Universitat Autònoma de Barcelona

Novel approaches for in silico identification of pathogenic variants in BRCA1 and BRCA2 hereditary breast and ovarian cancer predisposition genes

Natàlia Padilla Sirera

Novel approaches for *in silico* identification of pathogenic variants in BRCA1 and BRCA2 hereditary breast and ovarian cancer predisposition genes

Natàlia Padilla Sirera

PhD thesis

Doctorat en Biotecnologia

Director: Dr. Xavier de la Cruz

Tutor: Dr. Enric Querol

Unitat de Bioinformàtica Clínica i Translacional

Vall d'Hebron Institut de Recerca

Departament de Bioquímica i Biologia Molecular

Facultat de Medicina

Universitat Autònoma de Barcelona

Barcelona, 2020

Vall d'Hebron Institut de Recerca

UAB Universitat Autònoma de Barcelona

A la meva família

Agraïments

Primer de tot, volia agrair a la meva família, per creure en mi i estar sempre al meu costat, ajudar-me a créixer i ensenyar-me els valors de la vida.

També els hi volia agrair als meus amics, amb els que junts hem descobert el món i ens hem fet grans, a la Eva, David, Pau, Mònica, Sílvia, Josep, Joan, Christian, Ingrid i Laura del cau; i a la Júlia, Jusep, Alba, David, Ester, Gonzalo i Sarai del Pele.

Al meu supervisor, el Dr. Xavier de la Cruz, per haver-me donat aquesta gran oportunitat i haver-me guiat durant aquest camí, moltes gràcies!

Als companys de laboratori: a la Selen, la millor companya de laboratori que es pot desitjar i a la Luz, una gran científica, els hi desitjo molta sort amb els seus doctorats! I als que ja han marxat: al Josu pels grans debats i anàlisis, la Casandra, per introduir-me al món científic, a l'Elena i l'Òscar. Haver format grup amb vosaltres ha estat una gran sort!

A les col·laboracions del VHIO i del VHIR, especialment al grup de la Sara Gutiérrez per la seva incansable energia i valuosos comentaris; i al grup del Roger Colobran i a la Laura Viñas, per la seva dedicació i determinació. També als nostres veïns d'Estadística per tots els moments compartits.

A tot el professorat de la UAB, UPF i l'escola Joan Pelegrí, que sense donar-se compte, m'han ajudat a ser on sóc. A tots els amics que he fet durant al camí, especialment al màster de Bioinformàtica, les llicenciatures de Biologia i Bioquímica, i la comunitat de Python de Barcelona.

Per últim, volia agrair-te a tu Dani, per acompanyar-me en aquest viatge.

Gràcies a tots!

Caminante no hay camino,

se hace camino al andar

Antonio Machado

Abstract

Germline variants in BRCA1 and BRCA2 can disrupt the DNA protective role of these proteins resulting in an increased risk of developing hereditary breast and ovarian cancer (HBOC). Identification of those individuals carrying pathogenic variants will allow channeling them into specific programs of prevention and surveillance, incrementing their survival rates. For this purpose, first, it is necessary to identify which of the variants are pathogenic. Unfortunately, there is not always enough information to reach a conclusion. In this situation, pathogenicity predictors designed to computationally estimate the damage caused by variants, can provide valuable information.

Here, we present a novel family of pathogenicity predictors for BRCA1 and BRCA2. These predictors differ in their objective: one is trained to estimate the molecular impact of variants on the HDR function of BRCA1 and BRCA2, and the other is trained to estimate the clinical significance of a variant, that is, whether it should be classified as pathogenic or neutral. Their performances have been tested and are comparable to those of widely used predictors in the field. Additionally, we presented them to the ENIGMA challenge from the 5th Critical Assessment of Genome Interpretation (CAGI), finding that our predictors, especially those estimating the functional impact of variants, ranked in the top positions compared to other tools.

In order to disseminate this family of predictors to the scientific community, we have built the BRASS website (https://www.biotoclin.org/BRASS), where users can analyze their missense BRCA1 and BRCA2 variants. More advanced users can also interpret the predictions using a reliability metric and several plots contextualizing the score to that of a set of manually curated variants.

Independently, we applied our knowledge about pathogenicity predictors in a large international effort to characterize a novel pediatric neurologic disorder caused by pathogenic variants in histone H3.3. We combined the use of standard pathogenic predictors with evidence from structural analyses and biophysical computations to provide a mechanistic view of the impact of the causative variants.

Resum

Variants germinals a les proteïnes BRCA1 i BRCA2 poden alterar la funció protectora d'aquestes a l'ADN, incrementant el risc de desenvolupar càncer de mama i ovari hereditari (HBOC). Identificació d'aquells individus portadors de variants patogèniques permet canalitzar-los en programes específics de prevenció i vigilància, augmentant les seves taxes de supervivència. Per això, en primer lloc, cal identificar quines de les variants són patogèniques. Malauradament, no sempre hi ha prou informació per arribar a una conclusió. En aquesta situació, els predictors de patogenicitat dissenyats per estimar computacionalment el dany causat per les variants poden proporcionar una valuosa informació.

En aquest treball presentem una nova família de predictors de patogenicitat per BRCA1 i BRCA2. Aquests predictors difereixen en el seu objectiu: un està entrenat per estimar l'impacte molecular de les variants sobre la funció HDR de BRCA1 i BRCA2, i l'altre està entrenat per estimar la significació clínica d'una variant, és a dir, si la seva classificació és patogènica o neutra. Els seus rendiments han estat provats i són comparables als d'altres mètodes àmpliament utilitzats en el camp. Addicionalment, vam presentar els predictors al repte ENIGMA de la 5a Avaluació Crítica de la Interpretació del Genoma (CAGI), trobant que els nostres mètodes, especialment aquells que estimen l'impacte funcional de les variants, es classifiquen en les primeres posicions en comparació amb les altres eines.

Per tal de difondre aquesta família de predictors a la comunitat científica, hem construït el lloc web BRASS (https://www.biotoclin.org/BRASS), on els usuaris poden analitzar les seves variants de BRCA1 i BRCA2 amb canvi de sentit. Els usuaris més avançats també poden interpretar les prediccions

mitjançant una mètrica de fiabilitat i diversos gràfics contextualitzant la seva puntuació amb la d'un conjunt de variants curades manualment.

Independentment, hem aplicat els nostres coneixements sobre predictors de patogenicitat en un gran projecte internacional per caracteritzar un nou trastorn neurològic pediàtric causat per variants patogèniques a la histona H3.3. Vam combinar l'ús de predictors de patogenicitat estàndard amb evidències d'anàlisis estructurals i càlculs biofísics per proporcionar una visió mecanicista de l'impacte de les variants causals.

Table of contents

1. Introduction

1.1. Hereditary breast and ovarian cancer

Breast cancer is the second most commonly diagnosed cancer in the world and the first among women, with an estimated of 2.1 million new cases each year (Bray et al., 2018). It ranks as the fifth cause of death from cancer worldwide, and the first cause of cancer death among women. In Europe, the estimated numbers of new cancer cases indicate that breast cancer is the most commonly diagnosed and the main cause of cancer death among women, with the highest incidence observed in Western Europe, notably in Belgium, Luxembourg and The Netherlands; and in Northern Europe, particularly in United Kingdom, Sweden and Finland (Figure 1.1) (Bray et al., 2018; Ferlay et al., 2018).

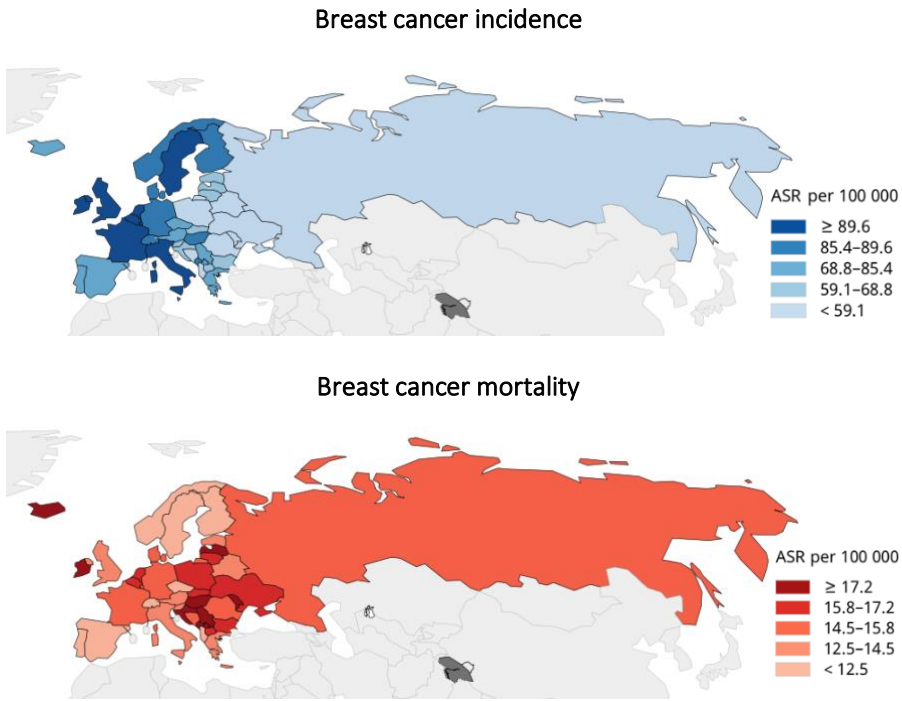


Figure 1.1 ***Breast cancer incidence and mortality among women in Europe***. *Age-standardized rate (ASR) estimates in females ages 0-74 from the Global Cancer Observatory http://gco.iarc.fr.*

Ovarian cancer is a less frequent cancer that causes 295,400 new cases every year, ranking as the eighth most diagnosed cancer among women worldwide (Bray et al., 2018). It also represents the eighth cause of female cancer death in the world. In Europe, it is the sixth most diagnosed cancer and the fifth cause of cancer death among women, with the highest incidence in Eastern Europe (Figure 1.2) (Bray et al., 2018; Ferlay et al., 2018).

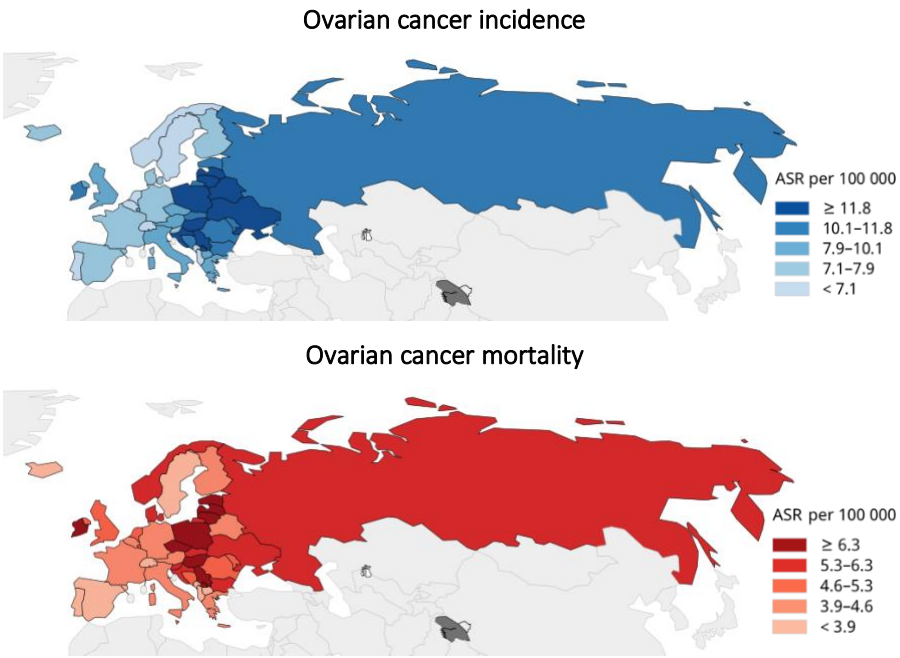


Figure 1.2 ***Ovarian cancer incidence and mortality among women in Europe***. *Age-standardized rate (ASR) estimates in females ages 0-74 from the Global Cancer Observatory http://gco.iarc.fr.*

Approximately, 5-10% of breast and 20% of ovarian cancer patients have a genetic component segregating in their family and are classified as familial or hereditary cancer (Newman, Austin, Lee, & King, 1988; Russo et al., 2009). Familial breast cancer is characterized by an early-onset of the disease, younger age of diagnosis, frequent bilateral cancer and high incidence in men (J. M. Hall et al., 1990). In contrast, the remaining 90% of breast and ovarian cancer cases occur without a family history and are referred as sporadic.

In the 1990s, an important step in the molecular understanding of breast and ovarian cancer took place. Numerous families affected by these cancers were studied by linkage analysis to identify high-risk susceptibility genes. Cosegregation of markers led to the identification of linkage to chromosomes 17q and 13q; and subsequent positional cloning led to the identification of *BRCA1* gene on 17q11 in 1994 and *BRCA2* gene on 13q12–q13 in the next year (J. M. Hall et al., 1990; Miki et al., 1994; Wooster et al., 1995).

BRCA1 and *BRCA2* are two tumour suppressor genes involved in DNA repair mechanisms. Germline pathogenic variants in one of these two genes result in hereditary breast and ovarian cancer (HBOC) syndrome (Roy, Chun, & Powell, 2012). HBOC is a disorder inherited in an autosomal dominant manner and with an incomplete penetrance. HBOC is characterized by an increased risk of breast and ovarian cancer (Ford et al., 1998; King, Marks, & Mandell, 2003). Women with HBOC have a lifetime risk of 46% - 87% of developing breast cancer and of 11% - 63% for ovarian cancer. HBOC also confers an slightly increased risk of male breast, prostate, pancreatic and melanoma cancer (Table 1.1) (Chen et al., 2006; Easton et al., 1995; Mavaddat et al., 2013; Moran et al., 2012; Deborah Thompson & Easton, 2003).

	Risk of developing cancer		
Cancer Type	**BRCA1 variant carrier**	**BRCA2 variant carrier**	**General population**
Breast	46% - 87%	38% - 84%	12%
Second primary breast	21% in 10 years	11% in 10 years	2% in 5 years
Ovarian	39% - 63%	11% - 27%	1% - 2%
Male breast	1.2%	6.8%	0.1%
Prostate	9% by age 65	15% by age 65	6% by age 69
Pancreatic	1% - 3%	2% - 7%	0.5%
Melanoma		3%	1.6%

Table 1.1 ***Risk of malignancy in BRCA1 and BRCA2 carriers.*** *Adapted from GeneReviews at https://www.ncbi.nlm.nih.gov/sites/books/NBK1247/.*

Interindividual variability in the risk of breast and ovarian cancer is attributed to both environmental and genetic factors, including the location and type of variants in *BRCA1* and *BRCA2* (Fackenthal & Olopade, 2007). In early reports, it was suggested that the location of nonsense and frameshift variants in the central regions of *BRCA1/2*, termed ovarian cancer cluster regions (OCCR), were associated with a greater risk of ovarian cancer than similar variants in the proximal and distal regions of each gene (D. Thompson & Easton, 2001).

Understanding the genetic component of HBOC plays an important role in the medical management of patients. We know that in order to improve the patient outcome and survival, early detection is critical. In fact, most countries recommend an annual screening for breast cancer in women at 50 - 74 years old for an early diagnosis (Shah & Guraya, 2017). The advent of Next Generation Sequencing (NGS), a highly scalable technology of massive sequencing, has represented an important advance, making genetic testing widely available. Genetic testing of *BRCA1/2* and other breast and ovarian cancer susceptibility genes enable an accurate risk assessment of variants and, importantly, the identification of those individuals carrying high risk-variants who can then benefit from enhanced screenings and prevention strategies (Castéra et al., 2014).

Thus, identification of pathogenic variants predisposing to cancer represent a breakthrough in the management of HBOC patients. However, to apply it massively, we need to understand the clinical significance of each variant. Unfortunately, this is still an unsolved problem, and the clinical interpretation of variants remains an open challenge, especially for those variants of uncertain significance (VUS). VUS are variants that have been identified but lack sufficient evidence to be classified as pathogenic or benign. As a consequence, patients carrying VUS may experience delays accessing preventive and therapeutic target measures.

1.1.1. Genetic landscape of HBOC

As we have seen before, *BRCA1* and *BRCA2* are the most important high-penetrant genes predisposing to HBOC. However, only 25% of families with HBOC have variants in *BRCA1* or *BRCA2* (Kast et al., 2016).

Studies of families testing negative for *BRCA1/2*, early-onset of the disease, and a high number of individuals affected, have led to the identification of other susceptibility genes such as the highly penetrant genes *TP53*, *PTEN*, *CDH1* and *STX11*; and the moderately penetrant genes *ATM*, *CHEK2*, and *PALB2* (Figure 1.3), which are related to the genome maintenance pathways of *BRCA1* and *BRCA2*.

Moreover, genome-wide association studies (GWAS) in a large number of breast cancer patients resulted in the identification of common genetic variants in 76 loci associated with small increases in the risk of breast cancer (Couch, Nathanson, & Offit, 2014). But all these predisposing genes and SNPs can only explain 50% of all familial cases affected by HBOC. The other half of inheritance still remains unknown (Couch et al., 2014).

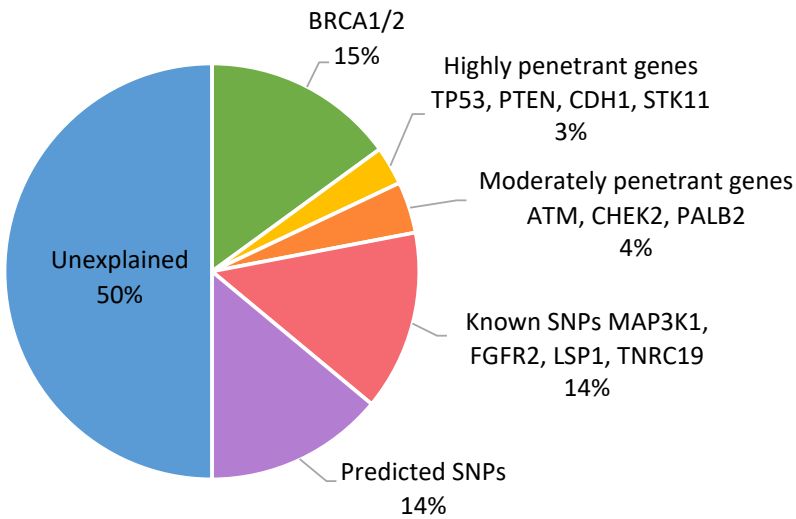


Figure 1.3 ***Estimated contributions of pathogenic variants in familial cases of HBOC.*** *Adapted from Couch et al. (Couch et al., 2014).*

1.1.2. BRCA1/2 genes: domain structure and function

In this work, we will focus on the study of *BRCA1* and *BRCA2* genes, since they are the most highly penetrant and well-studied genes. These genes code for two large nuclear proteins, BRCA1 and BRCA2, that act as tumour suppressors and show no homology to each other or to previously described proteins. In addition to having similar disease phenotypes, they both play a key role in maintaining genome integrity through several mechanisms such as repairing DNA double-strand breaks (DSBs) by homologous recombination (HR), protecting stalling DNA replication forks and controlling DNA damage in cell cycle checkpoints (Figure 1.4).

