the interactions between BRCA1 and the newly added genes, shown in table 6 in section 3.1.2, protein-protein binding and interaction sites on BRCA1 can be discovered and further analysed for changes within these regions. Disruptive mutations within a binding site can potentially destroy protein function by removing access to this interaction. The function in regards to BRCA1 of each new gene is detailed below, indicating potential areas for study of the interactome.

UBE21 is essential for the SUMOylation process, and through a reaction catalysed by the PIAS1,4 complex which acts as an E3-type small ubiquitin-like modifier (SUMO) ligase (Molinaro, Martoriati, Cailliau, 2021). UBE21 is found in three different complexes within the cell: SUMO1:C93-UBE21, SUMO2:UBE21, and SUMO3:UBE21. More SUMO2-BRCA1 and SUMO3-BRCA1 is observed in vivo than SUMO1-BRCA1 (Zhao et al., 2020). SUMOylation is a process that occurs as a response to cellular stress, with SUMOylation of SUMO2 and 3 seen more frequently when oxidative stress occurs and SUMO1 participating in normal cellular responses (Wang, Qian, Yang, & Gu, 2021). SUMOylation of BRCA1 increases its ubiquitin ligase activity, which in turn enhances the ability of BRCA1 to bind and regulate particular transcription factors.

The BRCA1:BARD1 heterodimer is necessary for maintaining the normal repair of dbDNA breaks by HR, the BRCA1-mediated tumour suppressor response, and normal development. BRCA1 contributes to the stability and maintenance of the chromosome through regulation of centrosomes (Yoshino, Fang, Qi, Kobayashi, & Chiba, 2021). BRCA1 and BARD1 both contain RING domains that have E3 ubiquitin ligase activity, which sees a dramatic increase when in the BRCA1:BARD1 heterodimer (Otsuka, Yoshino, Qi, & Chiba, 2020). The heterodimer ubiquitinates centrosomal proteins, such as γ-tubulin, in order to regulate centriole duplication. There are reports of pathogenic BRCA1 mutations that abolish the formation of the BRCA1:BARD1 heterodimer.

DMC1, CDK4, RAD51, BRCA2, and ATM are all involved in the same reaction pathway to generate the meiotic single-stranded DNA invasion complex. There are two RecA homologs found within this pathway: RAD51 and DMC1 (which is meiosis specific), ensuring faithful chromosome segregation (Okorokov et al., 2010). HR repair of dsDNA breaks generates a long overhanging 3' tail which can contain hundreds of nucleotide bases. Strand invasion and homologous pairing through presynaptic RAD51 initiates the strand synthesis of error-free repair (Short et al, 2016). The catalytic activity of RAD51 causes ATP hydrolysis, allowing components of HR repair to dissociate. Both DMC1 and RAD51 function in a similarway, but DMC1 is specifically implicated in meiotic crossing-over (Da Ines et al., 2013).
Knockouts of both DMC1 or RAD51 result in ineffective recombination, therefore both are required for function recombination. BRCA1 and BRCA2 are both responsible for the transport of RAD51 to the site of recombination, while only BRCA2 is responsible for the transport of DMC1 (Jimenez-Sainz, & Jensen, 2021). BRCA2 is required to enable the formation of RAD51 and DMC1 filaments.

While this project didn't address the exploration of analysing SNPs in specific locations, the impact of disruptive mutations in residues that are known to interact with other proteins remains an exciting prospect in classifying the clinical effect of a SNP. Recent studies into predicting functional impacts of variants using interaction network frames revealed that studying the interactome as well as the protein itself will benefit functional prediction (Ozturk, & Carter, 2021). Additionally, a recent study into cancer-related SNPs tested a combined approach of structure analysis and flexible protein-protein interactions with machine learning, which effectively predicted structural effects and changes within interactions (Lie et al., 2021). These recent studies suggest that further investigation into the direct protein interactions with BRCA1 would be beneficial to elucidating the functional impact of SNPs and thus the classification of unknown SNPs.

4.3 Regional Variant Analysis Across BRCA1

Using the table of domains seen in section 3.2.2, it can be seen that areas of high variants can often be observed within important functional regions. Referring to figure 11 in section 3.1.3, it can be seen that between 0 and 99 amino acids, there are a total of 69 variants. These positions lie within the zinc finger of BRCA1, a highly conserved region at the amino terminal of the protein. Zinc finger function is known to differ between protein families such as oncoproteins and regulatory proteins, it is implicitly linked to inhibition of apoptosis in BRCA1 (Johnson, & Kruk, 2002). The cysteine-aspartate specific protease (caspase) pathway is the most common pathway by which cells undergo apoptosis. Disruption of the BRCA1 gene via a 185del mutation within the zinc linker region leads to an increase in caspase-3 dependent apoptotic response. Exploration of variants within the zinc finger domain itself might reveal similar mechanics, thus implicating the caspase pathway as a drug target site.