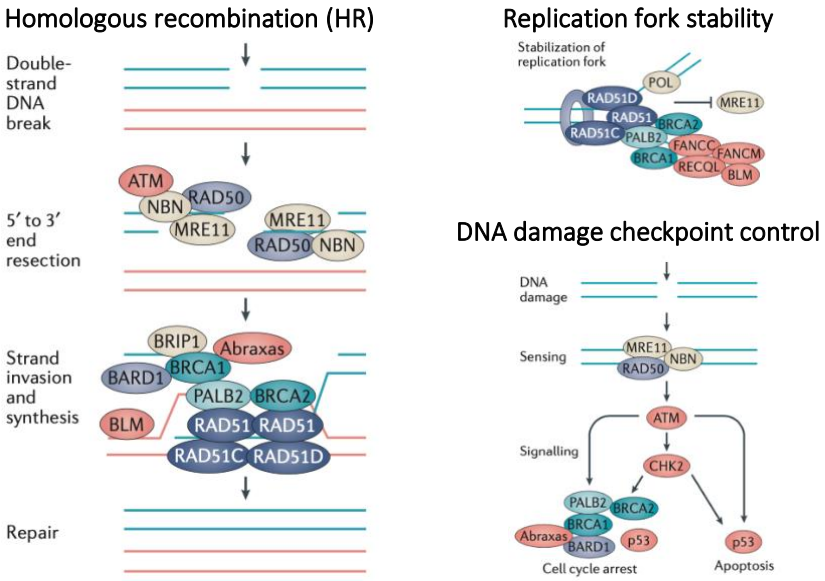


Figure 1.4 ***Genomic stability pathways and genes in HBOC.*** *Adapted from Nielsen et al. (Nielsen, Van Overeem Hansen, & Sørensen, 2016).*

In this thesis, we will focus in HR, a vital DNA repair process underlying the oncogenic impact of variants in BRCA1/2. HR appears to be the major mechanism protecting the integrity of the genome in proliferating cells, using the undamaged sister chromatid to carry out high-fidelity repair of double-strand breaks (DSBs) (Roy et al., 2012).

DSBs are considered to be the most threatening form of DNA damage, as the integrity of both strands of the DNA chromosome are compromised simultaneously. DSBs occur mainly during DNA replication, but also following exposition to ionizing radiation and genotoxic compounds. In mammalian cells, DSBs are repaired by HR (which is mostly error-free), or by non-homologous end-joining (NHEJ; which is error-prone). The genome is particularly susceptible to DNA damage during replication because damage on a single strand can be converted to double-strand damage and lead to replication fork collapse. In the absence of an intact HR pathway, these replication associated DSBs can result in chromosome rearrangements and hence, genomic instability (Roy et al., 2012).

HR repairs DSBs during the S and G2 phases of the cell cycle, when an intact sister chromatid can serve as a template for repair. The protection of the genome by HR involves damage recognition by the kinases ATM and ATR, signal mediation by CHEK2 and BRCA1, and initiation of repair by the effectors BRCA2 and RAD51. There are also several facilitators of the HR pathway, such as PALB2 and BRIP1. Each of these are predisposing genes for HBOC.

In the following, I describe the main molecular features of BRCA1 and BRCA2 proteins associated to their function, focusing on their domain structure and their ability to interact with multiple molecular partners that lead to their participation in different biological processes.

BRCA1

BRCA1 (Figure 1.5) is a gene located on chromosome 17q21.3, encompassing genomic positions 43,044,295 to 43,125,483 on GRCh38.p12. It is structured in 23 exons and encodes for a multi-domain protein of 1863 amino acids.

BRCA1 is a versatile protein that through its various functional domains, interacts with numerous proteins including DNA damage sensors, DNA repair

proteins and cell cycle regulators, carrying out diverse roles in DNA repair pathways (particularly, in HR and NHEJ) and checkpoint regulation (Nielsen et al., 2016; Yarden, Pardo-Reoyo, Sgagias, Cowan, & Brody, 2002). It has also been reported to function in transcriptional regulation and control centrosomal microtubule nucleation (Mullan, Quinn, & Harkin, 2006; Sankaran, Crone, Palazzo, & Parvin, 2007).

The domain structure of BRCA1 is very rich (Roy et al., 2012). In the N-terminal region, it has a zinc finger domain of type RING finger with an E3 ubiquitin ligase activity which catalyses protein ubiquitylation. In the C-terminal region, BRCA1 has a coiled-coil motif and then, a BRCT domain composed by 100 amino acids in tandem repeat, that acts as a phospho-protein binding domain (Figure 1.5). These domains underly different protein interactions that enable the multiple functions of BRCA1, as explained below.

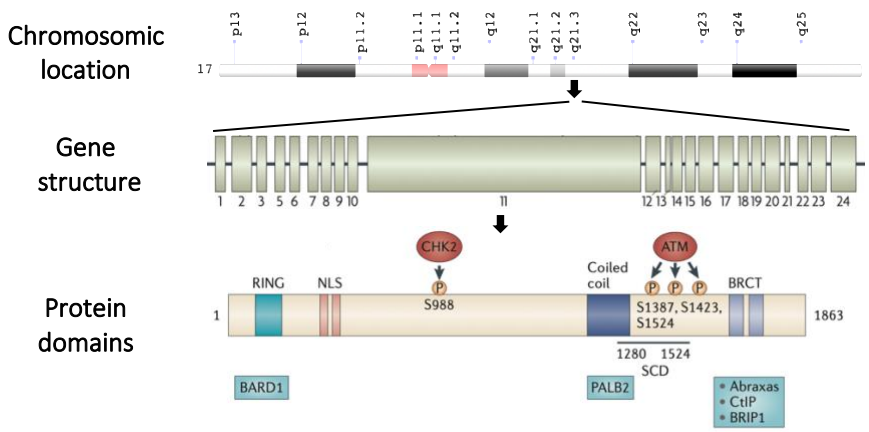


Figure 1.5 ***BRCA1 chromosomic location, gene structure and protein domains.***

Adapted from Fackenthal et al. (Fackenthal & Olopade, 2007; Roy et al., 2012).

The interaction of BRCA1 with BARD1 through their RING domains, enhances the E3 ubiquitin ligase activity of BRCA1. The BRCA1-BARD1 complex generates polyubiquitin chains at unconventional K6 that do not signal for protein degradation but rather mediate downstream signalling. BRCA1

ubiquitinates CtIP protein which is involved in the end resection of DSBs along with the MRN complex (Figure 1.6) (Roy et al., 2012).

Additionally, BRCA1–BARD1 complex is involved in the activation of G1/S, S-phase and G2/M checkpoints; BRCA1–BRIP1–TOPBP1 in the activation of S-phase checkpoint in response to stalled replication forks; and BRCA1–abraxas–RAP80 in the G2/M checkpoint in response to DNA damaged by ionizing radiation (Figure 1.4) (Roy et al., 2012).

In the C-terminal, the BRCT domain associates with proteins phosphorylated by ATM such as abraxas, BRIP1 and CtIP. These complexes carry out several functions in the DNA damage response: recruitment to DNA damage sites, DNA end resection and DNA repair during replication (Figure 1.6). Finally, the coiled-coil domain associates with PALB2 protein for the repair of DSB by HR (Figure 1.6) (Roy et al., 2012).

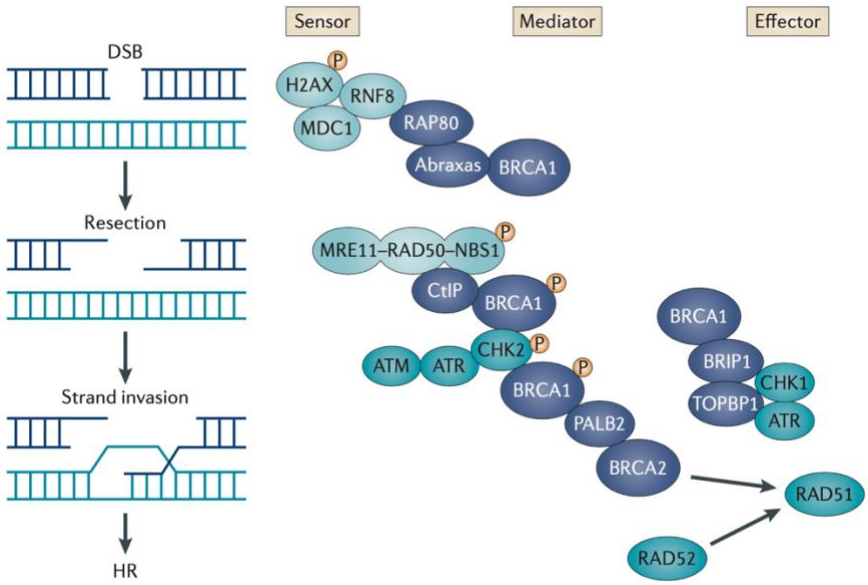


Figure 1.6 ***Repair of DSBs by HR pathway.*** *In response to DSBs, sensors detect DNA damage and signalling mediators recruit and activate effectors that repair the damage. Adapted from Roy et al. (Roy et al., 2012).*

The interacting ability of BRCA1 is key to its contribution to the HR pathway. BRCA1 is recruited to DSBs through its association with the complex abraxas–RAP80, which associates with ubiquitinated histones at DSBs. Next, BRCA1 is involved in processing DSBs through its interaction with CtIP (also known as RBBP8) and the MRN complex composed by MRE11, RAD50 and NBS1 proteins. The BRCA1–CtIP complex promotes CtIP-mediated 5'-end resection of DSBs. Afterwards, BRCA1 is also required for RAD51 recruitment to the sites of DNA damage through its interactions with PALB2 and BRCA2, which appears to be mediated by CHK2 phosphorylation on BRCA1 (Figure 1.6).

BRCA2

BRCA2 (Figure 1.7) is a gene located on chromosome 13q13, encompassing genomic positions 32,315,480 to 32,399,672 on GRCh38.p12. It is structured in 27 exons that encode for a protein of 3418 amino acids. This protein is composed of eight BRC repeats that recruit RAD51 at sites of DNA damage sites and a DNA-binding domain (DBD) that binds single-stranded (ssDNA) and double-stranded DNA (dsDNA). The DBD contains an α-helical domain, three oligonucleotide binding (OB) folds that are ssDNA-binding modules, and a tower domain (T) that protrudes from OB2 and binds dsDNA. Finally, there is a nuclear localization sequence (NLS).

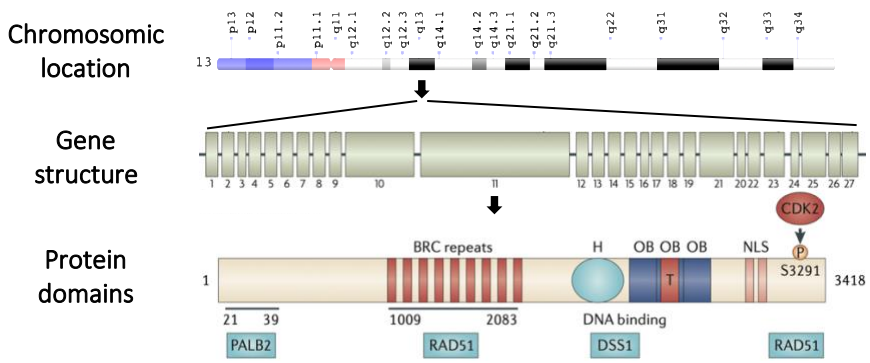


Figure 1.7 ***BRCA2 chromosome location, gene structure and protein domains***.
Adapted from Fackental et al. (Fackenthal & Olopade, 2007; Roy et al., 2012).

BRCA2 main function is to repair double-strand DNA breaks by HR. BRCA2 mediates the recruitment of the recombinase RAD51 to DSBs through its BRC repeats. Afterwards, BRCA2 mediates RAD51 filament formation at the appropriate sites of ssDNA and prevents it from binding to dsDNA. In addition, BRC repeats accelerate the displacement of protein RPA from ssDNA by RAD51, block RAD51 nucleation at dsDNA and facilitate RAD51 filament formation on ssDNA by maintaining the active ATP-bound form of RAD51 on ssDNA (Figure 1.6) (Roy et al., 2012).

1.1.3. Variant landscape of BRCA1 and BRCA2

Many variants have been described in BRCA1 and BRCA2, due to their large size and their relationship to HBOC. In fact, it is estimated that in the general population, the prevalence of BRCA1/2 pathogenic variants is between 0.1-0.3% for BRCA1 (1:200), and 0.1-0.7% for BRCA2 (1:400) (Ponder et al., 2000), but varies across ethnic groups and geographical areas, with much higher frequencies in certain founder populations, such as the Ashkenazi Jewish with a prevalence of 1:40 (King et al., 2003).

In a worldwide study carried out by the CIMBA consortium (Rebbeck et al., 2018) involving 18435 BRCA1 and 11351 BRCA2 families, it was found that the most common pathogenic variants are c.68_69del and c.5266dup for BRCA1, and c.5294del for BRCA2, accounting for 33% and 19% of all BRCA1 and BRCA2 variants respectively. In the same study, authors reported that the majority of pathogenic variants were frameshift followed by nonsense (Rebbeck et al., 2018).

Founder variants have been described in almost every population studied. The best known are in the Ashkenazi Jewish population, with 3% of individuals carrying one of the three founder variants: BRCA1 c.68_69del (1%), BRCA1 c.5266dup (0.13%), or BRCA2 c.5946del (1.52%) (Oddoux et al.,

1996). Other examples are BRCA2 c.771_775del in Iceland; BRCA1 c.4327C>T and BRCA2 c.8537_8538del in French Canada; and BRCA1 c.181T>G and c.4034del in Central-Eastern Europe (Rebbeck et al., 2018).

In the Spanish population, recurrent pathogenic variants include BRCA1 c.187_188del, c.330A>G, c.5236G>A, c.5242C>A and c.589_590del; and BRCA2 c.3036_3039del, c.6857_6858del, c.9254_9258del, and c.9538_9539del. BRCA1 c.330A4G has a Galician origin and BRCA2 c.6857_6858del and c.9254_9258del probably originated in Catalonia (Díez et al., 2003).

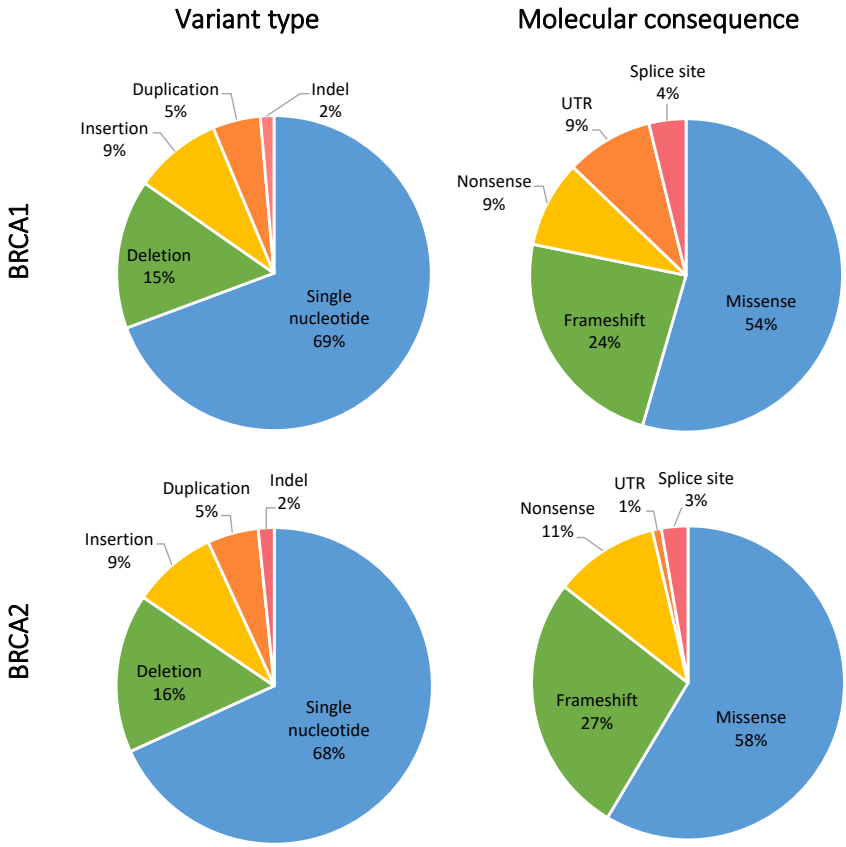


Figure 1.8 ***Distribution of BRCA1 and BRCA2 variants reported in ClinVar according to variant type and molecular consequence***. *Data obtained from ClinVar https://www.ncbi.nlm.nih.gov/clinvar/ ascertain by May 2020.*

A general overview of BRCA1/2 variants can be obtained from the data stored in ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), a large database relating human variation and phenotype that has >11000 and ~12000 variants for BRCA1 and BRCA2, respectively (Figure 1.8). Most of them are single nucleotide variants, followed by deletion and insertion variants, and, in a smaller proportion, duplications and indels. Regarding their molecular consequence, the majority are missense variants, followed by frameshift, nonsense, UTR and splice site variants (Figure 1.8).

Frameshift and nonsense variants mainly lead to premature truncation and loss of function, consistent with the tumour suppressor model. UTR variants may affect gene expression, reducing protein levels and causing loss of function. Splicing site variants commonly result in aberrant proteins unable to carry out their function properly. Large genomic rearrangements, which are more prevalent in BRCA1 than in BRCA2 due to its large number of Alu repeats, usually have devastating consequences (Judkins et al., 2012).

Missense variants, which are the focus of this thesis, tend to happen at some specific locations in the domain structure of the protein. For example, missense high-risk variants of BRCA1 are located primarily in the RING finger and BRCT domains, which are critical for the DNA repair activity of BRCA1. In BRCA2, highly penetrant pathogenic missense variants are located predominantly in the DNA binding domain (Castilla et al., 1994; Guidugli et al., 2014). It seems as if, a priori, we could use structural location of the variants to identify high-risk variants. However, this is not quite the case. In fact, we also find a substantial number of neutral variants in BRCA1/2, that is, variants that have no detectable impact on molecular function. These variants localize along the protein and can be found in the catalytic domains of BRCA1/2, making very difficult distinguishing pathogenic from neutral variants on the basis of structural location uniquely (Figure 1.9).

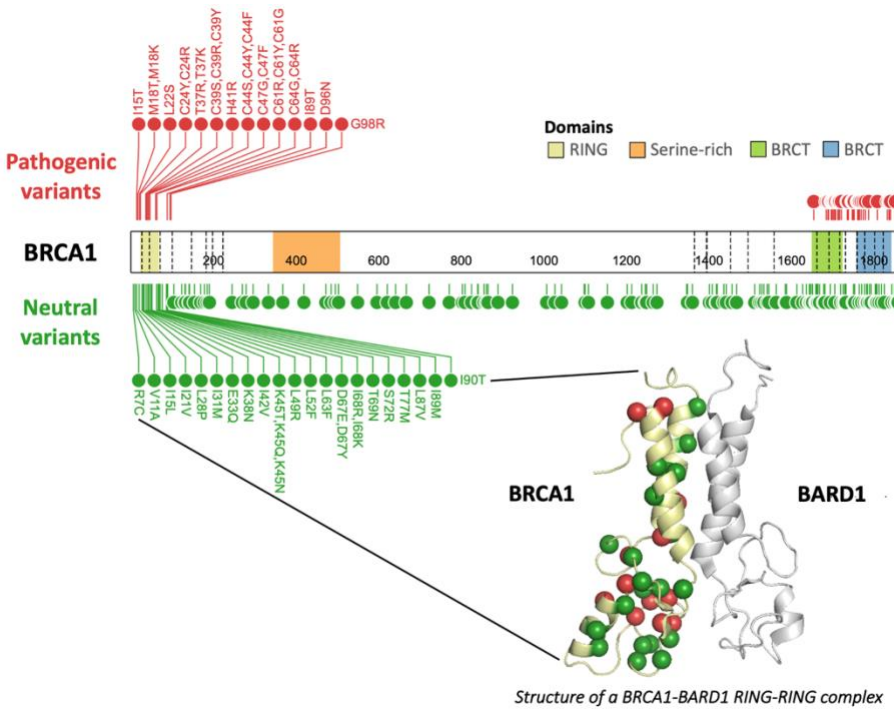


Figure 1.9 ***Distribution of pathogenic and neutral missense variants along the protein sequence of BRCA1 and in the three-dimensional structure of the BRCA1-BARD1 heterodimer complex****. Variants were obtained from Padilla et al. (Padilla et al., 2019) and coloured in red (pathogenic) and green (neutral). The three-dimensional structure of the BRCA1-BARD1 heterodimer complex was obtained from Brzovic et al. (Brzovic, Rajagopal, Hoyt, King, & Klevit, 2001). The RING domain of BRCA1 is coloured in yellow and the BARD1's RING domain in grey.*

To discriminate between both variant types, we can use segregation analysis of variants between the cancer affected members of a family. However, this is not feasible for many missense variants (Toland & Andreassen, 2017). Frequently, missense variants lack sufficient familial data to determine their pathogenicity and thereby, end up classified as VUS, delaying the access of their carriers to preventive and target therapies. When this happens, two additional approaches can provide more information: functional assays and *in silico* pathogenicity predictors.

1.1.4. Functional assays for BRCA1 and BRCA2

Functional assays experimentally measure the impact of missense variants on a function of the protein (Starita et al., 2015). They provide valuable information for the risk assessment of VUS, especially when other sources of information are not available.

For BRCA1, several functional assays are a priori available to healthcare professionals. For example, variants within the RING and BRCT domains have been characterized through assays based on rescue proliferative defects, transcription activation, ubiquitin ligase activity, measure of HR activity, resistance to DNA damage, protein-protein interaction or sensitivity to PARP inhibition or platinum drugs (Bouwman et al., 2013; Ransburgh, Chiba, Ishioka, Toland, & Parvin, 2011).

For BRCA2, variants within the N-terminal PALB2-binding domain and the C-terminal DBD domain have been characterized by functional assays measuring the HR activity, resistance to DNA damage or BRCA2-dependent assembly of RAD51 foci, protein-protein interaction or centrosome amplification (Biswas et al., 2012; Guidugli et al., 2013; K. Wu et al., 2005).

Up to date, the assay that best correlates with the tumour suppression function of BRCA1 and BRCA2 is the Homology-Directed Repair (HDR) assay. The HDR assay measures the capacity of a mutated BRCA1/2 to repair an induced DSB by means of the HDR mechanism. The most common form of HDR is the HR pathway (Roy et al., 2012), which involves the RING domain in BRCA1 and the DBD domain in BRCA2 (section 1.1.2). The HDR assay is a rescue assay performed on BRCA1/2 deficient cells with a green fluorescent protein (GFP) where the DSB is induced. Complementation of the deficient cells by the mutated BRCA1/2 cDNA expression repairs the DSB and reconstitutes the GFP, resulting in the recovery of the fluorescence.

Quantification of the proportion of GFP positive cells by flow cytometry, gives a measure of the HR rescue activity of the variant under study (Guidugli et al., 2013; Starita et al., 2015; K. Wu et al., 2005).

Although all these assays provide valuable information about the impact of variants, they are technically demanding, labour-intensive and time-consuming (Starita et al., 2017), making their current use limited to research.

Moreover, functional assays have some limitations when used to assess the pathogenicity of a variant. The first caveat is that they focus on one aspect of the protein's function, usually associated with a single domain. Therefore, for some variants, a functional assay may not cover them. Moreover, the multifunctional nature of BRCA1/2 makes necessary a combination of assays to fully characterize the impact of a variant in the different functions of the protein (Toland & Andreassen, 2017).

From these considerations, we see that to approach the variant interpretation problem from other directions may be beneficial. In this context, a promising option corresponds to the use of *in silico* pathogenicity predictors, which we present in the following section and to which an important part of this thesis is devoted.

1.2. An *in silico* approach for the variant interpretation problem

Massive application of NGS in routine clinical diagnosis has unveiled an important problem: our inability to establish the clinical significance of the variants identified by this technique. This is what we know as the variant interpretation problem. As we can see in the ClinVar database, the number of VUS is very high, representing more than 30% and 38% of the variants deposited for BRCA1 and BRCA2 (Figure 1.10), respectively.

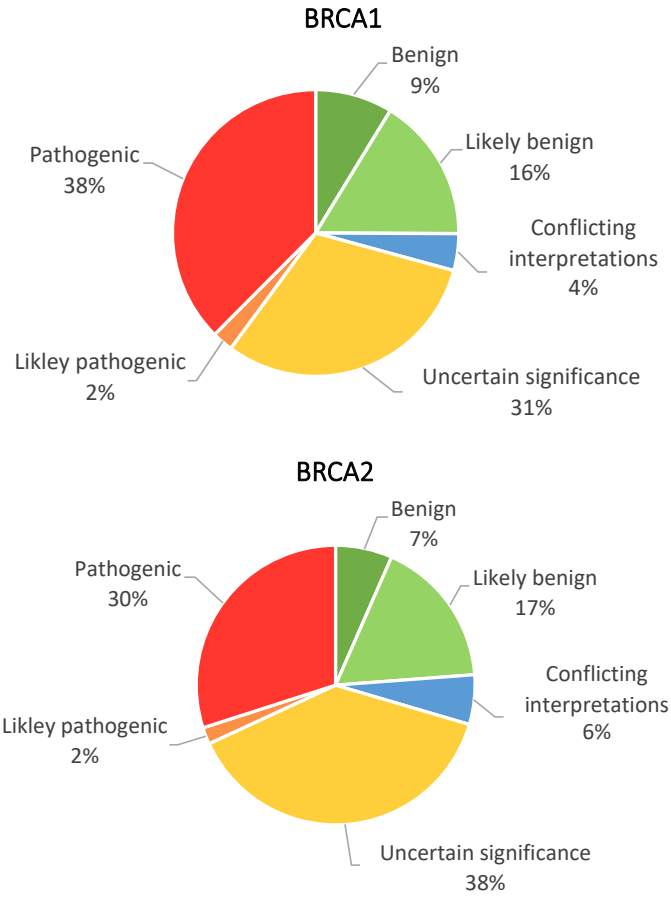


Figure 1.10 ***Clinical significance of BRCA1 and BRCA2 variants***. *Data obtained from ClinVar https://www.ncbi.nlm.nih.gov/clinvar/ ascertain by May 2020.*

A promising approach for the variant interpretation problem is the use of *in silico* tools to estimate the variants' impact, using what we known as pathogenicity predictors (Shendure, Findlay, & Snyder, 2019). However, this approach is far from easy.

Why is this so? Why is it so difficult to establish the relationship between a genetic variant and its clinical phenotype? The answer is simple: because it is a deep problem whose solution requires to address several scientifically hard questions related with the functional impact of variants. Furthermore, the issues addressed by these questions vary depending on the type of variant we are considering, e.g., single-nucleotide variants in the coding region, small insertions, large deletions in the non-coding regions of the genome, inversions, translocations, etc.

In this thesis, we will focus on missense variants, since single amino acid replacements ranks as one of the first causes of HBOC (Figure 1.7) (Dines et al., 2020) and the scientific knowledge behind them has reached an important level of maturity. In the next section, I describe the main aspects of the *in silico* approach to the pathogenicity prediction of missense variants.

1.2.1. Prediction of variant pathogenicity

Pathogenicity predictors are *in silico* tools that aim to predict the functional impact of variants using supervised algorithms. Supervised algorithms constitute a family of machine learning techniques designed to address both regression and classification problems, by learning from a set of examples. They are typically trained to discriminate between two classes of objects, for example, in the case of missense variants, they are trained to distinguish pathogenic from neutral ones, using a set of features derived from molecular biology, biophysics and biochemistry.

The development of a pathogenicity predictor follows four standard steps (Figure 1.11) (Riera, Lois, & De la Cruz, 2014). First, the collection of a group of pathogenic and neutral variants to train the predictor. Second, the selection of a set of features discriminant between these types of variants. Third, the training of a supervised algorithm with the variants and their discriminant features. Fourth, the estimation of the performance of the model, based on an independent set of variants. In the next section (1.2.2), we will focus on the selection of the discriminant features, which respond to our scientific view of the problem; and in the following section (1.2.3), I describe the remaining steps, which are of a more technical nature and are related to the actual construction of the predictor.

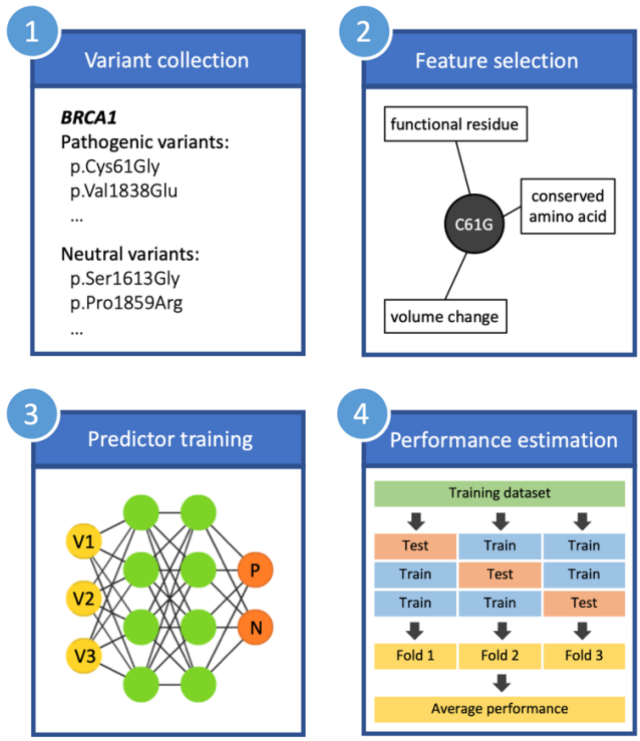


Figure 1.11 ***Development of a pathogenicity predictor***. *It follows four steps: (1) collection of a group of pathogenic and neutral variants, (2) selection of a set of discriminant features, (3) training of the predictor, and (4) estimation of the performance of the model.*

1.2.2. Characterizing the functional impact of missense variants

The working hypothesis behind these predictors is that an important part of the variant's pathogenicity depends on the molecular impact they have on protein function, structure and/or stability.

For this reason, the approach to predict the pathogenicity of these variants is based on features that reflect this molecular impact. They are divided into four groups (Riera et al., 2014): (i) features related to the functional residues of the protein, (ii) features strictly depending on the amino acid replacement, (iii) features related to the change in protein stability upon mutation, and (iv) features measuring the disruption in the conservation pattern of the multiple sequence alignment (MSA) of the protein family (S. R. Sunyaev, 2012).

Features reflecting the impact of the variants in functional residues

These features take into account the functional residues, which are generally located in the surface of the protein and carry out several functions such as substrate binding sites; catalytic sites where chemical reactions occur; post-translational modification sites that undergo phosphorylation, glycosylation and other covalent modifications; and protein-protein and protein-DNA interactions that allow a variety of functions like signal transduction, membrane transport or transcription regulation (Fernández-Recio, 2011).

Features reflecting the differences between the native and mutant amino acids

These are descriptors of the changes in hydrophobicity, volume, charge, etc., resulting from the amino acid replacement. Usually, large values indicate an important molecular impact, whereas small changes are better tolerated by the protein. Interestingly, these properties are summarized by substitution

matrices like Blosum62 (Henikoff & Henikoff, 1992), whose values correspond to disruptive changes when negative and to conservative when positive.

Features measuring the impact of variants on protein stability

Protein stability (ΔΔG) is a thermodynamic property that measures the separation between the native state of the protein and other non-native competing states. Several studies show how this fundamental property depends on the nature of the variant, and how pathogenic variants behave differently from neutral variants (Carles Ferrer-Costa, Orozco, & de la Cruz, 2002; Guerois, Nielsen, & Serrano, 2002; Riera et al., 2014). The impact on protein stability is intimately related to the 3D structure of the protein, e.g. the loss of native atomic interactions (e.g. hydrogen bonds, salt bridges, etc.), the secondary structure where the native residue is located, etc.

Features characterizing the impact of the variant on the conservation pattern of the protein family as represented in the MSA

A MSA is an alignment of several proteins sequences that usually share a common ancestor (J. D. Thompson, Higgins, & Gibson, 1994). Different conservation measures are currently used in pathogenicity prediction. Here we will comment on two of them that have been broadly utilized by our group: Shannon's entropy and Position-Specific Scoring Matrix (PSSM) (Riera, Padilla, & de la Cruz, 2016). Shannon's entropy (Cover & Thomas, 2006) is used to estimate the compositional diversity at the location of the variant in the MSA of the protein family. It is equal to $-\Sigma_i p_i.\log_2(p_i)$, where the index *i* runs over all the amino acids at the variant's MSA column. Low values of entropy are characteristic of highly conserved amino acid among species and suggest a low tolerance for change. On the contrary, high values of entropy indicate point to better tolerance for a change. PSSM measures the frequency

of the native amino acid at the variant location, normalized by the frequency of the amino acid in the whole MSA. It is equal to $\log_2(f_{nat,i}/f_{nat,MSA})$, where $f_{nat,i}$ is the frequency of the native amino acid at the locus *i* of the variant and $f_{nat,MSA}$ is the frequency of the same amino acid in the whole MSA. High values of PSSM indicate disruptive changes, whereas low values indicate more tolerable changes. A virtue of sequence conservation-based features is that they rely only on sequence information and can be applied to proteins for which we lack structural information.

1.2.3. Building a pathogenicity predictor

In this section, I will describe the three technical steps followed to build a predictor: the construction of a dataset of variants', the training of a predictor and the estimation of the performance of the predictor.

Construction of the variant dataset

This dataset must reflect the types of variants we aim to predict: neutral and pathogenic; and the problem we want to solve: e.g. if we want to classify BRCA1 variants, we will collect variants from this protein; if we want to obtain a general predictor, we will gather variants from different proteins, etc.

Pathogenic variants are obtained from databases such as UniProt/SwissProt (Bateman et al., 2017a), HGMD (Stenson et al., 2012), or ClinVar (Landrum et al., 2016), which are periodically updated and manually curated. It has to be noted however, that care must be exercised when using these sources, since they utilize different variant annotation and curation protocols, and in some cases the pathogenicity annotations may be incorrect (MacArthur et al., 2014), leading to contradictions between databases.

Neutral variants can be retrieved from projects that aim to sequence natural variation, such as the 1000 Genomes Project (Altshuler et al., 2010), ExAC or gnomAD (Lek et al., 2016). However, for some genes, the number of variants obtained this way may not be enough to train a predictor. In these cases, neutral variants can also be retrieved from protein sequence divergence data, that is, sequence differences between human proteins and close homologs (C. Ferrer-Costa, Orozco, & De La Cruz, 2004; S. Sunyaev et al., 2001). A comparative study carried by Wei et al. (Wei & Dunbrack, 2013) shows that both models give comparable results in the training of pathogenicity predictors.

Training the predictor

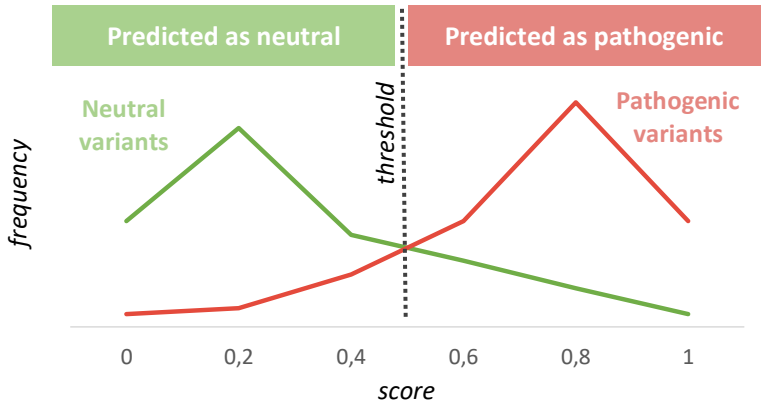
First, we need to select one of the available supervised machine learning algorithms, such as Neural Networks, Support Vector Machine, Random Forest, etc. Beforehand, there is not a better algorithm than another. So, for each problem, it has to be studied which is the most suitable. However, when several algorithms fit to the problem, the most interpretable and simplest should be favoured to prevent overfitting problems (Rudin, 2019). Overfitting problems occur in many cases when the size of the training dataset is too small or when the composition of pathogenic and neutral variants is very imbalanced (P. Baldi & Brunak, 2001). Then, the algorithm memorizes so well the data that it learns its noise and hence, fails to predict new data.