Referring back to the same figure 11 within 3.1.3, it can be seen that across the BRCT domainof BRCA1, there are 2 higher than average regions of amino acids. Across the region of 1600 – 1699 there were 49 variants, and across the region of 1700 – 1799 there were 79 variants. Figure 12 in section 3.1.3 shows that these variants are located in particular towards the end of the first region, and the start of the second region, where some of the functional activity is located. The BRCT domain at the BRCA1 C-terminus acts as a transcriptional activator of cell-cycle regulated genes, and also as a corepressor of transcriptional factors during the cell cycle G2/M checkpoint through association with zinc-finger protein 350 (ZNF350) (Zeng, & Sang, 2017). The BRCA1 BRCT/ZNF350 complex represses expression of gene commonly involved in the proliferation and vascular formation of tumours, particularlyangiopoitin-1 and high-mobility group AT-hook2. Variants that lie within the BRCT finger can potentially disrupt the association with ZNF350, leading to a dampening of the repression function of the complex. Additionally, the BRCT domain of BRCA1 is commonly known to interact with phospho-ligands and in particular form hydrogen bonds with these phosphor ligands at S1655 residue (Billing et al., 2018). Pathogenic mutations at this S1655 residue leads to a disruption of the interaction between BRCA1 and its phospho-ligands. Further investigations into the particular positions of these variants within the BRCT domain might reveal new information knowledge about the effects of pathogenic mutations within importantfunctional regions, and potentially allow for new drug targets.

4.4 Heuristic Function Analysis

The heuristic functions of van der Waals clashes, contacts, and hydrogen bonds were selected for testing within mutated proteins. These functions are based on physicochemical entities that are vital for maintaining the structural integrity of proteins, and therefore for maintaining protein function. Assessing the change in heuristic functions between pathogenic and benign variants can reveal information about the underlying molecular causation underlying a disruptive mutation.

Van der Waals contacts are crucial for stabilising the tertiary (and sometimes quaternary) structure of a protein, often shaping structurally significant areas such as protein binding regions. Additionally, these forces are responsible for the structural stability of the protein, essentially holding together protein secondary structures and maintaining folded structures. Despite the weak nature of van der Waals forces, the high number of these forces within a protein make them significant in protein folding (Pollard, Earnshaw, Lippincott-Schwartz, & Johnson, 2017). Interrupting these contacts can potentially induce a change of conformation, resulting in the loss of structure or function of a protein. The resulting data collected across 32 variants from both pathogenic and benign datasets shows that the level of change within the pathogenic variant set is significantly different to that of the benign variant set.

Not only is there a significant difference between the pathogenic variant and benign variant datasets in regard to the van der Waals contacts, it can also be seen that there is difference in the strict hydrogen bonds. In order to be classified as a strict hydrogen bond, Chimera needs to consider atom types and geometric criteria (UCSF Computer Graphics Laboratory, 2014). As above, hydrogen bonds are critical for maintaining structure right down to the secondary structure level. Disruption of hydrogen bonds can greatly decrease structure stability and function, as each hydrogen bond has been demonstrated to contribute, on average, around 1 kcal mol^{-1} per bond (Pace et al., 2014) indicating that they significantly contribute to protein stability.

From table 10 in section 3.3.7, it can be seen that there is a statistically significant difference between the absolute change in the stabilising van der Waals contacts between the benign and pathogenic variants datasets. The significant difference between these implies that either a large increase or decrease in these stabilising contacts can alter the function of the protein. Decreasing these stabilising forces will disrupt the stability of the protein, and introduce more flexibility into functionally important regions which reduces the ability to maintain the suitable conformation required for specific function denaturing. Equally as destructive, an increase of these forces can lead to misfolding and contacts between parts of the protein not seen in the wildtype, resulting in insufficient flexibility to allow sufficient conformational freedom required for specific function.

The pathogenic variant dataset from section 3.3.3 can be seen to show statistically significant changes across nearly all functions tested when compared to the corresponding wildtype. Across all 6 functions tested, only one did not show significant change: the van der Waals contacting atoms. Meanwhile, a large change can be observed across the other functions. This indicates that the majority of these functions are vital to maintaining the correct function of the BRCA1 protein, and that alteration in these particular heuristic functions can lead to damaging mutations.

Where significant changes are observed within the benign variant dataset, these changes can be seen to be much less significant than the pathogenic variant dataset, as seen in table 9 in section 3.3.4. For this dataset, the van der Waals disruptive clashes and the lenient hydrogen bonds are found to be significantly different to the wildtype. The decrease in significant changes could further help us discern which functions are crucial for maintaining correct function and structure; if a significant change can be seen within the benign variant dataset where function is broadly maintained, then it can be assumed that this physicochemical parameter contributes less to the overall stability as function hasn't been lost, or that structural maintenance in that particular location is not so important for the maintenance of function.