Once the algorithm and parameters are chosen, we can train the predictor with the collection of variants and their discriminant features. Afterwards, the predictor is ready to predict new variants. Usually, it provides a continuous numerical score (typically comprised between 0 and 1) and using a decision threshold, discretizes it between pathogenic and neutral variant classes (Figure 1.12a).

Estimating the performance of the predictor

Before the predictor is delivered to the biomedical/clinical community, we must estimate its predictive performance. This is relevant to determine the suitability of the tool for specific applications, which may have very concrete quality requirements. There are different metrics to measure the performance of a pathogenicity predictor, which reflect the predictor's success in solving different aspects of the binary classification problem.

a) Distribution of training variants according to their score



b) Confusion matrix

		Predicted: Pathogenic	Predicted: Neutral
Observed	Pathogenic	True Positive	False Negative
	Neutral	False Positive	True Negative

Figure 1.12 ***Outcome and performance of pathogenicity predictors****. a) Distribution of training pathogenic and neutral variants according to the predicted score. Most of the pathogenicity predictors present their score as a continuous value between 0 and 1, which is discretized by means of a decision threshold. b) The result of comparing observed and predicted values can be summarized in a confusion matrix which its values are at the basis of most performance descriptors.*

Most of the performance measures of a predictor are generally obtained from a confusion matrix (Figure 1.12b), where the successes and failures of the method are summarized. Successes correspond to True Positive (TP) and True Negative (TN) amounts, which are the numbers of pathogenic and neutral variants correctly predicted, respectively. Misclassification errors are represented by False Positive (FP) and False Negative (FN) amounts, which are the numbers of neutral variants predicted as pathogenic and vice versa, respectively. As mentioned before, these four numbers (TP, TN, FP, FN) are the basis of most performance metrics for binary classifiers (Vihinen, 2012). In this thesis we will use the sensitivity, specificity, positive predictive value, negative predictive value, accuracy and Matthews correlation coefficient metrics. Since they are broadly used, I will only briefly describe them.

Sensitivity (also known as True Positive Rate (TPR) or recall) and specificity (also known as True Negative Rate (TNR)) focus on complementary aspects of the predictive performance. Sensitivity measures the proportion of observed positive cases which are correctly predicted as positive, whereas specificity measures the proportion of observed negative cases correctly predicted as negative. They are expressed as:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Positive predictive (PPV) and negative predictive (NPV) values. PPV measures the proportion of predicted positive cases which are actually positive cases, and NPV measures the proportion of predicted negative cases which are actually negative cases. They are expressed as:

$$\text{PPV} = \frac{TP}{TP + FP}$$

$$\text{NPV} = \frac{TN}{TN + FN}$$

Although these metrics are very valuable, they are also sensitive to the dataset composition, making difficult their use for comparing the work of authors working with different datasets.

Finally, we have a couple of performance metrics that describe the success rate of a predictor for both classes simultaneously: Accuracy and Mathews Correlation Coefficient.

Accuracy corresponds to the overall fraction of successful predictions. It gives a general view of the performance, however, when there is a class imbalance in the mutation dataset, that is, when one class is more frequent than the other, accuracy can be misleading (P. Baldi, Brunak, Chauvin, Andersen, & Nielsen, 2000). It is expressed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Mathews Correlation Coefficient (MCC) is another performance measure that is highly cited in the literature (Vihinen, 2012). It is a correlation coefficient, with values comprised between -1 and 1. These two extremes reflect a complete disagreement and agreement in the predictions, respectively; and 0 corresponds to a random predictor (P. Baldi et al., 2000). MCC is considered more informative than the previous measures since it takes into account the four primary quantities in a balanced way (Chicco, 2017):

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

As we have seen, each of these descriptors has its virtues and defects and, when presenting a new predictor, it is recommended to use several of them to describe completely its success rate (P. Baldi et al., 2000; Vihinen, 2012).

Once a predictor has been developed and calibrated it is ready for its use by the scientific community. Many predictors are available nowadays (Ghosh, Oak, & Plon, 2017; Niroula & Vihinen, 2016) and their continued use has unveiled some problems with the *in silico* approach that is worth mentioning. For example, in some cases predictors have imbalanced sensitivities and specificities (Ernst et al., 2018) and their performances vary between genes (Riera et al., 2016). These issues impede the stand-alone use of pathogenicity predictors. However, their potential has been recognized and are currently employed in the clinical setting, although always in combination with other sources of biomedical evidence (Richards et al., 2015).

1.2.4. State-of-the-art trends in pathogenicity prediction

In spite of the huge amount of simplifications involved, pathogenicity predictors work surprisingly well. Indeed, present day predictors have an average predictive power over 85% (Riera et al., 2014), indicating that structure- and conservation-based parameters, capture some essential aspects of the molecular impact of mutations.

Since the first predictors were developed in the early 2000s, the strategies to develop pathogenicity predictors have gradually changed. Initially, general predictors aiming to predict all variants from any protein were developed. More recently, after noticing that predictors had a different performance depending on the gene (Figure 1.13) (Riera et al., 2016), a part of the development efforts have shifted towards the obtention of protein specific tools.

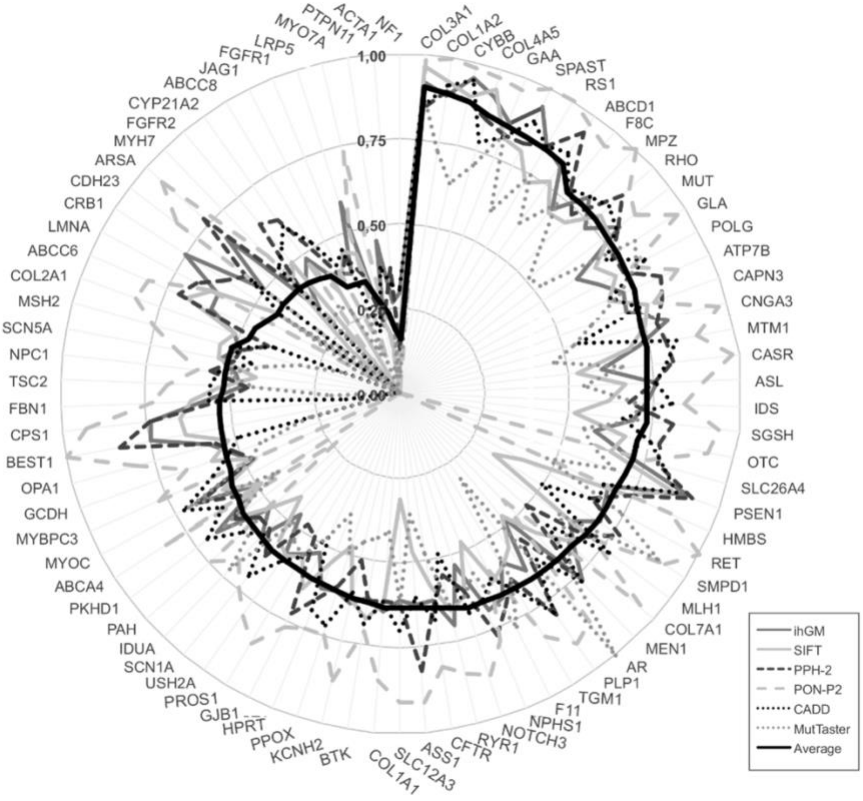


Figure 1.13 ***Performance of general predictors among several proteins.*** *Adapted from Riera et al. (Riera et al., 2016).*

Protein specific predictors are produced in different ways, from *de novo* development (Riera et al., 2016) to adapting pre-existing predictors, by generating specific decision cutoffs for the protein of interest (Itan et al., 2016). Comparative studies support the idea that these protein specific tools can compete in performance with general predictors (Riera et al., 2016).

Interestingly, and in parallel with the development of specific predictors, these recent years have witnessed the birth of a new generation of methods. These new predictors, referred to as metapredictors or ensemble predictors (Vihinen, 2014), are conceptually different from fundamental predictors. Rather than using pre-existing scientific knowledge on the molecular impact of variants, they combine the scores of pre-existing methods to generate

their predictions. Metapredictors have good performances, but they may also have some unwanted biases towards the predictions of certain tools frequently used to build them, such as SIFT or PolyPhen2 (Vihinen, 2014).

Finally, one of the most promising trends in the development of pathogenicity predictors has been the proposal of Masica et al. (David L. Masica & Karchin, 2016). These authors suggest to focus on intermediate phenotypes, known as endophenotypes, rather than on the final, clinical phenotype. Endophenotypes are quantitative measures of clinical relevance, like catalytic activities or others. They are closer to the genotype than clinical phenotypes and, thereby, less influenced by the genetic background and environment, which may result in predictors with higher success.

The main goal of this thesis is to combine two of these strategies, the development of protein-specific tools and the close genotype-endophenotype relationship, to build a protein specific prediction tool for BRCA1 and another for BRCA2, aimed at predicting the values of the HDR assay for these proteins.

2. Objectives

Objectives

The aim of this thesis is to advance the field of pathogenicity predictors, by means of understanding the impact of missense variants on the function of BRCA1 and BRCA2 proteins, and how we can predict it.

To accomplish this goal, the thesis addresses the following objectives:

1. Model the impact of missense variants in the HDR function of BRCA1 and BRCA2 proteins and construction of a pathogenicity predictor using molecular properties related to sequence conservation and amino acid replacement.
2. Disseminate the knowledge and methodology of the protein-specific pathogenic predictors for BRCA1 and BRCA2 among the scientific and clinical community by building a user-friendly website.
3. Understanding the computational information of pathogenicity predictors in a practical case of clinical research: characterization of a novel pediatric neurologic disorder caused by variants in histone H3.3.

3. Building a protein specific pathogenicity predictor for BRCA1 and BRCA2

The results presented in this chapter have been published in Padilla et al. (Padilla et al., 2019)

3.1. Introduction

Germline variants disrupting the DNA protective role of BRCA1 and BRCA2 (BRCA1/2) result in an increased risk of developing hereditary breast and ovarian cancer (HBOC) (Roy et al., 2012; Venkitaraman, 2014). Identification of the individuals carrying these pathogenic variants is clinically relevant since it allows channeling them to surveillance, prevention programs and targeted therapies (Paluch-Shimon et al., 2016). As a result, the survival rates of these patients may be increased.

However, not all of them benefit equally, because we lack the exact knowledge of the functional impact of the majority of BRCA1/2 variants. In these cases, a straightforward decision can only be taken when the variant is overtly deleterious (insertions, deletions, and substitutions codifying truncated proteins). When the variant has an uncertain effect on protein function (e.g., missense, synonymous, intronic, and 5'UTR or 3'UTR variants) the best course of action becomes unclear.

Solving this problem is not easy since familial data is usually scarce and functional assays are technically challenging for a systematic application (Starita et al., 2015). The most widely used experiment is the homology-directed DNA repair (HDR) assay of BRCA1/2, a cell-base experiment that requires a complex rescue assay (Guidugli et al., 2013; Millot et al., 2012) to measure the impact of variants in the HR activity of BRCA1/2.

In these circumstances, *in silico* pathogenicity predictors like Align-GVGD (Tavtigian et al., 2006), PolyPhen-2 (I. Adzhubei et al., 2010), SIFT (Kumar, Henikoff, & Ng, 2009), PON-P2 (Niroula, Urolagin, & Vihinen, 2015), etc; are employed as an inexpensive, easy-to-use alternative. The predictions obtained are applied to prioritize the variants for experimental evaluation

and as a contribution to decision models that integrate different sources of evidence (Karbassi et al., 2016; Lindor et al., 2012; Moghadasi, Eccles, Devilee, Vreeswijk, & van Asperen, 2016; Vallée et al., 2016).

However, the moderate success rate of these tools is an obstacle for their extended use in the clinical environment (Riera et al., 2014). Ernst et al. (Ernst et al., 2018) after testing the performance of Align-GVGD, SIFT, PolyPhen-2, MutationTaster2 on a set of 236 BRCA1/2 variants of known effect, suggested that *in silico* results cannot be used as stand-alone evidence for diagnosis. In terms of molecular effect, two independent massive functional assays of BRCA1 variants (Findlay et al., 2018; Starita et al., 2015) show that *in silico* predictors provide only a limited view of the functional impact of these variants. In summary, there is an urge to improve the predictive power of these tools, if we want to increase their usage in the clinical setting and augment their value for healthcare stakeholders.

The slow progression in performance displayed by pathogenicity predictors along time shows that improving them is a difficult task (Riera et al., 2014). In this scenario, the use of rigorous performance estimates becomes an important factor, since improvements are expected to be small and hard to establish. Generally, these estimates are obtained using a standard N-fold cross-validation procedure (Pierre Baldi, Brunak, Chauvin, Andersen, & Nielsen, 2000; Riera et al., 2014; Vihinen, 2012).

However, given the increasing availability of variant data, independent testing of predictors is emerging as a valuable option to complement cross-validated performance estimates. Sometimes this testing is done in specific systems for which new variants with impact annotations become available, either at specific/general databases or through experimental testing of their function. For example, Riera et al. (Riera et al., 2015) cross-validate their Fabry-specific predictor with a set of 332 pathogenic and 48 neutral variants,

and provide an independent validation, using a set of 65 pathogenic variants obtained from an update of the Fabry-specific database. Wei et al. (Wei & Dunbrack, 2013) test five *in silico* predictors using an independent set of 204 variants (79 deleterious, 125 neutral) of the human cystathionine beta-synthase whose impact they establish with an *in vitro* assay. Large variant sets, including data from different genes, are also frequently used to assess and compare the performance of several predictors simultaneously (Niroula & Vihinen, 2016).

While relevant, the value of these approaches to validation is limited by different factors, such as the fact that the performance evaluation between works may vary, their dataset of variants may differ, etc. In this situation, CAGI (Critical Assessment of Genome Interpretation) (Hoskins et al., 2017), a community meeting where developers can assess the performance of their methods in specific challenges, offers an excellent opportunity to obtain an independent view on their work. For users, it allows having an idea on the state of the art of the predictors for the protein or disease of their interest.

In this chapter, I present a novel family of pathogenicity predictors for scoring BRCA1 and BRCA2 missense variants and their performance in the recently held ENIGMA challenge in CAGI 5.

The four tools described here, two for BRCA1 and two for BRCA2, are protein-specific (Crockett et al., 2012; C. Ferrer-Costa et al., 2004; Pons et al., 2016; Riera et al., 2016), that is, only variants for the given protein are used to train the predictors. These two protein specific predictors differ on their objective: one is trained to estimate the molecular impact of variants on the HR function of BRCA1/2 as measured in the HDR assay and the other is trained to estimate the clinical significance of variants, that is, whether it is pathogenic or neutral. Technically, due to our small training dataset, we employed simple

algorithms. For the first predictor, we used a standard multiple linear regression and for the second, a neural network model with no hidden layers.

Once these predictors were obtained, they were applied to the BRCA1/2 variants of the ENIGMA consortium (Spurdle et al., 2012) in the CAGI 5 experiment. This was done following a protocol that combined the pathogenicity predictions of both splicing and protein function impact of missense variants (Figure 3.1). Given a variant, it was first tested for its effect on the splicing pattern, using a recently developed approach by Moles-Fernández et al. (Moles-Fernández et al., 2018). If the variant had no detectable effect, it was subsequently tested for its impact on protein function, using the predictors here presented.

Our results show that all our protein-specific predictors can discriminate (with different degrees of success) between pathogenic and neutral variants, for both BRCA1 and BRCA2 proteins. For this binary discrimination problem, their performances are comparable to, or better than, those of general predictors (CADD, PolyPhen-2, PON-P2, PMut, SIFT). When applied to the variants of the CAGI challenge, where the goal is to classify them in one of the IARC 5-tier classes, we see the same trend, although with a decrease in performance, like the rest of predictors. Nonetheless, our methods are able to predict the biased composition of the CAGI dataset, which is enriched towards neutral variants; especially, our predictors that estimate the molecular impact of variants on the HR function. For the correct identification of the pathogenic variants, it is particularly important the prediction of the splicing impact, which enhances the final success rate.

3.2. Materials and methods

First, I describe the overall prediction protocol (Figure 3.1), which integrates predictions of splicing and protein impact; then, the development of the pathogenicity predictors for BRCA1/2 missense variants; and finally, the application of these tools in the ENIGMA challenge of the CAGI 5.

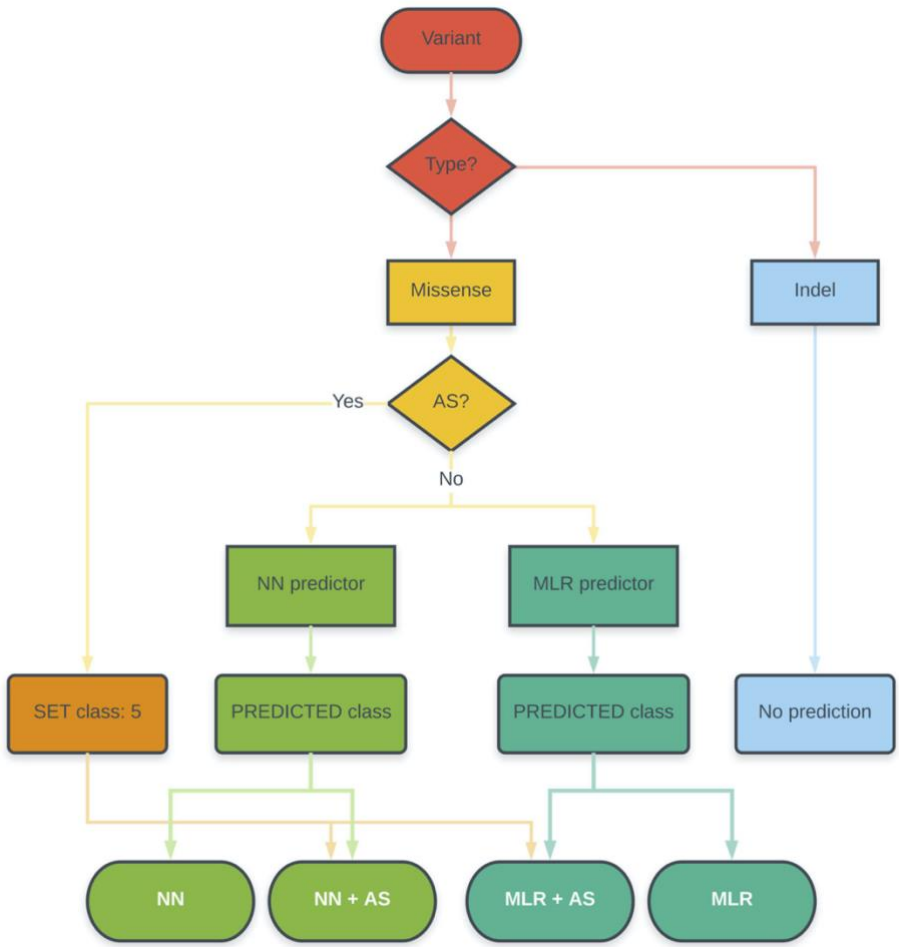


Figure 3.1 ***Protocol to assess the impact of BRCA1/2 variants on splicing and protein function used in the ENIGMA challenge of the CAGI 5 experiment.*** *MLR and NN refer to our two protein-specific predictors, based on a multiple linear regression (MLR) and a neural network (NN), respectively. AS refers to the procedure to predict variants affecting splicing (Moles-Fernández et al., 2018).*

3.2.1. Overall prediction protocol

In Figure 3.1, I describe the protocol followed in our contribution to CAGI 5, an experiment that presents several challenges revolving around a central theme (Hoskins et al., 2017): the prediction of pathogenicity of variants and its applications.