Disruption of both hydrogen bonds and van der Waals forces can cause loss of function within a protein directly by interrupting protein-protein interactions, removal of enzymatic activity, or though inducing structural instability leading to degradation and misfolding (Birolo et al., 2021). Destabilisation can be seen in many neurodegenerative disorders, such as Parkinson's disease. However, thermodynamics within human proteins is still not fully understood, and so the process for assessing if a variant disrupting protein stability is disease causing or not, is not fully developed (Sanavia et al., 2020). Integration of a process for determining the thermodynamic change in a protein from pre- and post-mutation would greatly improve the insight this tool can provide.

4.5 Potential as a SNP Classification System

Currently, there are many SNP classification systems in use and development, focusing on different areas for classification. Nearly all approaches utilise machine learning methods, training on particular properties within known variant datasets. It is worth noting that there currently exists no SNP classification system that considers structural parameters, making this project a novel area of study and an exciting approach to not only classifying SNPs, but also furthering understanding of their function and effects. A recent literature review into the classification of SNPs in complex brain disease diagnosis using machine learning has shown that there is huge potential for the use of machine learning in biological settings despite certain drawbacks (Ahmed, Alarabi, El-Sappagh, Soliman, & Elmogy, 2021), such as small datasets and the intensity of such computing methods. Therefore, integration of a larger dataset from other sources besides ClinVar would provide the necessary dataset richness Classification of susceptibility to asthma based on SNPs within a person's genome has shown that prediction of genotype-phenotype association can be achieved with high accuracy when integrating various machine learning methods (Gaudillo et al., 2019). A study into classifying SNPs for breast cancer diagnosis demonstrated that a variety of machine learning techniques are needed to achieve high accuracy within classification (Boutorh, & Guessoum, 2015). Reliance on a single technique produced poor accuracy, due to the large number of features but relatively small data sample, otherwise known as the dimensionality problem.

Through further work, a classification algorithm could be developed from the premise of this project which serves as a proof of concept. The structural effects of deleterious SNPs in the human RASSF5 gene has been studied (Hossain, Roy, & Islam, 2020), where it was discovered that SNPs in the binding region crucial to the protein's function reduce the affinity for the ligand. This further supports using structural effects for SNP prediction, where incorporation of structurally important regions could be considered. Additionally, the consideration of amino acid side chain structure could be incorporated into the system to provide further scope for investigation of structural alterations. This is especially useful for changes such as from a small side chain into a large one.

4.6 Conclusion

Structural analysis by application of heuristic functions has a strong and powerful potential in use of characterisation of SNPs with unknown clinical significance. Through additional work and expansion of the pipeline/system, this tool could be used to provide insights into pathogenic variants, the damage these cause, and ultimately for identifying targets for treatment or diagnosis of breast cancer and other diseases linked to BRCA1. The benefits of this tool expand beyond use with the BRCA1 gene, as the pipeline can be adapted for any gene and protein due to the nature of the input into the command line as it is a generic approach. The heuristic functions of a protein associated with structural changes appear to be a significant factor in determining the pathogenic effects of a SNP, where significant differences were observed between the wildtype and pathogenic data, and also between the benign and pathogenic absolute deviation dataset. Further investigation into both the role of associate genes taken from pathway analysis and consideration of the location of a SNP could provide further detail of what precise mechanism determines pathogenicity. Additionally, integrating physics systems to quantify thermodynamic change within a protein would yield insightful information about the impacts of changes in protein stability.

Further work on the project pipeline to scale data collection up to include more SNPs, a wider range of studied properties, and more comparisons between these properties would yield a deeper insight into the structural effects of SNPs, both in disease and benign variants. Furthermore, integration of machine learning techniques seen within current systems in section 4.4 would be highly beneficial for making the tool user friendly; fully automating the system to classify mutations as pathogenic or benign without such high reliance on clinical and laboratory investigations.

References

Abramowicz, A., & Gos, M. (2018). Splicing mutations in human genetic disorders: examples, detection, and confirmation. *Journal of Applied Genetics, 59,* 253 – 268

Adhikari, A.N., Peng, J., Wilde, M., Xu, J., Freed, K. F., & Sosnick, T. R. (2012). Modeling large regions in proteins: Applications to loops, termini, and folding. *Protein Sci, 21,* 107 – 121.

Ahmed, H., Alarabi, L., El-Sappagh, S., Soliman, H., & Elmogy, M. (2021). Genetic variations analysis for complex brain disease diagnosis using machine learning techniques: opportunities and hurdles. *PeerJ Computer Science, 7,* e697.

Ahn, S., Woo, J. W., Lee, K., & Park, S. Y. (2019). HER2 status in breast cancer: changes in guidelines and complicating factors for interpretation. *Journal of Pathology and Translational Medicine, 54,* 34 – 44.