We focused our efforts on the ENIGMA challenge to predict the increased risk of breast cancer of a collection of BRCA1 and BRCA2 missense variants provided by the ENIGMA consortium (Spurdle et al., 2012). We submitted four sets of predictions per protein (Annex 8.1). These sets correspond to the different combinations of our tools to predict the effect of a variant, including the impact on splicing and protein function/structure.

To predict the variant's impact on splicing, we used the method of Moles-Fernández et al. (Moles-Fernández et al., 2018) labelled here as AS. For the prediction of the variant's impact on protein function/structure, we used our two developed methods: a multiple linear regression (MLR) model that predicts the variant's impact on the HR function of BRCA1/2, and a neural network (NN) model that predicts the clinical significance of a variant. The protocol of these four sets of predictions are the following:

1. **MLR + AS**: predicts AS impact followed by protein impact with MLR.
2. **NN + AS**: predicts AS impact followed by protein impact with NN.
3. **MLR + nAS**: predicts only protein impact with MLR, no AS is used.
4. **NN + nAS**: predicts only protein impact with NN, no AS is used.

The submission format was the same for each set of predictions and was provided by the organizers. It comprised the following information per variant: three fields for the identification of the variant: gene, DNA variant, protein variant; three fields for the prediction of pathogenicity of the variant: predicted IARC 5-tier class, probability of the variant being pathogenic (p),

confidence of each prediction probability (sd); and one field for comments. The predict class of pathogenicity corresponds to the IARC 5-tier classification system (Goldgar et al., 2008; Plon et al., 2008) which includes the classes 1=Not pathogenic, 2=Likely not pathogenic, 3=Uncertain, 4=Likely pathogenic and 5=Pathogenic.

For the sets MLR+AS and NN+AS, any missense variant predicted as pathogenic by the AS predictor was arbitrarily assigned values of p=1 and sd=0, and class 5. Otherwise, the variant was annotated using our protein impact predictors, which were obtained as explained below. That is, the protein impact was estimated only if the variant had no predicted effect on AS. One can distinguish these situations by the text in the comments column: splicing, which means that the variant is annotated with the AS predictor; protein, which means that the variant is annotated with the protein-based predictors (MLR or NN); and arbitrary, used for variants for which we do not have a predictor, since they are not missense (annotation is arbitrarily set to the following: class=5, p=0.5, sd=0.5).

For the sets MLR+nAS and NN+nAS we did not use the AS predictor. All the missense variants are annotated using our protein impact predictors (obtained as explained below). As before, these situations are distinguished in the comments field with the label protein.

3.2.2. Prediction of the variants' impact on splicing

To score the effect of variants on splicing, we used the recent method of Moles-Fernández et al. (Moles-Fernández et al., 2018). They identified the best combination of predictors available in the package Alamut Visual v2.10, for predicting splice site alterations. More precisely, they showed that the HSF+SSF-like combination (with Δ-2% and Δ-5% as thresholds, respectively) for donor sites and the SSF-like (Δ-5%) for acceptor sites, exhibited an optimal

performance in a benchmark combining RNA *in vitro* testing and a dataset of variants retrieved from public databases and reported in the literature.

For the CAGI challenge, a variant predicted to produce splice site alterations was arbitrarily assigned class 5, p=1 and sd=0; and identified as splicing in the comments column. Variants giving no signal for splice site alterations were directly channelled to the protein predictors.

3.2.3. Prediction of the variants' impact on protein function

We developed two methods for predicting the impact of missense variants in BRCA1 and BRCA2. One is trained on a neural network (NN) and produces a binary output reflecting the clinical significance of variant, that is, the high/low risk of cancer. The other method is based on a multiple linear regression (MLR) and is trained to predict the value of the HDR assay of a variant, that can be further discretized between high/low risk variants. Both methods are protein-specific: there is a version of MLR for BRCA1 and another for BRCA2, and the same for NN.

3.2.4. The NN predictor

We followed our approach to produce protein-specific predictors (Riera et al., 2016), which comprises the following steps: (i) collection of the variant set, (ii) selection of the discriminant features, and (iii) training of the predictor.

Collection of BRCA1/2 variants with known clinical significance

Missense variants with a known clinical significance were selected manually by reviewing several gene-specific databases that collect BRCA1 and BRCA2 variants along with published literature: Leiden Open Variation Database

(LOVD) describing functional studies of specific BRCA1 and BRCA2 variants (http://databases.lovd.nl), LOVD-IARC dedicated to variants that have been clinically reclassified using an integrated evaluation (http://hci-exlovd.hci.utah.edu), BRCA Share™ (formerly Universal Mutation Database UMD-BRCA mutations database http://www.umd.be/), ClinVar that provides clinical relevance of genetic variants (https://www.ncbi.nlm.nih.gov/clinvar/) and *BRCA1* CIRCOS which compiles and displays functional data on all documented *BRCA1* variants (https://research.nhgri.nih.gov/bic/circos/). Finally, each variant was validated by combining these different sources of evidence.

Variants for which the pathogenic role was attributable to splice site alterations (assessed using Alamut Visual biosoftware 2.6, from Interactive Biosoftware) were eliminated. This was done to ensure, as far as possible, that our model was trained using variants whose pathogenic/neutral nature was a consequence of their impact in protein function/structure only.

The final dataset of missense variants with annotated clinical significance was constituted by 77 pathogenic and 149 neutral variants in BRCA1; and 36 pathogenic and 105 neutral variants in BRCA2 (see Table 3.1).

Selection of discriminant features

We used a total of 6 discriminant features that we previously employed for the development of protein-specific predictors (Riera et al., 2016).

Two of these features are extracted from multiple sequence alignments (MSAs): Shannon's entropy and position-specific scoring matrix (PSSM). Shannon's entropy is equal to $-\Sigma_i p_i.\log_2(p_i)$, where the index i runs over all the amino acids at the variant's MSA column. PSSM is equal to $\log_2(f_{nat,i}/f_{nat,MSA})$, where $f_{nat,i}$ is the frequency of the native amino acid at the locus i of the variant and $f_{nat,MSA}$ is the frequency of the same amino acid in the whole MSA.

We used two different MSAs: psMSA and oMSA, which resulted in two different versions of the NN predictor. psMSA was obtained using the same protocol utilized for the protein-specific predictors (Riera et al., 2015, 2016) which, briefly, consists of three steps: (i) recovery of BRCA1/2 homologs using a query search of UniRef100; (ii) elimination of remote homologs (<40% sequence identity); (iii) alignment of the remaining sequences with Muscle (Edgar, 2004). The oMSA was obtained from the group of Sean Tavtigian (Tavtigian, Greenblatt, Lesueur, & Byrnes, 2008) at their website of Huntsman Cancer Institute (http://agvgd.hci.utah.edu/alignments.php), and comprise only orthologs of BRCA1 and BRCA2. The NN predictions submitted to CAGI were those obtained with the method developed using the psMSA, although results for the second predictor are mentioned below.

Three other features measure the change of a physicochemical property of the amino acid replacement, the difference between the native and the mutant amino acid of the Van der Waals volume (Bondi, 1964), the hydrophobicity value (estimated from water/octanol transfer free energy measurements) (Fauchere & Pliska, 1983) and Blosum62 value (Henikoff & Henikoff, 1992).

Finally, a sixth feature is a boolean (True/False) that summarizes the functional/structural role of the native residue at the protein from the UniProt database. It is set to True when the native residue has an annotated function on the database, and False otherwise.

Training the neural network predictor

The neural network model was trained using WEKA (v3.6.8) (M. Hall et al., 2009). Following our experience in the development of protein-specific predictors with small datasets (Riera et al., 2016), we employed the simplest neural network model: a single-layer perceptron. Sample imbalances in the

training set were corrected with SMOTE (Chawla, Bowyer, Hall, & Kegelmeyer, 2002).

The NN predictor provides two outputs: (i) a binary prediction of the class of the variant: pathogenic or neutral; and (ii) a continuous numerical score, comprised between 0 and 1, that reflects the probability of pathogenicity.

Finally, a Leave-one-out cross-validation (LOOCV) was carry out to assess the performance of the model also using WEKA (v3.6.8) (M. Hall et al., 2009).

Obtention of the CAGI output

As mentioned above, the CAGI submission requires three pieces of information for each variant prediction: the predicted IARC 5-tier class, the probability of pathogenicity (p) and the reliability (sd). We took as p the numerical score from the NN, which varies between 0 (minimal probability of pathogenicity) and 1 (maximal probability of pathogenicity). For the sd value, we used the following formula (C. Ferrer-Costa et al., 2004): $sd= 0.5-|0.5-p|$. It goes from 0 (maximal reliability) to 0.5 (minimal reliability). Finally, the predicted IARC 5-tier class was obtained from p, using the ENIGMA conversion table at the CAGI site (class 5: $p>0.99$; class 4: $0.95<p<0.99$; class 3: $0.05<p<0.95$; class 2: $0.001<p<0.49$; class 1: $p<0.001$).

3.2.5. The MLR predictor

This method aims to predict the impact of variants on the molecular function of BRCA1/2 in the HR pathway as measured in the HDR (homology-directed DNA repair) assay. Since the output of the HDR assay is a continuous value, we opted for using a multiple linear regression as a modeling tool, as implemented in the python package Scikit-learn (Pedregosa et al., 2011), that we also used to perform the LOOCV.

For a given variant, the output of the MLR predictor is HDR_{pred}, the predicted value of the HDR assay. In the few cases that the result was a slightly negative number, the predicted value was set to 0, since the output of the HDR experiment is always a positive number.

To train our model, we used the experimental HDR values available on the literature: 44 variants for BRCA1 (Starita et al., 2015) and 185 variants for BRCA2 (Guidugli et al., 2013, 2018). However, to reinforce the strength of the signal relative to experimental noise, we did not employ the full data sets. The training dataset was constituted by those variants used to build the NN predictor (see the previous section) for which HDR values were available. The final number of training HDR values was 28 for BRCA1 and 92 for BRCA2. The HDR values for BRCA2 corresponded to 56 unique variants since some variants had been tested twice (Guidugli et al., 2013, 2018).

Given the small size of the variant datasets and to try to minimize potential overfitting problems, we used the three most discriminative features of the previous ones: Shannon's entropy, PSSM, and Blosum62, as independent variables in the regression model (Figure 3.2). Like for the NN methods, the MSA-based features were computed with both the psMSA and the oMSA, thus, leading to two versions of the MLR. Only the predictions for the oMSA-based MLR where submitted to CAGI; however, the results for the second predictor are also provided here.

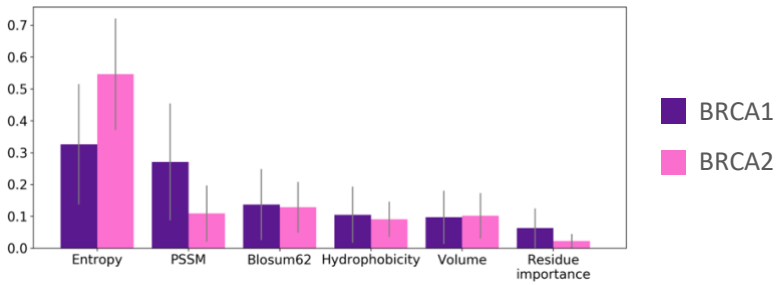


Figure 3.2 ***Feature importance of BRCA1 (violet) and BRCA2 (pink) variants.***

Calculated with the Extremely Randomized Trees Classifier from scikit-learn.

Obtention of the CAGI output

To adapt the HDR predicted value to the CAGI numerical score from 0 to 1, we us used the following steps:

1. Obtention of the HDR predicted values for the variants in the BRCA1 and BRCA2 training datasets with the protein specific MLR.
2. For each protein, compute the mean (m) and standard deviations (sd) of the predicted HDR values of the training pathogenic and neutral variants separately, obtaining four metrics per protein: m_P, sd_P, m_N, sd_N.
3. Compute CAGI's p as follows:

$$\frac{N(x; m_P, sd_P)}{N(x; m_P, sd_P) + N(x; m_N, sd_N)}$$

   where N(x; m, sd) represents a normal probability distribution of mean m and standard deviation sd. The resulting value is comprised between 0 (neutral) and 1 (pathogenic), and reflects the probability of a variant being pathogenic according to our model.
4. Obtain the sd value as for the NN methods, using the following formula (C. Ferrer-Costa et al., 2004): sd= 0.5-|0.5-p|.

3.2.6. Performance assessment

The performance was estimated using a standard LOOCV procedure (Riera et al., 2016) for both MLR and NN predictors and both BRCA1 and BRCA2 proteins. The metrics used to measure the success rate of the predictors depended on the number of classes predicted.

For instance, during the development of the predictors, the NN and MLR methods predicted only two classes: pathogenic and neutral variants; whereas in subsequent validations, including the CAGI submissions, three and five classes of variants were considered. Below, we describe the performance parameters employed in each case.

Binary performance estimation

Binary performance was measured with four commonly employed metrics for binary classifications (Pierre Baldi et al., 2000; Vihinen, 2013): sensitivity, specificity, accuracy and Matthews correlation coefficient. They are computed as follows:

- Sensitivity (SN):

$$\frac{TP}{TP + FN}$$

- Specificity (SP):

$$\frac{TN}{TN + FP}$$

- Accuracy (ACC):

$$\frac{TP + TN}{TP + FP + TN + FN}$$

- Matthews Correlation Coefficient (MCC):

$$\frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FN) \cdot (TN + FP) \cdot (TP + FP) \cdot (TN + FN)}}$$

where TP (True Positive) and FN (False Negative) are the numbers of correctly and incorrectly predicted pathological variants; TN (True Negative) and FP (False Positive) are the numbers of correctly and incorrectly predicted neutral variants, respectively.

Multiclass performance estimation

Multiclass performance was used when the predicted numerical score was discretized into five and three classes. This happened when assessing the CAGI submission, where we predicted five classes; and during the application of the MLR predictor to the recently published functional assay of BRCA1 variants (Findlay et al., 2018), where we predicted three classes.

For multiclass performance estimation, the number of metrics available is smaller than for binary (Pierre Baldi et al., 2000; Vihinen, 2013). Here, we utilized the confusion matrix, accuracy per class, overall accuracy, and multiclass MCC (Gorodkin, 2004; Jurman, Riccadonna, & Furlanello, 2012).

For a multiclass problem with M classes, the confusion matrix $C=(c_{ij})$ is an (MxM) matrix where c_{ij} is the number of times that a class i input is predicted as class j. The sum of the c_{ij} corresponds to the sample size N, which in our case is the total number of variants predicted. This matrix provides the simplest description of the performance of a predictor, its diagonal and off-diagonal elements correspond to the predictor's successes and failures, respectively. If we normalize each diagonal element by its row total ($c_{ii}/ \Sigma_j c_{ij}$, where j=1,M) we obtain the accuracy of the predictor for class i. If we add all the diagonal elements and divide the result by N ($\Sigma_i c_{ii}/N$, where i=1,M), we obtain the overall accuracy.

The multiclass MCC (Gorodkin, 2004; Jurman et al., 2012) was obtained using the implementation in the python package Scikit-learn (Pedregosa et al., 2011).

3.3. Results

Here, I describe the obtention of a novel family of pathogenicity predictors specific for BRCA1/2 proteins (MLR and NN) and their application to the unknown variants of the CAGI challenge, within a protocol that also includes AS predictions (Figure 3.1).

As shown in the section 3.2 of Materials and Methods, we considered the use of two different MSAs (psMSA and oMSA) to develop our predictors. However, I center our descriptions on the versions employed for the CAGI challenge: MLR based on oMSA and NN based on psMSA. For completeness, the performance of our methods when developed using psMSA (for MLR) and oMSA (for NN) is also provide in Table 3.3 and Figure 3.6.

3.3.1. Variant datasets

The size of the datasets of variants employed in this work is shown in Table 3.1a and the overlap between the CAGI and the remaining datasets is reported in Table 3.1b. Note that the CAGI class information on each variant was made public only after the challenge was closed.

	NN	MLR	CAGI	SGE
BRCA1	226 (P=77/N=149)	28	144	1837
BRCA2	141 (P=36/N=105)	56	174	-

Table 3.1a ***Size of the datasets of variants used in this work.***

	NN-CAGI	MLR-CAGI	MLR-SGE
BRCA1	18 (P=7/N=11)	2	28
BRCA2	5 (P=2/N=3)	4	-

Table 3.1b ***Overlap between variant datasets.***

Training datasets for the NN and MLR predictors

The number of variants in the NN training sets (BCA1: 226; BRCA2: 141) is comparable to that used for developing protein-specific predictors with the same neural network model and variant features in Riera et al. (Riera et al., 2016). The situation is quite different for the MLR training sets, which are small (BRCA1: 28; BRCA2: 56), consequently, imposing a severe limitation in the number of features that can be used in the model (see section 3.2 Material and Methods).

SGE, a validation dataset for the BRCA1 MLR predictor

This set is obtained from the results of a recently published experiment for *BRCA1* (Findlay et al., 2018). The authors functionally score a large number of single nucleotide variants (SNVs), from which we retrieved the 1837 cases corresponding to missense variants. We refer to this dataset as SGE (from Saturation Genome Editing). We used SGE to further test the performance of our BRCA1 MLR because Findlay et al. find that there is a correspondence between their functional score and the score of the HDR assay (Findlay et al., 2018).

CAGI dataset

Their size (BRCA1: 144; BRCA2: 174) is of the same magnitude as that of the NN training datasets. In Table 3.2, I provide two partitions of these datasets, corresponding to: (i) the original, 5-class ENIGMA partition; and (ii) a reduced, 3-class partition. For the latter, the Pathogenic and Likely pathogenic classes have been unified into a single Pathogenic class, and the Likely not pathogenic and Not pathogenic classes have been unified into a single Neutral class. The Uncertain class (or Unknown) has been left untouched. It must be noted the high compositional imbalance of the CAGI dataset, with

the total of classes 1 and 2 being 10 and 25 times higher than that of the remaining classes, for BRCA1 and BRCA2, respectively. In particular, the absolute numbers of variants for classes 3, 4 and 5 are so low that they can hardly lead to reliable estimates for class-dependent parameters. For example, there are only two variants of class 3 for both BRCA1 and BRCA2; two and three variants for classes 4 and 5, respectively, in BRCA2; and four and seven variants for classes 4 and 5, respectively, in BRCA1.

BRCA1					
IARC 5 class	**1**	**2**	**3**	**4**	**5**
CAGI	31	100	2	4	7
3 class		**Neutral**	**Unknown**	**Pathogenic**	
CAGI		131	2	11	
BRCA2					
IARC 5 class	**1**	**2**	**3**	**4**	**5**
CAGI	31	136	2	2	3
3 class		**Neutral**	**Unknown**	**Pathogenic**	
CAGI		167	2	5	

Table 3.2 ***Composition of the variant dataset of ENIGMA challenge in CAGI 5***

3.3.2. Predicting the functional impact of variants: the MLR predictor

We developed two MLR methods, one per each BRCA1/2 protein. The goal of these methods is to predict the impact of a given variant on the molecular function of the protein, as measured by the HDR experiment. To this end, we trained them with a set of variants with known experimental values for the HDR assay and the discriminant features chosen related to the variant's effect on protein structure, protein-protein interactions, sequence conservation, etc. (Carles Ferrer-Costa et al., 2002; Riera et al., 2014).

In Figure 3.3, we see the correlation between observed vs. predicted (LOOCV) HDR values which is statistically significant (BRCA1: 0.72, p-value=$1.5x10^{-5}$; BRCA2: 0.73, p-value=$3.3x10^{-17}$). Visual inspection reveals that the variants tend to group into two clusters, showing that MLR predictions approximately reproduce the bimodal pattern of HDR assays (Guidugli et al., 2013; Starita et al., 2015). We also show in grey color, the variants which were left outside of the training set (see Materials and Methods). These are more scattered than those forming the training set, illustrating how the filtering worked.

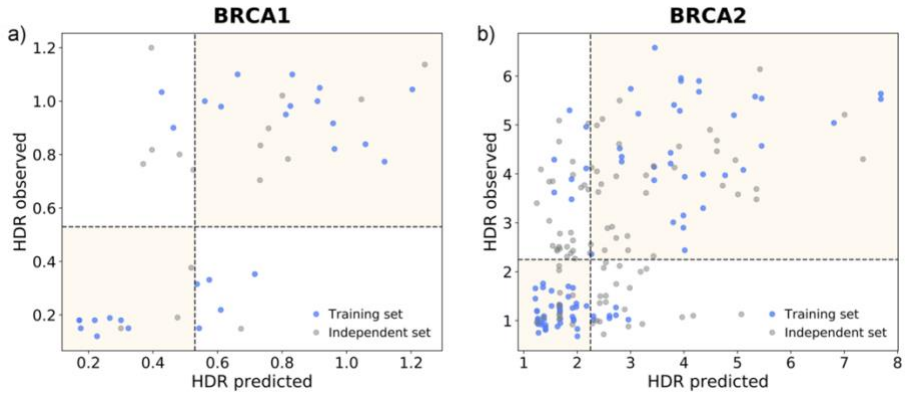


Figure 3.3 ***Observed versus predicted HDR values for BRCA1 and BRCA2****. In blue, we show the variants used for the training of our MLR method. The HDR predicted values are cross-validated (LOOCV). For completeness, we show in grey the points from the original HDR experiments that were excluded from the training process after applying our filtering procedure (see section 3.2).*

We explored how good is this level of accuracy for a standard two-class (pathogenic/neutral) prediction of the variant's pathogenicity. To this end, we discretized the predictions applying a decision boundary: a variant was called pathogenic or neutral when its predicted HDR score was below or above a given threshold, respectively. These thresholds that were taken from the experimental papers, are 0.53 for BRCA1 (Starita et al., 2015) and 2.25 for BRCA2 (Guidugli et al., 2013).

In Table 3.3 we give the parameters measuring the success rate of the discretized MLR methods. Their accuracies, 0.75 for BRCA1 and 0.86 for BRCA2, fall within the 0.79-0.99 accuracy range for protein-specific predictors (Riera et al., 2016); the same happens for the MCC, 0.50 for BRCA1 and 0.71 for BRCA2. We detect that specificity (0.85) and sensitivity (0.86) are closer for BRCA2 than for BRCA1 (SP: 0.87, SN: 0.62). Actually, for BRCA1 sensitivity tends to be small when compared to that of protein-specific predictors (Riera et al., 2016). Overall, these results indicate that the continuous HDR predictions of our MLR model can be transformed into binary predictions preserving a non-random prediction power, comparable to that of predictors trained with binary encodings (pathogenic/neutral) of the variant impact.

Protein	Method		SN	SP	ACC	MCC
BRCA1	MLR	(psMSA)	0.692	0.933	0.821	0.651
	MLR-CAGI	(oMSA)	0.615	0.867	0.75	0.502
	NN	(oMSA)	0.922	0.852	0.876	0.746
	NN-CAGI	(psMSA)	0.857	0.718	0.765	0.546
BRCA2	MLR	(psMSA)	0.828	0.741	0.786	0.571
	MLR-CAGI	(oMSA)	0.862	0.852	0.857	0.714
	NN	(oMSA)	0.75	0.867	0.837	0.592
	NN-CAGI	(psMSA)	0.75	0.771	0.766	0.473

Table 3.3 ***Two class (binary) performance of our MLR and NN predictors***

3.3.3. Validation of the BRCA1 MLR predictor with functional data

The recent publication (Findlay et al., 2018) of a massive functional assay of BRCA1 variants has given us the opportunity to check the performance of our MLR model on a set of 1837 variants. The output of this experiment is a continuous value measuring the impact of sequence variants on BRCA1 function. When we represent these values against our HDR predictions

(Figure 3.4a), we observe two clusters of points (below and above SGE=-1) that reflect the bimodal behavior of both assays, with a statistically significant rank correlation (Spearman's ρ=0.47, p-value~0).

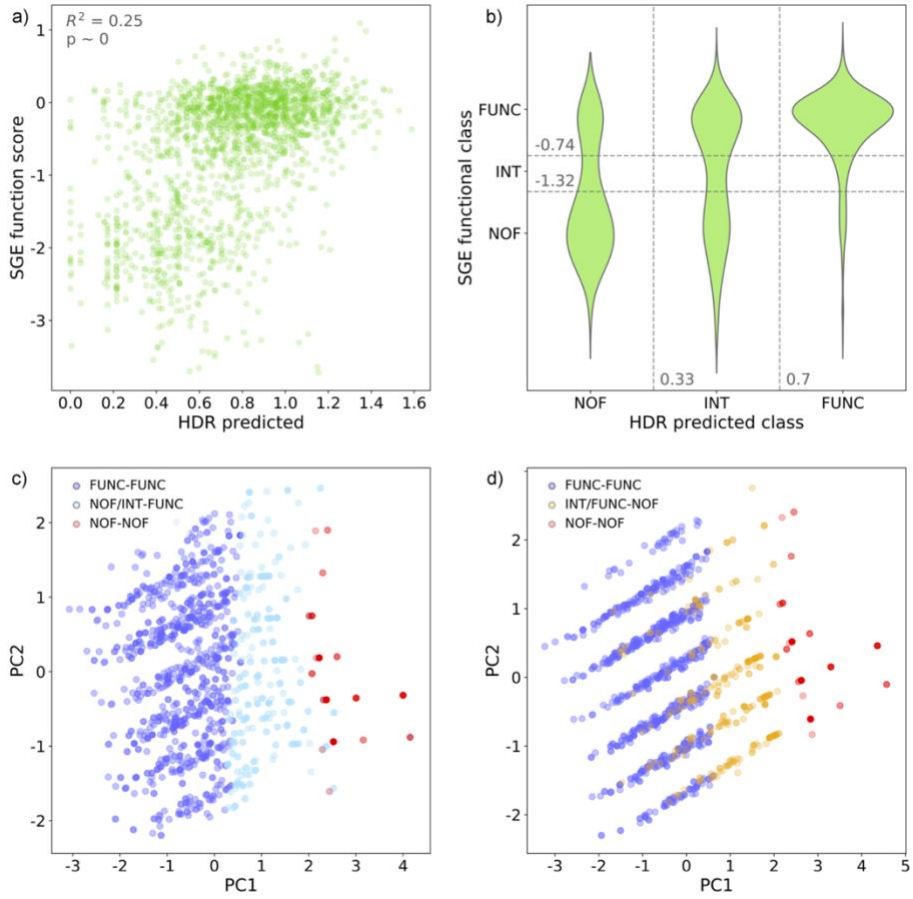


Figure 3.4 ***Prediction of the "saturation genome editing" (SGE) experiment in BRCA1.*** *a) Scatterplot representing SGE values versus HDR predictions for the 1,837 missense variants from (Findlay et al., 2018). b) Violin plot showing the distribution of variants for the different combinations of SGE and HDR functional categories. c) Principal component analysis of three variant populations (HDR-SGE classes): FUNC-FUNC (dark blue), NOF-NOF (red) and the outliers NOF-FUNC plus INT-FUNC (light blue). d) Principal component analysis of three variant populations (HDR- SGE classes): FUNC-FUNC (dark blue), NOF-NOF (red) and the outliers INT-NOF plus FUNC-NOF (yellow).*

This overall coincidence is limited by a substantial scatter. Part of it may be due to technical/biological (inter-exon normalization procedures, impact of RNA levels, etc.) differences between the SGE and HDR experiments that introduce some dispersion in the comparison between both experiments (see Figure 9m from Extended Data Section in (Findlay et al., 2018)). Another part of the scatter is due to limitations of our model.

To better understand these, we divided the SGE-HDR plane into 9 regions (Figure 3.4b), corresponding to the 3x3 combinations of SGE (functional, intermediate and non-functional) (Findlay et al., 2018) and HDR (High, Int, Low) (Starita et al., 2015) equivalent functional classes. The main blocks of outliers correspond to the two top-left and the two bottom-right regions.

We separately used the variants inside each block for a principal component analysis (PCA) (Figures 3.4c and 3.4d), using as variables the three features in our model. As a reference, for each PCA we also included the variants from the upper (functional) and lower (non-functional) diagonal regions. In the plane of the first two principal components (PC1 and PC2) the chosen variants adopt a three-layered disposition, where we successively find the functional, the outliers and the non-functional ones. This disposition reflects the contrast between the bimodal nature of the SGE experiment and the smoother nature of our model.

In fact, in Figure 3.5 we can see that those outlier variants indeed tend to have intermediate values (comprised between those of the functional and non-functional populations) for the features in our model. This suggests that for these variants we need to improve our representation of protein impact with new properties, to reproduce more accurately the results of the SGE experiment. However, it may also indicate the need to consider the effect of variants on other aspects of gene function, like RNA levels (Findlay et al., 2018).

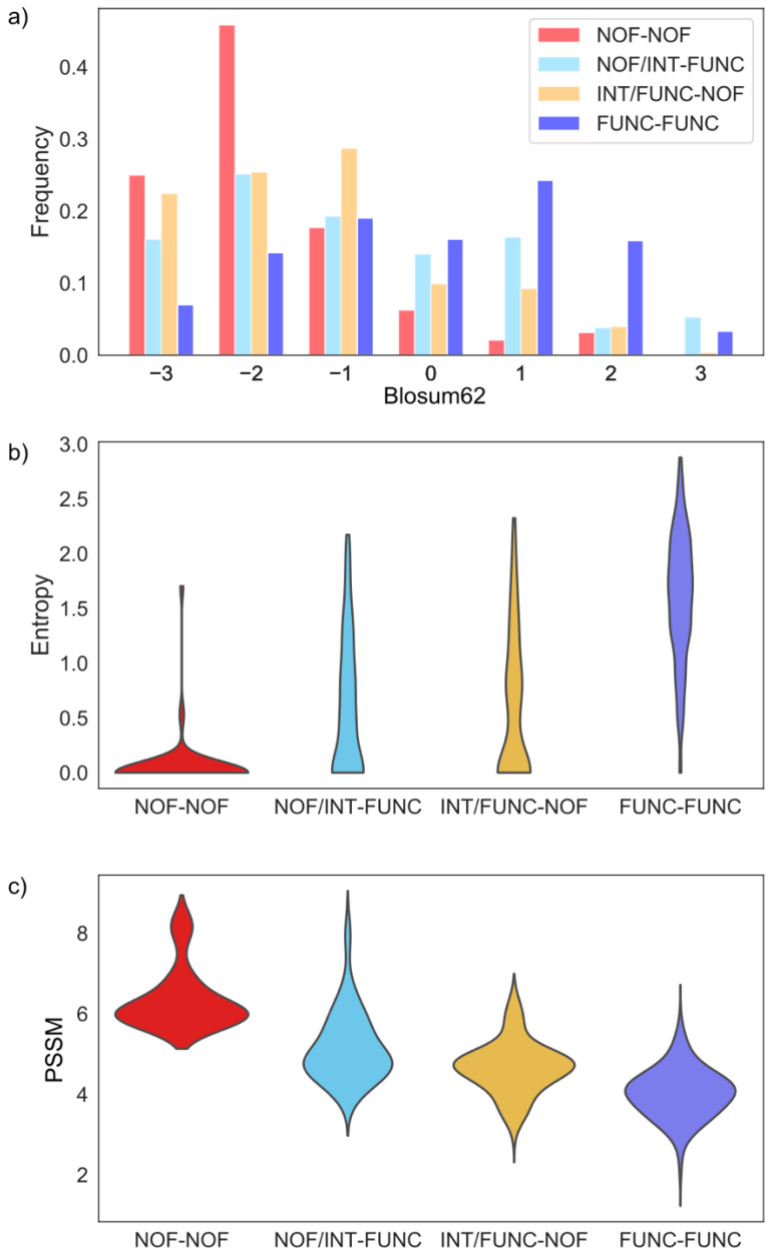


Figure 3.5 ***Feature distribution for the outlier populations***. *The three figures represent value distributions for a) Blosum62, b) Entropy, and c) PSSM. In each figure, we display four variant populations from the following regions in Figure 3.4: FUNC-FUNC (dark blue), NOF-NOF (red), the outliers from the quadrants NOF-FUNC plus INT-FUNC (light blue), and the outliers from the quadrants INT-NOF plus FUNC-NOF (orange).*

3.3.4. Predicting the clinical impact of variants: the NN predictor

We developed two NN predictors, one per protein. These methods were trained with the idea of predicting the clinical significance of a given variant. To this end, during the training process, each variant was labeled with a binary version of this clinical impact: pathogenic/neutral. Here, the larger amount of training data compared to our previous MLR methods (Table 3.1a), allowed us to work with three additional features, fully adhering to our protocol for the obtention of protein-specific predictors (Riera et al., 2016).

As for the MLR predictors, the performance obtained with the NN predictors (Table 3.3) are comparable to those of other protein-specific predictors. Their accuracies, 0.77 for both BRCA1 and BRCA2, are almost within the 0.79-0.99 accuracy range for protein-specific predictors; the same happens for the MCC, 0.55 for BRCA1 and 0.47 for BRCA2. The sensitivities and specificities are more balanced for both BRCA1 (SP: 0.72, SN: 0.86) and BRCA2 (SP: 0.77, SN: 0.75) when compared with what happened for the MLR predictors.

Overall, as in the case of MLR, the results indicate that the more clinically flavored NN predictors have a prediction power comparable to that of other protein-specific predictors (Riera et al., 2016).

3.3.5. Comparison with general pathogenicity predictors

To contextualize the performance of our protein-specific predictors, we give the results of our cross-validate predictions along with the outcomes of a representative set of general predictors: CADD (Kircher et al., 2014), PolyPhen-2 (I. Adzhubei et al., 2010), SIFT (Kumar et al., 2009), PON-P2 (Niroula et al., 2015) and PMut (López-Ferrando, Gazzo, De La Cruz, Orozco, & Gelpí, 2017) (Figure 3.6).

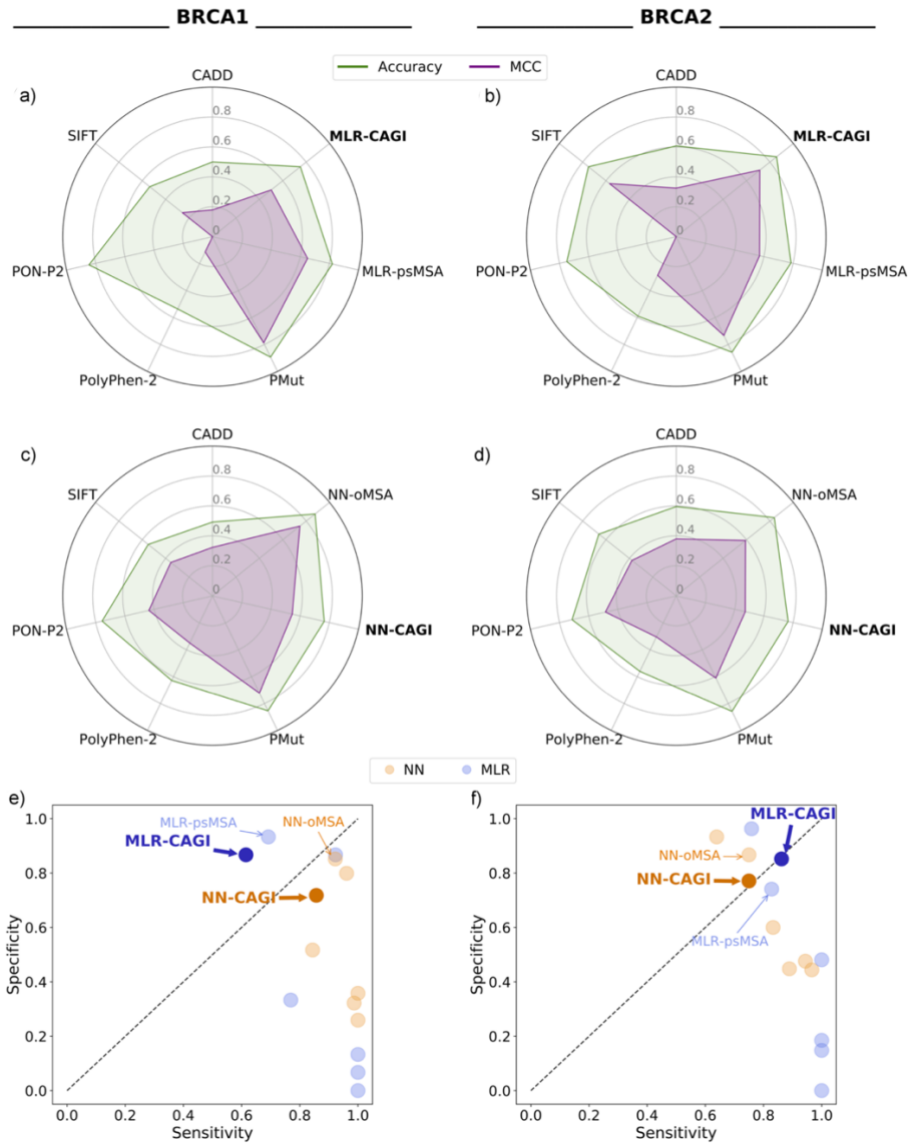


Figure 3.6 ***Binary cross-validated performance of our MLR and NN predictors along with that of the general predictors CADD, PolyPhen-2, PMut, PON-P2 and SIFT.*** *Predictor's accuracy and MCC are calculated based on the training variant dataset of a) BRCA1 MLR b) BRCA2 MLR c) BRCA1 NN d) BRCA2 NN and shown in a radar plot. Predictor's sensitivity and specificity are calculated based on the training variant dataset of NN (orange) and MLR (blue) of e) BRCA1 and f) BRCA2 and shown in a scatterplot.*

Care must be exercised when considering the results of this comparison, since the variants in our datasets can be found in databases like UniProt (Bateman et al., 2017b), commonly used to develop pathogenicity predictors (Riera et al., 2014). Therefore, it is likely that some of these variants have been used in the training of the general methods, thus leading to optimistic estimates of their performance. An additional limitation of the comparison is the small sample size involved in the case of MLR (Figures 3.6a, 3.6b).