Alsøe, L., Sarno, A., Carracedo, S., Domanska, D., Dingler, F., Lirussi, L., … & Nilsen, H. (2017). Uracil Accumulation and Mutagenesis Dominated by Cytosine Deamination in CpG Dinucleotides in Mice Lacking UNG and SMUG1. *Scientific Reports, 7,* 7199.

Bailey, R. (2020). *An Introduction to DNA Transcription.* Retrieved from https://www.thoughtco.com/dna-transcription-373398

Bayat, A. (2002). Science, Medicine, and the Future: Bioinformatics. *The BMJ, 324,* 1018 – 1022.

Bethesda. (2005). *SNP Class Definitions*. Retrieved from: https://www.ncbi.nlm.nih.gov/books/NBK44488/

Bhagavan, N. V., & Ha, C. E. (2015). *Essentials of Medical Biochemistry*. (2nd ed.). Academic Press

Billing, D., Horiguchi, M., Wu-Baer, F., Taglialatela, A., Leuzzi, G., Nanez, S. A., … & Baer, R. (2018). The BRCT Domains of the BRCA1 and BARD1 Tumor Suppressors Differentially Regulate Homology-Directed Repair and Stalled Fork Protection. *Molecular Cell, 1,* 127 – 139.

Birolo, G., Benevenuta, S., Fariselli, P., Capriotti, E., Giorgio, E., & Sanavia, T. (2021). Protein Stability Perturbation Contributes to the Loss of Function in Haploinsufficient Genes. *Frontiers in Molecular Biosciences, 8,* 620793

Boundless General Microbiology. (2021). *Elongation and Termination in Eukaryotes*. Retrieved from https://bio.libretexts.org/@go/page/9279

Boutorh, A., & Guessoum, A. (2015). Classication of SNPs for breast cancer diagnosis using neural-network-based association rules. *12th International Symposium on Programming and Systems (ISPS).*

Brown, D. K., & Bishop, O. T. (2017). The role of structural bioinformatics in drug discovery via computational SNP analysis – a proposed protocol for analyzing variation at the protein level. *Global Heart, 12,* 151 – 161.

Budowle, B., & van Daal, A. (2018). Forensically relevant SNP classes. *Biotechniques, 44.*

Chatzou, M., Magis, C., Chang, J. M., Kemena, C., Bussotti, G., Erb, I., & Notredame, C. (2015). Multiple sequence alignment modeling: methods and applications. *Briefings in Bioinformatics,* 1 – 15.

Cheriyedath, S. (2019). *Protein Folding.* Retrieved from https://www.news-medical.net/life-sciences/Protein-Folding.aspx

Clancy, S. & Brown, W. (2008) Translation: DNA to mRNA to Protein. *Nature Education, 1*, 101. Retrieved from https://www.nature.com/scitable/topicpage/translation-dna-to-mrna-to-protein-393/

Clancy, S. (2008). Genetic Mutation. *Nature Education, 1,* 187.

Clark, D. P., Pazdernik, N. J., & McGhee, M. R. (2019). Mutations and Repair. In *Molecular Biology*. (3rd ed., 832 – 879). Academic Cell.

Coignard, J., Lush, M., Beesley, J., O'Mara, T. A., Dennis, J., Tyrer, J. P., … & Antoniou, A. C. (2021). A case-only study to identify genetic modifiers of breast cancer risk for BRCA1/BRCA2 mutation carriers. *Nature Communications, 12,* 1078.

Cui, D., Ou, S., & Patel, S. (2014). Protein-spanning water networks and implications for prediction of protein–protein interactions mediated through hydrophobic effects. *Proteins, 82,* 3312 – 3326.

Da Ines, O., Degroote, F., Goubely, C., Amiard, S., Gallego, M. E., & White, C. I. (2013). Meiotic Recombination in Arabidopsis Is Catalysed by DMC1, with RAD51 Playing a Supporting Role. *PLoS Genetics, 9,* 1003787.

Deng, N., Zhou, H., Fan, H., & Yuan, Y. (2017). Single nucleotide polymorphisms and cancer susceptibility. *Oncotarget, 8,* 110635 – 110649.

Ding, X., & Cheng, J (2011). MSACompro: protein multiple sequence alignment using predicted secondary structure, solvent accessibility, and residue-residue contacts. *BMC Bioinformatics, 12,* 472.

Dong, X., & Zhang. Y. (2011). Improving the Physical Realism and Structural Accuracy of Protein Models by a Two-Step Atomic-Level Energy Minimization. *Biophysical Journal, 101,* 2525 - 2534.

Eichler, J. (2019). Protein glycosylation. *Current Biology, 29,* 229 – 231.

EMBL-EBI. (2021). *InterPro.* Retrieved from https://www.ebi.ac.uk/interpro/

Erdoğan, O., & Son, Y. A. (2014). Predicting the Disease of Alzheimer With SNP Biomarkers and Clinical Data Using Data Mining Classification Approach: Decision Tree. *Studies in Health Technology and Informatics, 205,* 511 – 515.

Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M.Y., Pieper, U., & Sali, A. (2016). Comparative protein structure modeling using Modeller. *Curr Protoc, 5,* 5 -6.

Frenkel, D., & Smit, B. (2007). *Understanding molecular simulation : from algorithms to applications* (2nd ed.) San Diego : Academic Press.

Fu, J., Gao, J., Liang, Z., & Yang, D. (2021). PDI-Regulated Disulfide Bond Formation in Protein Folding and Biomolecular Assembly. *Molecules, 26,* 171.

Gao, C., & Wang, Y. (2020). mRNA Metabolism in Cardiac Development and Disease: Life After Transcription. *Physiological Reviews, 100,* 673 – 694.

Gaudillo, J., Rodriguez, J. J. R., Nazareno, A., Baltazar, L. R., Vilela, J., Bulalacao, R., Domingo, M., & Albia, J. (2019). Machine learning approach to single nucleotide polymorphism-based asthma prediction. *PLOS One, 14,* e0225574.

Gheorghiu, A., Coveney, P. V., Arabi, A. A. (2020). The influence of base pair tautomerism on single point mutations in aqueous DNA. *Interface Focus, 10.*

Gong, S., Worth, C. L., Cheng, T. M. K., Blundell, T. L. (2011). Meet Me Halfway: When Genomics Meets Structural Bioinformatics. *Journal of Cardiovascular Translational Research, 4,* 281 – 303.

Grover, A., & Sharma, P. C. (2013). Development and use of molecular markers: past and present. *Critical Reviews in Biotechnology, 2,* 290 – 302.

Grumezescu, A.M. (2018). *Fullerens, Graphenes and Nanotubes : A Pharmaceutical Approach* (1st. ed.) Retrieved from http://ebookcentral.proquest.com/lib/swansea-ebooks/detail.action?docID=5405543

Guirouilh-Barbat, J., Lambert, S., Bertrand, P., & Lopez, B. S. (2014). Is homologous recombination really an error-free process?. *Frontiers in Genetics, 5,* 175.

Haim, A., Neubacher, S., & Grossmann, T. N. (2021). Protein Macrocyclization for Tertiary Structure Stabilization. *ChemBioChem*.

Health Sciences Library System. (2014). *SuperLooper - Loop Structure Predictor.* Retrieved from https://www.hsls.pitt.edu/obrc/index.php?page=URL1250175982

Helyar, S. J., Hemmer-Hanson, J., Bekkevold, D., Taylor, M. I., Ogden, R., Limborg, M. T., … & Nielsen, E. E. (2011). Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Molecular Ecology Resources, 11*, 123-136

Holroyd, L. F., & van Mourik, T. (2015). Stacking of the mutagenic base analogue 5-bromouracil: energy landscapes of pyrimidine dimers in gas phase and water. *Physical Chemistry Chemical Physics, 17,* 30364-30370.

Husain, M. A., Ishqi, H. M., Sarwar, T., Rehman, S. U., & Tabish, M. (2017). Interaction of indomethacin with calf thymus DNA: a multi-spectroscopic, thermodynamic and molecular modelling approach. *MedChemComm, 8,* 1283 – 1296.

IBM Cloud Education. (2019). *Relational Databases.* Retrieved from https://www.ibm.com/cloud/learn/relational-databases

Jimenez-Sainz, J., & Jensen, R. B. (2021). Imprecise Medicine: BRCA2 Variants of Uncertain Significance (VUS), the Challenges and Benefits to Integrate a Functional Assay Workflow with Clinical Decision Rules. *Genes (Basel), 12,* 780.

Jin, B., Li, Y., & Robertson, K.D. (2011). DNA Methylation. *Genes & Cancer, 6,* 607 – 617.

Jun, S., H., Warner, B.A., & Murakami, K. S. (2013). RNA Polymerase Reaction in Bacteria. *Encyclopaedia of Biological Chemistry,* 167 – 172. Retrieved from https://www.sciencedirect.com/science/article/pii/B9780123786302006290

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols, 10,* 845 – 858.

Khan Academy. (2021). *Overview of transcription*. Retrieved from https://www.khanacademy.org/science/ap-biology/gene-expression-and-regulation/transcription-and-rna-processing/a/overview-of-transcription

Khor, B. Y., Tye, G. J., Lim, T. S., & Choong, Y. S. (2015). General overview on structure prediction of twilight-zone proteins. *Theoretical Biology and Medical Modelling, 12.*

Khoury, G. A., Baliban, R. C., & Floudas, C. R. (2011). Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Scientific Reports, 1,* 90.

Kim, J. H., Park, S., Park, H. S., Park, J. I., Lee, S-T., Kim, S-W … & Jung, S. H. (2021). Analysis of BRCA1/2 variants of unknown significance in the prospective Korean Hereditary Breast Cancer study. *Scientific Reports, 11.*

Krissinel, E. (2007). On the relationship between sequence and structure similarities in proteomics. *Bioinformatics, 23,* 717 - 723.

Landrum, M., Chitipiralla, S., Brown, G. R., Chen, C., Gu, B., Hart, J., … & Kattman, B. L. (2020). ClinVar: improvements to accessing data. *Nucleic Acids Research, 48,* D835 – D844.

Landrum, M., J., Lee, J., M., Benson, M., Brown, G., Chao, C., Chitipiralla, S.,… Maglott D., R. (2015). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research, 44,* D862–D868

Leech, M. D., & Brown, A. J. P. (2012). Posttranslational Modifications of Proteins in the Pathobiology of Medically Relevant Fungi. *Eukaryotic Cell, 11,* 98 – 108.

Levene M. (2005) *The Nested Universal Relational Database Model* (1st ed.). Retrieved from https://link.springer.com/book/10.1007/3-540-55493-9#toc

Li, D., Fedeles, B. I., Singh, V., Peng, C. S., Silvestre, K. J., Simi, A. K., … & Essigmann, J. M. (2014). Tautomerism provides a molecular explanation for the mutagenic properties of the anti-HIV nucleoside 5-aza-5,6-dihydro-2′-deoxycytidine. *PNAS, 111.*

LibreTexts. (2021). *Protein Folding.* Retrieved from https://chem.libretexts.org/@go/page/496

Liu, D., Gao, Y., Li, L., Chen, H., Bai, L., Qu, Y… & Zhao, Y. (2021). Single nucleotide polymorphisms in breast cancer susceptibility gene 1 are associated with susceptibility to lung cancer. *Oncology Letters, 21,* 424

Liu, J-J., Yu, C-S., Wu, H-S., Chang, Y-J., Lin, C-P., & Lu, C-H. (2021). The structure-based cancer-related single amino acid variation prediction. *Scientific Reports, 11, 13599.*

Liu, L., & Lu, L-Y. (2020). BRCA1 and homologous recombination: implications from mouse embryonic development. *Cell & Bioscience, 10.*

Makigaki S., & Ishida, T. (2020). Sequence alignment using machine learning for accurate template-based protein structure prediction. *Bioinformatics, 36,* 104 – 111.

Moes, G., & Sheldon, R. (2005). *Beginning MySQL* (1st ed.). Retrieved from https://www.book-info.com/isbn/0-7645-7950-9.htm

Molinaro, C., Martoriati, A., & Cailliau, K. (2021). Proteins from the DNA Damage Response: Regulation, Dysfunction, and Anticancer Strategies. *Cancers (Basel), 13,* 3819.

Mulloy, B., Dell, A., Stanley, P., & Prestegard, J. H. (2017). Chapter 50 Structure Analysis of Glycans. *Essentials of Glycobiology* (3rd. ed.). Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK453059/

Murthy, P., & Muggia, F. (2019). Women's cancers: how the discovery of BRCA genes is driving current concepts of cancer biology and therapeutics. *Ecancermedicalscience, 13.*

Mylavarapu, S., Das, A., & Roy, M. (2018). Role of BRCA Mutations in the Modulation of Response to Platinum Therapy. *Frontiers in Oncology, 8,* 16.

Nakamori, M., Panigrahi, G. B., Lanni, S., Gall-Duncan, T., Hayakawa, H., Tanaka, H., … & Pearson, C. E. (2020). A slipped-CAG DNA-binding small molecule induces trinucleotide-repeat contractions in vivo. *Nature Genetics, 52,* 146 – 159.

National Human Genome Research Institute. (2020). *Genome-Wide Association Studies Fact Sheet.* Retrieved from https://www.genome.gov/about-genomics/fact-sheets/Genome-Wide-Association-Studies-Fact-Sheet

NCBI ClinVar. (2019). *How to search ClinVar.* Retrieved from https://www.ncbi.nlm.nih.gov/clinvar/docs/help/

NCBI ClinVar. (2021). *Search results for BRCA1[gene] Sort by: Position Filters: Single nucleotide.* Retrieved from https://www.ncbi.nlm.nih.gov/clinvar/?term=BRCA1%5Bgene%5D on 1st September 2021

NCBI ClinVar. (2021). *Search Results for Search human Sort by: Position Filters: Single nucleotide.* Retrieved from https://www.ncbi.nlm.nih.gov/clinvar on 1st September 2021.

NCBI dbSNP. (2021). *Search results for homo sapien[Organism].* Retrieved from https://www.ncbi.nlm.nih.gov/snp/?term=homo%20sapien%5BOrganism%5D on 1st September 2021.

Nguyen, T-T., Huang, J. Z., Wu, Q., Nguyen, T. T., Li, M., J. (2015). Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests. *BMC Genomics, 16,* S5.

Okorokov, A. L., Chaban, Y. L., Bugreev, D. V., Hodgkinson, J., Mazin, A. V., & Orlova, E. V. (2010). Structure of the hDmc1-ssDNA filament reveals the principles of its architecture. *PLoS One, 5,* 8586.