In general, we observe that our specific methods have success rates comparable to those of general methods. For MCC, our methods are only surpassed by PMut. For BRCA2, our NN is slightly surpassed by PON-P2 (MCC of 0.47 vs. 0.49), but our MLR surpasses PON-P2 (MCC of 0.71 vs. 0). The sensitivities and specificities of our methods are generally smaller and larger, respectively, than those of other methods. However, our methods have an equilibrated performance for pathogenic and neutral variants (Figures 3.6e, 3.6f), since they display the smallest differences between sensitivity and specificity, 0.14 (BRCA1) and 0.021 (BRCA2) for NN, respectively, and 0.25 (BRCA1) and 0.01 (BRCA2) for MLR. Only PMut has closer values for the MLR training set of BRCA1, 0.06.

3.3.6. Results of the predictors in the CAGI experiment

Here, I present the results of our prediction protocol (Figure 3.1) at the CAGI experiment. For each protein, we submitted four predictions: MLR+AS, NN+AS, MLR, and NN. For simplicity, I will restrict our analysis to the complete protocols (MLR+AS, NN+AS), mentioning protein predictions (MLR, NN) only for discussing the contribution of the AS predictors. Performance was assessed using the class assignments provided by the CAGI organizers after the challenge was closed. More precisely, we computed the ability of our protocols to correctly assign a variant to its class in two different classification

schemes. One is the IARC 5-tier classification system (Goldgar et al., 2008; Plon et al., 2008), which was the one requested by the organizers; the other is a 3-class version of this system (see section 3.2 Materials and Methods).

The fact that we must consider the performance for more than two classes makes the evaluation problem more difficult: in multiclass problems confusion matrices retain their explanatory power, but summary measures are not easy to generalize, nor to interpret (Pierre Baldi et al., 2000; Vihinen, 2012). In our case, the severity of this problem is augmented by the compositional imbalance in the CAGI dataset (Table 3.2). For these reasons, we focus our analysis mainly on the confusion matrices (Figure 3.7) because they provide the basal information in any prediction process and allow a direct interpretation. More concretely, we consider: (i) the diagonal elements to see how good our predictions are; and (ii) the off-diagonal elements to see how incorrect predictions distribute among classes. We treat separately BRCA1 and BRCA2 cases since the performance of both specific and general pathogenicity predictors is protein-dependent (Riera et al., 2016).

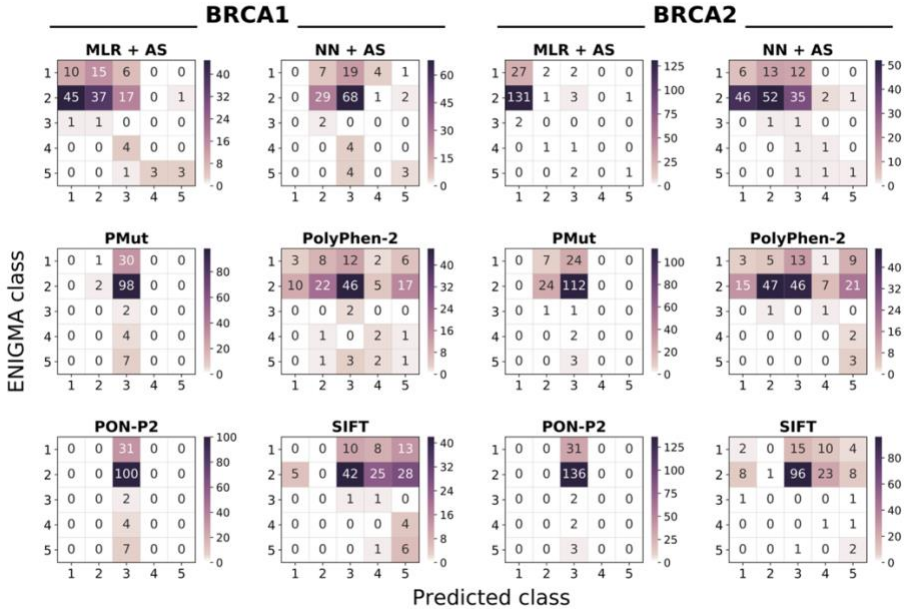


Figure 3.7 ***Confusion matrices of IARC 5-tier predictions on the CAGI dataset****. We provide six heatmaps per protein, two for our predictors MLR+AS and NN+AS, and four for the general predictors PolyPhen-2, PON-P2, PMut, and SIFT. The vertical and horizontal axes correspond to the observed variant class (provided by CAGI) and the predicted IARC 5-tier class, respectively. Diagonal and off-diagonal elements correspond to successful and failed predictions. Given the range differences in predictions, each plot has its own colour scale.*

BRCA1 variants

Looking at the diagonals of the confusion matrices (Figure 3.7), we observe that MLR+AS and NN+AS can recognize, with varying accuracies, members from three (1,2,5) and two classes (2,5), respectively. This overall trend is reflected in the class accuracies, which are higher for MLR-based protocols than for NN-based ones (Table 3.4). If AS predictions are not included, the two methods also fail to recognize class 5 variants (Table 3.4). In fact, for MLR+AS and NN+AS protocols AS predictions are responsible for the accuracy of class 5, which is 0.43 (3 out of 7 correctly predicted variants) in both cases; AS predictions lead to a single failure, for a class 2 variant.

To understand the distribution of incorrect predictions among classes, we consider the off-diagonal elements (Figure 3.7). For MLR+AS, incorrect predictions mostly group at positions adjacent to the diagonal, with only 9 out of 144 variants breaking this trend. For NN+AS this number grows to 31 and predictions (both correct and incorrect) seem to cluster around class 3.

Now, if we analyze the predictions within the unified 3-class framework, we find that class accuracies increase for MLR+AS to 0.82 and 0.56 for Neutral and Pathogenic classes, respectively. For NN+AS, this is not the case due to the previously mentioned clustering of predictions around class 3. Accuracy for the Unknown class is the same as that for IARC 5-tier class 3, since the classes are the same.

BRCA1

IARC 5 class	**1** (<0.001)	**2** (0.001-0.049)	**3** (0.05-0.949)	**4** (0.95-0.99)	**5** (>0.99)
MLR	0.323	0.37	0	0	0
MLR + AS	0.323	0.37	0	0	0.429
NN	0	0.29	0	0	0
NN + AS	0	0.29	0	0	0.429
3 class	**Neutral**		**Unknown**	**Pathogenic**	
MLR	0.817		0	0.273	
MLR + AS	0.817		0	0.545	
NN	0.275		0	0	
NN + AS	0.275		0	0.273	

BRCA2

IARC 5 class	**1** (<0.001)	**2** (0.001-0.049)	**3** (0.05-0.949)	**4** (0.95-0.99)	**5** (>0.99)
MLR	0.871	0.007	0	0	0
MLR + AS	0.871	0.007	0	0	0.333
NN	0.194	0.382	0.5	0.5	0
NN + AS	0.194	0.382	0.5	0.5	0.333
3 class	**Neutral**		**Unknown**	**Pathogenic**	
MLR	0.97		0	0	
MLR + AS	0.964		0	0.2	
NN	0.701		0.5	0.4	
NN + AS	0.701		0.5	0.6	

Table 3.4 ***Class accuracies of our predictors for the CAGI variants.*** *The colour shading reflects the correspondence between both class systems.*

Finally, we compare the performance of our predictors with that for the general predictors for which the output directly corresponded to a probability of pathogenicity (we only excluded CADD, because the score has another scale) (Figure 3.7).

For the chosen predictors (PMut, PolyPhen-2, PON-P2, and SIFT) their score is a probability of pathogenicity that can be transformed into an equivalent of the IARC 5-tier classes, using the ENIGMA conversion table (see Materials and Methods). Focusing on the most frequent CAGI variants (31 from class 1; 100 from class 2), we see that MLR+AS performs better than general methods; for class 5, all general methods, except SIFT, identify fewer correct variants. The case of SIFT is of interest since some of the class 5 variants appear to be splicing variants according to our AS predictions: at this point, and without further evidence, it is unclear which is the correct view, the amino acid view provided by SIFT or the nucleotide view provided by AS predictions.

For classes 3 and 4, the size of the sample, two and four variants, respectively, limits the value of the results, which are: for the two variants of class 3, MLR+AS performs worse than general methods; for the four variants of class 4, only PolyPhen-2 correctly identifies two of them. A remarkable feature of MLR+AS, relative to general methods, is that its predictions form a band around the diagonal, while general methods either scatter their predictions (PolyPhen-2, SIFT) or cluster them around class 3 (PON-P2 and PMut).

Comparison of NN+AS with general methods (Figure 3.7) shows similarities with PON-P2 and PMut, and a failure to identify members of class 1 that is shared with all general methods, except PolyPhen-2; again, AS predictions favor our method for class 5, except in the case of SIFT.

The comparison within the three-class framework (Figure 3.8) confirms the previous trends, with MLR+AS having the largest class accuracy for Neutral, 0.82, well over that of general methods (0.33 for PolyPhen-2; 0.04 for SIFT, 0.02 for PMut and 0 for PON-P2). MLR+AS displays the second best accuracy for Pathogenic, together with PolyPhen-2 and behind SIFT. NN+AS again shows a performance below that of these two general methods, but above that of PON-P2 and PMut.

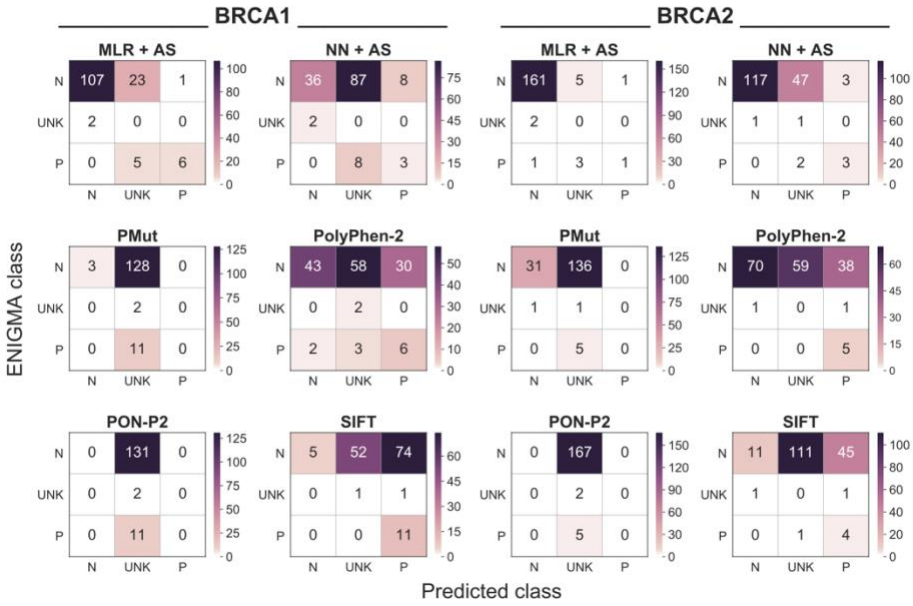


Figure 3.8 ***Confusion matrices of 3-tier predictions on the CAGI dataset.*** *We provide six heatmaps per protein, two for our predictors MLR+AS and NN+AS, and four for the general predictors PolyPhen-2, PON-P2, PMut, and SIFT. In all plots, the vertical and horizontal axis respectively correspond to the observed variant class (provided by CAGI organizers) and the predicted classes in the 3-class reduced version of the IARC 5-tier classification. Diagonal and off-diagonal elements correspond to successful and failed predictions, respectively. Given the range differences in the predictions, each plot has its own colour scale.*

BRCA2 variants

For BRCA2, the situation is somewhat different. The diagonal elements of the confusion matrix (Figure 3.7) show that NN+AS can recognize variants from the five classes, with varying accuracies (Table 3.4), while MLR+AS recognizes only variants from classes 1, 2 and 5. Additionally, for the most frequent classes (1, 2) NN+AS is more balanced than MLR+AS (Figure 3.7, Table 3.4): 0.19 (1) and 0.38 (2) vs. 0.87 (1) and 0.01 (2), respectively.

Inspection of the off-diagonal elements shows that wrong predictions are more spread for NN+AS than for MLR+AR. For example, for MLR+AS, essentially all (97%) the incorrect predictions of class 2 go to class 1, while this figure drops to 55% for NN+AS. As before, the tiny number of variants in the remaining classes reveals no clear trends. The AS predictions result in one correctly identified member of class 5 for the two versions of our protocol; AS predictions lead to a single failure, for a class 2 variant.

As for BRCA1, reduction of the five IARC 5-tier classes to a 3-class system reveals a reversion in the previous trend, with a high class accuracy for Neutral, higher for MLR+AS (0.96) than for NN+AS (0.70). Accuracy for the Unknown class is the same as that for IARC 5-tier class 3, because the classes are the same. For the Pathogenic class, NN+AS still performs better than MLR+AS (Figure 3.7, Table 3.4).

Finally, we compare the performance of our predictors with that for the general predictors for which the output directly corresponded to a probability of pathogenicity (we only excluded CADD, because the score has another scale) (Figure 3.7).

Focusing on the most frequent CAGI variants (31 from class 1; 136 from class 2), we see that NN+AS performs better than general methods; MLR+AS is only better for class 1; for class 2 its accuracy is low, the same as SIFT, and below

that of PolyPhen-2 and PMut. For classes 3, 4 and 5, the sample size is smaller than that of BRCA1 (2, 4, 7 vs. 2, 2, 3 variants for BRCA1 and BRCA2, respectively); for this reason, we believe that for these variants it is preferable to wait for next rounds of the CAGI challenge to assess the performance the different *in silico* tools, including ours.

The comparison within the three-class framework (Figure 3.8) confirms the previous trends, showing that for the Neutral class (167 out of 174 CAGI variants) both MLR+AS and NN+AS surpass general methods (Figure 3.8). For the Pathogenic class (5 variants), PolyPhen-2 and SIFT have the best performances, while our methods rank third (MLR+AS) and fourth (MLR+AS).

3.4. Discussion

Obtaining good estimates of the functional impact and cancer risk of BRCA1 and BRCA2 sequence variants plays a vital role in the diagnosis and management of inherited breast and ovarian cancers (Eccles et al., 2015; Findlay et al., 2018; Guidugli et al., 2018; Moreno et al., 2016; Paluch-Shimon et al., 2016). A priori, *in silico* tools can be used to obtain these estimates; however, their moderate success rate restricts their applicability (Ernst et al., 2018). In this work, we have addressed this issue focusing on the problem of predicting the pathogenicity of BRCA1/2 missense variants, applying the protein-specific approach (Riera et al., 2014). We validated the performance of the resulting BRCA1- and BRCA2-specific tools in two different ways: (i) in isolation, using manually curated sets of functionally and clinically annotated variants; and (ii) in combination with predictors of splicing impact (Figure 3.1), to interpret the variants from the ENIGMA challenge of the CAGI 5 experiment.

3.4.1. Performance of the predictors in isolation

When tested in isolation, we find that our two methods (MLR and NN) are competitive when compared with general methods (section 3.3.5, Table 3.3 and Figure 3.6), for both BRCA1 and BRCA2. In particular, their specificities are among the best, a property desirable from the point of view of HBOC diagnosis requirements (Ernst et al., 2018); they also have the best balances between specificity and sensitivity, with the only exception of PMut in BRCA1, which has slightly better figures for the MLR training set.

General methods also show good success rates in our training sets (Figure 3.6), in contrast with the usually lower performance estimates cited in the literature. For example, the last version of PMut displays an MCC of 0.31 for

both BRCA1 (63 variants) and BRCA2 (104 variants) (López-Ferrando et al., 2017). In the same work, we find MCC values for other tools, computed on the same dataset: for BRCA1 they vary between 0.17 (PROVEAN) and 0.38 (LRT); for BRCA2 they vary between 0.01 (PROVEAN) and 0.19 (Mutation Assessor). In a previous study, using a small dataset of BRCA2 variants, Karchin et al. (Karchin, Agarwal, Sali, Couch, & Beattie, 2008) find that general tools display good sensitivities but low specificities. A similar trend has been recently reported by Ernst et al. (Ernst et al., 2018), after testing PolyPhen-2, SIFT, AlignGVGD and MutationTaster2 in a set of 236 BRCA1/2 variants.

These authors express concern about the moderate performance observed, particularly about the low specificities observed relative to HBOC diagnosis requirements (e.g., PolyPhen-2: 0.56 and 0.72 for BRCA1 and BRCA2, respectively). We believe that our higher estimates for general predictors (Table 3.3 and Figure 3.6), relative to those in the literature, may partly result from the overlap between their training sets and our dataset.