O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D. J, & Notredame, C. (2004). 3DCoffee: Combining Protein Sequences and Structures within Multiple Sequence Alignments. *Journal of Molecular Biology, 340,* 385 – 395.

Otsuka, K., Yoshino, Y., Qi, H., & Chiba, N. (2020) The Function of BARD1 in Centrosome Regulation in Cooperation with BRCA1/OLA1/RACK1. *Genes (Basel), 11,* 842.

Ozturk, K., & Carter, H. (2021). Predicting functional consequences of mutations using molecular interaction network features. *Human Genetics, pre-print.*

Pace, C. N., Fu, H., Fryar, K. L., Landua, J., Trevino, S. R., Schell, D. … & Grimsley, G. R. (2014). Contribution of hydrogen bonds to protein stability. *Protein Science, 23,* 652 – 661.

Padmanabhan, S. (2014). *Handbook of Pharmacogenomics and Stratified Medicine* (1st. ed.) Retrieved from http://ebookcentral.proquest.com/lib/swansea-ebooks/detail.action?docID=1683294

Pandey, N. V. (2020). DNA viruses and cancer: insights from evolutionary biology. *Virusdisease, 31,* 1 – 9.

Parker, N., Schneegurt, M., Thi Tu, A. H., Lister, P., & Forster, B. M. (2017). *Microbiology*. OpenStax CNX

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., & Ferrin, T.E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem, 25* 1605 - 1612.

Phillips, C. (2012). Application of Autosomal SNPs and Indels in Forensic Analysis. *Forensic Science Review, 24,* 43 – 62.

Pollard, T. D., Earnshaw, W. C., Lippincott-Schwartz, J., & Johnson, G. T. (2017). *Cell Biology* (3rd ed.). Elsevier.

Pollard, T., Earnshaw, W. C., Lippincott-Schwartz, J., & Johnson, G. T. (2017) *Cell Biology.* (3rd ed.). DOI: https://doi.org/10.1016/C2014-0-00272-9

Raval, A., Piana, S., Eastwood, M.P., Dror, R.O., & Shaw, D.E. (2012). Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. *Proteins, 80*, 2071 - 2079.

Reactome Organisation. (2021). *What is Reactome?* Retrieved from https://reactome.org/what-is-reactome

Reily, C., Stewart, T. J., Renfrow, M., B., & Novak, J. (2019). Glycosylation in Health and Disease. *Nature Reviews Nephrology, 15,* 346 – 366.

Roberts, J. (2019). Mechanisms of Bacterial Transcription Termination. *Journal of Molecular Biology, 431,* 4030 – 4039.

Roy, A., Kucukural, A., & Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols, 5,* 725–738.

Rozewicki, J., Li, S., Amada, K. M., Standley, D. M., Katoh, K. (2019). MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Research, 47,* 5 – 10.

Sanavia, T., Birolo, G., Montanucci, L., Turina, P., Capriotti, E., & Fariselli, P. (2020). Limitations and challenges in protein stability prediction upon genome variations: towards future applications in precision medicine. *Computational and Structural Biotechnology Journal, 18,* 1968 – 1979.

Saxena, A., Sangwan, R. S., & Mishra, S. (2013). Fundamentals of Homology Modeling Steps and Comparison among Important Bioinformatics Tools: An Overview. *Science International, 1,* 237-252

Schenker, G. N. (2020). *Learn Docker - Fundamentals of Docker 19.x* (2nd ed.). Retrieved from https://web.archive.org/web/20110406121920/http://java.sun.com/developer/Books/jdbc/ch07.pdf

Short, J. M., Liu, Y., Chen, S., Soni, N., Madhusudhan, M. S., Shivji, M. K. K., & Venkitaraman, A. R. (2016). High-resolution structure of the presynaptic RAD51 filament on single-stranded DNA by electron cryo-microscopy. *Nucleic Acid Research, 44,* 9017 – 9030.

Soto C. S., Fasnacht, M., Zhu, J., Forrest, L., & Honig, B.(2008). Loop modeling: Sampling, filtering, and scoring. *Proteins, 70,* 834 – 843.

Sterne-Weiler, T., & Sanford, J. R. (2014). Exon identity crisis: disease-causing mutations that disrupt the splicing code. *Genome Biology, 15,* 201.

SQLite Organisation. (2021). *SQLite is a Zero-Configuration Database.* Retrieved from https://sqlite.org/zeroconf.html#:~:text=SQLite%20Is%20A%20Zero-Configuration%20Database%20SQLite%20does%20not,that%20needs%20to%20be%20started%2C%20stopped%2C%20or%20configured.

Taly, J.F., Magis, C., Bussotti, G., Chang. J.M., Di Tommaso, P., Erb, I., Espinosa-Carrasco, J., Kemena, K., Notredame , C. (2011). Using the T-Coffee package to build multiple sequence alignments of protein, RNA, DNA sequences and 3D structures. *Nature Protocols, 6,* 1669 – 1682.

Tang, K., Zhang, J., & Liang, J. (2014) Fast Protein Loop Sampling and Structure Prediction Using Distance-Guided Sequential Chain-Growth Monte Carlo Method. *PLoS Comput Biol*, *10*.

Tavares, E. M., Wright, W. D., Heyer, W-D., Le Cam, E., & Dupaigne, P. (2019). In vitro role of Rad54 in Rad51-ssDNA filament-dependent homology search and synaptic complexes formation. *Nature Communications, 10,* 4058.

Thompson, J. D., Linard, B., Lecompte, O., & Poch, O. (2011). A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods: Current Challenges and Future Perspectives. *PLoS One, 6.*

UCSF Computer Graphics Laboratory. (2014). *FindHBond.* Retrieved from https://www.cgl.ucsf.edu/chimera/docs/ContributedSoftware/findhbond/findhbond.html

UniProt Consortium. (2021). *About UniProt*. Retrieved from https://www.uniprot.org/help/about on 26th August 2021

UniProtKB/Swiss-Prot Consortium. (2021). *UniProtKB/Swiss-Prot protein knowledgebase release 2021_03 statistics.* Retrieved from https://web.expasy.org/docs/relnotes/relstat.html#top on 26th August 2021

UniProtKB/TrEMBL Consortium. (2021). *Current Release Statistics.* Retrieved from https://www.ebi.ac.uk/uniprot/TrEMBLstats on 26th August 2021

Uversky, V. N. (2013). Brenner's Encyclopedia of Genetics. (2nd Ed.). ISBN: 978-0-08-096156-9

vlab.amrita.edu. (2012). Homology Modeling using Modeller. Retrieved from vlab.amrita.edu/?sub=3&brch=275&sim=1481&cnt=1

Voet, D. (2006). *Fundamentals of Biochemistry.* (2nd ed.). Hoboken : Wiley

Voet, D., Voet, J.G., & Pratt, C.W. (2016). *Principles of Biochemistry* (Fifth ed.). Wiley.

Wang, L., Qian, J., Yang, Y., & Gu, C. (2021). Novel insights into the impact of the SUMOylation pathway in hematological malignancies (Review). *International Journal of Oncology, 59,* 73.

Warren, B. (2020). *Translation of DNA.* Retrieved from https://teachmephysiology.com/biochemistry/protein-synthesis/dna-translation/

Widen, E., Raben, T. G., Lello, L., & Hsu, S. D. H. (2021). Machine Learning Prediction of Biomarkers from SNPs and of Disease Risk from Biomarkers in the UK Biobank. *Preprints, 1.*

Wu, L., Chu, X., Zheng, J., Xiao, C., Zhang, Z., Huang, G., … & Xiong, B. (2019). Targeted capture and sequencing of 1245 SNPs for forensic applications. *Forensic Science International: Genetics, 42,* 227 – 234.

Xu, G-P., Zhao, Q., Wang, D., Xie, W-Y., Zhang, L-J., Zhou, H… & Wu, L-F. (2018). The association between BRCA1 gene polymorphism and cancer risk: a meta-analysis. *Oncotarget, 9*, 8681 – 8694.

Yang, J., & Zhang, Y. (2015). I-TASSER server: new development for protein structure andfunction predictions*. Nucleic Acids Research, 43*, 174-181.

Yoshino, Y., Fang, Z., Qi, H., Kobayashi, Q., & Chiba, N. (2021). Dysregulation of the centrosome induced by BRCA1 deficiency contributes to tissue-specific carcinogenesis. *Cancer science, 112,* 1679 – 1687.

Zeng, Y. F., & Sang, J. (2017). Five zinc finger protein 350 single nucleotide polymorphisms and the risks of breast cancer: a meta-analysis. *Oncotarget, 8,* 107273-107282.

Zenkin N. (2014). Ancient RNA stems that terminate transcription. *RNA Biology, 11,* 295 – 297.

Zhang, C., Freddolino, P. L., & Zhang, Y. (2017) COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information. *Nucleic Acids Research, 45,* 291 - 299.

Zhang, J. (2013). The role of BRCA1 in homologous recombination repair in response to replication stress: significance in tumorigenesis and cancer therapy. *Cell & Biosciences, 3,* 11.

Zhang, S-Y., & Liu, S-L. (2013). Bioinformatics. In Maloy, S., & Hughes, K. (Eds.) *Brenner's Encyclopaedia of Genetics* (2nd ed., 338 – 340). Academic Press

Zhao, Q., Ma, Y., Li, Z., Zhang, K., Zheng, M., & Zhang, S. (2020). The Function of SUMOylation and Its Role in the Development of Cancer Cells under Stress Conditions: A Systematic Review. *Stem Cells International, 2020,* 8835714.