Presently, stand-alone use of *in silico* methods for HBOC diagnosis is discouraged (Ernst et al., 2018). Nonetheless, it is considered that these methods can be fruitfully combined with the results of functional assays, to provide an alternative to multifactorial models in the absence of family information (Guidugli et al., 2018).

The tools presented in this work are easily amenable to this type of approach because of their extreme simplicity and interpretability. This is a consequence of the small number of features utilized (3 and 6 for MLR and NN, respectively) and of the low complexity of our models (Riera et al., 2014). Additionally, our MLR models allow a direct interpretation of a variant's impact at the molecular level, because they produce estimates of the HDR assay for the target variant. In this sense, the MLR approach resembles that of Starita et al. (Starita et al., 2015) who estimate HDR values using the results

of other functional assays (E3 ligase scores and BARD1-binding scores). In our case, we use instead a few sequence-based features, with two conservation measures (Shannon's entropy and PSSM) standing among them given their recognized predictive power (C. Ferrer-Costa et al., 2004).

Conceptually, MLR methods implement the idea of addressing pathogenicity prediction problems focusing on endophenotypes, rather than on clinical phenotypes. Endophenotypes are quantitative measures of intermediate phenotypes with clinical relevance (D.L. Masica & Karchin, 2016); they are closer to the genotype and, for this reason, may result in predictors with higher success rates, given the small contribution of genetic background and environmental effects to the outcome of the variant.

In general, this is the case when looking at clinical performance (Table 3.3, Figure 3.6). However, BRCA1 sensitivity (0.62) is low compared to specificity (0.87); while this may be a consequence of the discretization of the HDR prediction, it may also be a consequence of the simplicity of our model.

When testing the MLR model with SGE data we observe a significant correlation (Spearman's ρ=0.47, p-value~0), comparable to that of Align-GVGD (ρ=0.46) and better than that of CADD (ρ=0.40), PhyloP (ρ=0.36), SIFT (ρ=0.36) and PolyPhen-2 (ρ=0.28) (values obtained from (Findlay et al., 2018), Extended Data Figure 9). However, visual inspection shows the presence of substantial deviations from a monotonic relationship (Figures 3.4a and 3.4b).

If we analyze the population of outliers using PCA and value distributions of the features in our model (Figure 3.5) we see that, generally, they have an intermediate behavior between functional and non-functional variants for all features. This points to an aspect of the variant's impact that is poorly represented by our present set of features, like the effect of the mutation in RNA levels.

Finally, it is worth mentioning that our MLR predictors have been trained with small sets of variants that are concentrated in a reduced region of BRCA1 and BRCA2, the domains responsible to carry out the homology direct repair function (Figure 3.9). This is in contrast with the broader range of positions covered by the NN and the CAGI datasets. The fact that, in spite of this situation, the MLR tools are competitive suggests that the discriminative features allow them to capture some general effect of variants on protein function/structure, like impact on stability (Yue, Li, & Moult, 2005).

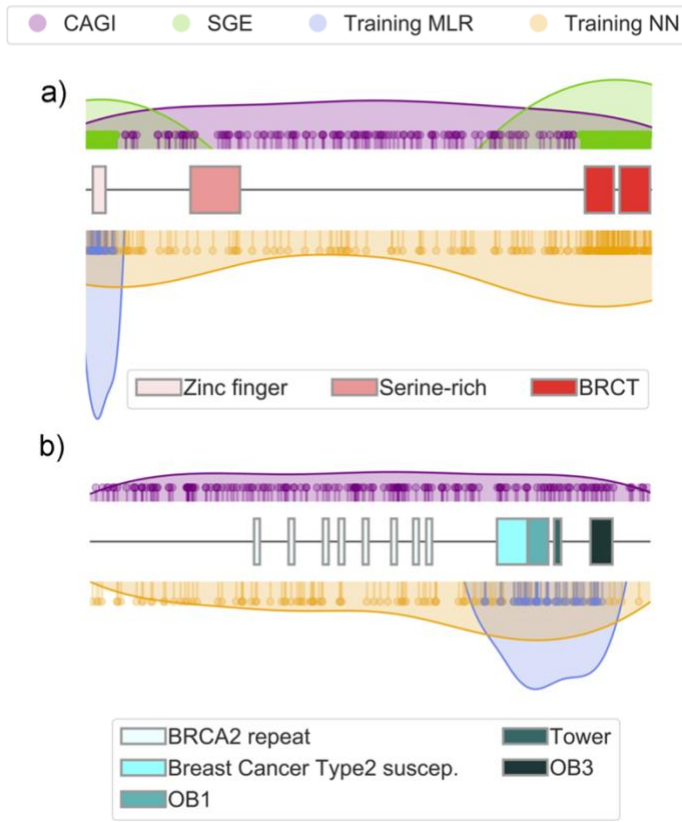


Figure 3.9 ***Distribution of the variants along the BRCA1 and BRCA2 sequences****. Each variant used in this work is represented with a pin indicating its location, and a coloured surface that provides a general, smoothed view of the distribution. The different functional domains in each protein are represented with boxes. For representation purposes, BRCA1 (1863 amino acids) and BRCA2 (3418 amino acids) are displayed with the same length.*

3.4.2. Performance of the predictors in the CAGI 5 experiment

The ENIGMA challenge within the CAGI experiment provides a good opportunity to independently validate the performance of pathogenicity predictors for BRCA1/2. Two aspects are specific of the ENIGMA challenge. First, if some of the target variants are pathogenic, the participants do not know what molecular effect originates their pathogenicity: it can be the impact on protein function, but it can also be the impact on splicing (Eccles et al., 2015). For this reason, we decided to combine predictions for these two effects in our protocol (Figure 3.1).

A second, distinctive aspect of the challenge is that the submissions had to provide the predicted IARC 5-tier class for each variant (see section 3.2.1). This is relevant since this classification is strongly related to the clinical actions associated to each class (Goldgar et al., 2008; Moghadasi et al., 2016; Plon et al., 2008) which are in turn related to factors such as impact on the counselee or cost to the healthcare system. Collective consideration of these factors crystallizes into five decision regions (Plon et al., 2008) that are applied to the posterior probability of pathogenicity, a probability obtained after integrating different sources of clinical/biomedical evidence.

In our case, this probability was estimated using only molecular information; nonetheless, to adapt our output to the CAGI requirements we directly applied the ENIGMA boundaries (section 3.2.4 and 3.2.5). We computed our performances on the basis of this assignment; however, we also obtained the performances for a simplified version of the ENIGMA classification, separately collapsing its neutral and pathogenic classes (Table 3.2).

Assessment of the results obtained (Figure 3.7, Figure 3.8, Table 3.4 and Table 3.5) shows some clear trends. For the 5-class problem, all the methods (both

ours and the general methods) have poor per class performances; however, our methods are more successful at reproducing the compositional bias of the sample and outperform general methods for the most abundant classes (1 and 2) in BRCA1/2, with only one exception, for class 2 in BRCA2, both PolyPhen-2 and PMut surpass MLR+AS; our methods also have a better distribution of wrong predictions among classes, because they tend to cluster nearby the correct class. These trends are reinforced when reducing the number of classes from five to three.

BRCA1								
IARC 5 class	**MLR**	**MLR+AS**	**NN**	**NN+AS**	**PMut**	**PolyPhen-2**	**PON-P2**	**SIFT**
ACC	0.326	0.347	0.201	0.222	0.028	0.208	0.014	0.049
MCC	-0.041	0.006	0.015	0.056	-0.002	0.031	0	0.021
3 class	**MLR**	**MLR+AS**	**NN**	**NN+AS**	**PMut**	**PolyPhen-2**	**PON-P2**	**SIFT**
ACC	0.764	0.785	0.25	0.271	0.035	0.354	0.014	0.118
MCC	-0.237	0.354	-0.012	0.055	0.026	0.136	0	0.123
BRCA2								
IARC 5 class	**MLR**	**MLR+AS**	**NN**	**NN+AS**	**PMut**	**PolyPhen-2**	**PON-P2**	**SIFT**
ACC	0.161	0.167	0.345	0.351	0.144	0.305	0.011	0.034
MCC	-0.109	-0.068	-0.017	-0.006	-0.029	0.078	0	0.017
3 class	**MLR**	**MLR+AS**	**NN**	**NN+AS**	**PMut**	**PolyPhen-2**	**PON-P2**	**SIFT**
ACC	0.931	0.931	0.69	0.695	0.184	0.431	0.011	0.086
MCC	0.18	0.277	0.185	0.213	-0.013	0.125	0	0.022

Table 3.5 ***Overall accuracies (ACC) and MCC in the CAGI dataset*** *for our two methods with and without splicing; and four general methods.*

Overall, the results for the CAGI challenge show that our methods can identify low-risk variants with an accuracy higher than that of general methods, a desirable property for HBOC diagnosis (Ernst et al., 2018). Part of this improved performance could be attributed to an unequal effect of applying the ENIGMA decision boundaries to the posterior probability generated by general methods. We believe that this mapping procedure may play a role,

but not a determining one since comparison of the original, binary predictions of the general methods with those of the binary versions of our tools (MLR scores binarized as explained in section 3.2.5) gives a similar result (Table 3.7) again. MLR+AS has the top specificities for BRCA1/2 and high sensitivities; NN+AS has the same sensitivities but lower specificities, nonetheless these are only surpassed by PMut.

BRCA1							
2 class	**MLR+AS**	**NN+AS**	**CADD**	**PMut**	**PolyPhen-2**	**PON-P2**	**SIFT**
ACC	0.347	0.201	0.222	0.028	0.208	0.014	0.049
MCC	-0.041	0.015	0.056	-0.002	0.031	0	0.021
BRCA2							
2 class	**MLR+AS**	**NN+AS**	**CADD**	**PMut**	**PolyPhen-2**	**PON-P2**	**SIFT**
ACC	0.167	0.345	0.351	0.144	0.305	0.011	0.034
MCC	-0.068	-0.017	-0.006	-0.029	0.078	0	0.017

Table 3.7 ***Binary performances for our predictors and general predictors.***

In summary, we have applied the protein-specific approach to building a pathogenicity predictor for BRCA1/2 variants, using either clinical phenotypes or endophenotypes. The results obtained from our methods indicate that this approach can contribute to improve our ability to discriminate between high- and low-risk variants for BRCA1/2. Of particular interest is the MLR+AS tool, because it gives an estimate of the molecular impact of a sequence replacement that is easy to interpret since it corresponds to an *in silico* version of the HDR assay. Participation in the CAGI experiment has allowed us to obtain independent estimates of the performance of our predictors, to compare them with other predictors and to help us clarify the classification level at which *in silico* tools could be useful for HBOC diagnosis. This participation has also underlined the role that splicing predictions can play in the correct annotation of BRCA1/2 variants, particularly when integrated in protocols that combine different views of a variant's impact.

3.5. Conclusions

Germline variants in BRCA1 and BRCA2 may disrupt the DNA repair function of these proteins increasing the risk of hereditary breast and ovarian cancers. Correct assessment of these variants thereby, becomes clinically relevant as it may increase the survival rates of its carriers. Unfortunately, we are still unable to systematically predict the impact of BRCA1/2 variants.

Here, we presented a family of *in silico* predictors that address this problem, using a protein-specific approach. For each BRCA protein, we developed two predictors that estimate at two different levels the impact of a variant: the molecular function and the clinical significance. The performance of these predictors with different datasets was good, in spite of the small number of predictive features and the limited size of the variant sets used for training.

Additionally, these tools were applied to the BRCA1/2 variants of the ENIGMA challenge in CAGI 5 experiment. We found that these predictors, particularly those estimating the functional impact of variants, have a good performance, being able to predict the large compositional bias towards neutral variants in the CAGI sample. The performance is further improved when incorporating to the prediction protocol, the estimates of the variant's impact on splicing.

4. Dissemination of BRCA1/2 specific pathogenicity predictor among the scientific community

4.1. Introduction

A germline variant in the high susceptibility genes BRCA1 or BRCA2 is the greatest risk factor for HBOC (Roy et al., 2012), which produces a higher than normal levels of breast, ovarian and additional cancers.

Genetic testing of these genes often reveals a set of variants that are not easily classified. Identification of those that are pathogenic is the key to channel its carriers to the proper programs of prevention, surveillance and target therapies (Paluch-Shimon et al., 2016). But this becomes a hard problem when variants do not have a clearly damaging effect on the protein function like missense variants.

In this case, segregation analysis in cancer affected families can be used to decipher the effect of these variants. But frequently, this is not a feasible approach (Toland & Andreassen, 2017). Then, functional studies can be carried out to measure the impact of variants in a specific protein function, although these assays can be technically challenging and also time and cost demanding (Starita et al., 2015). In this scenario, *in silico* methods constitute an inexpensive alternative to provide new information to facilitate the variant interpretation process. However, these methods are not always easy to use or, many times, lack the context that would allow clinical users a fruitful interpretation.

Here, we address this problem and present an open access website which can be found at https://www.biotoclin.org/BRASS, with the purpose of facilitating the usage of our methodology to the community. It includes our recently developed predictors of the impact of missense variants in BRCA1 and BRCA2 genes (Padilla et al., 2019) (see Chapter 3). No software

installation is needed, and the results can be accessed from both a computer and a smartphone with Internet access.

Once in our website, there are different options to address the prediction problem: the researcher can query either a single variant or a set of them at once, and the pathogenicity predictions appear promptly, since the predictions of the more than 100,000 possible missense variants of BRCA1 and BRCA2 have all been pre-computed.

Moreover, in order to facilitate the interpretation of the predictions, the user will find a measure of the reliability of the prediction, many links to different sources of biomedically related information, and a description of the explanatory features used by the predictor. Thus, the user can have a more complete view of the variant's impact that may help him/her interpret the pathogenic/neutral prediction.

4.2. Materials and Methods

I developed the website https://www.biotoclin.org/BRASS in Python, using the Django framework (djangoproject.com) (Figure 4.1). Django is based on the Model-View-Controller pattern where the Controller (URL configuration, form data processing...) is separated from the View components. The View components implement the business logic and retrieve the information from the Models, which map to the PostgreSQL relational database storing the predictions of the missense variants. Subsequently, the View components pass the information to the Templates which constitute the webpages made of HTML, CSS, Bootstrap, JavaScript and D3.js. The website runs on a Heroku server and uses Nginx to serve static assets.

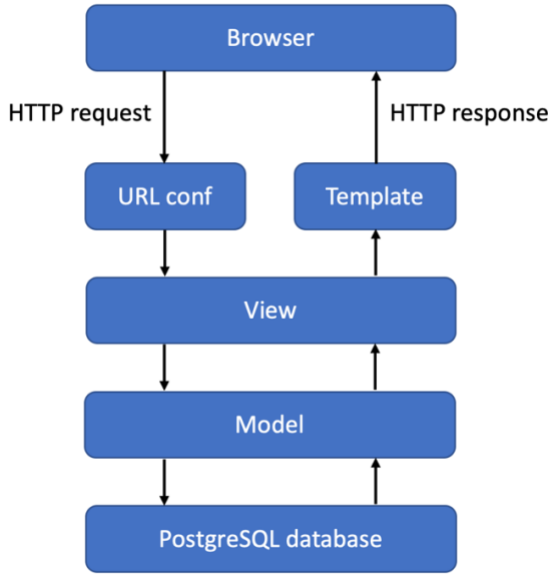


Figure 4.1 ***Website architecture based on the Django framework.*** *In a nutshell, Django receives an HTTP request instance with an URL. The URL is parsed by the URL configuration that redirect it to the proper View function. The View function retrieves the data from the correspondent model, which gets it from the PostgreSQL database. Then, the View passes the information to the Template where it is merged with the HTML and it is sent to the user's browser.*

4.3. Results and Discussion

4.3.1. Obtention of the pathogenicity prediction

BRcA Specific Software (BRASS) at https://www.biotoclin.org/BRASS is a comprehensive website where the user can access a novel family of protein specific pathogenicity predictors to assess the impact of missense variants in BRCA1 and BRCA2 genes (Padilla et al., 2019). This family of pathogenicity predictors score the molecular impact of variants on protein function as measured in the HDR assay and the clinical significance of a variant as collected in the literature (see Chapter 3).

Figure 4.2 ***Main landing page at https://www.biotoclin.org/BRASS.***

In the main landing page (https://www.biotoclin.org/BRASS) (Figure 4.2), the user can find the first step in the prediction process: the selection of the protein to predict: BRCA1 or BRCA2. Additionally, there is an overview of BRASS predictors (NN/MLR), binary predictions and the numerical score behind them, prediction reliability and predictor performance.

After choosing a protein, a new form appears in which the user selects the predictor he/she wants to utilize (NN or MLR) and introduces the variant or list of variants to predict.

After our software validates the variants introduced, the predictions are shown to the user. In the case of a list of variants, a summary table is provided, displaying in each case, the prediction (pathogenic/neutral), the score of the prediction (0-1), and a link to a more complete webpage with rich information on the variant and its protein context (Figures 4.3 - 4.8). This page is shown directly if the user submitted a single variant only.

The variant's prediction webpage (Figures 4.3 - 4.8) is structured into three parts: (i) the prediction of the variant, (ii) the understanding of the prediction and (iii) additional information for interpreting the variant.

The prediction of the variant (Figure 4.3) includes the predicted class, which is binary and can be pathogenic or neutral; the numerical score, which ranges from 0 to 1, being 1 the most pathogenic value; and reliability, which ranges from 0 to 5 being 5 the highest reliability.

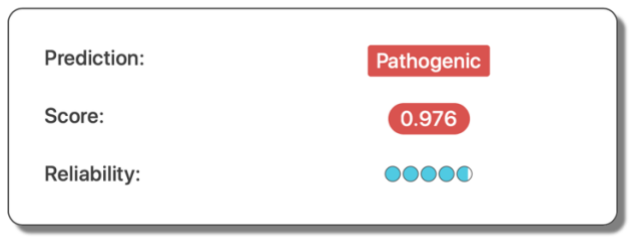


Figure 4.3 ***Prediction of the variant BRCA1 p.Cys61Gly with the NN method.***
Adapted from https://www.biotoclin.org/predictor/BRCA1/NN-P38398-C61G.

Next to it, there is a table summarizing the prediction of other frequently used predictors such as Align-GVGD (Tavtigian et al., 2006), PolyPhen-2 (I. Adzhubei et al., 2010), SIFT (Kumar et al., 2009), PON-P2 (Niroula et al., 2015) or CADD (Kircher et al., 2014). Additionally, in the FAQs section (https://www.biotoclin.org/predictor/BRCA1/help/), we can find the performance of these predictors (Figure 4.4), obtained as described in Chapter 3.

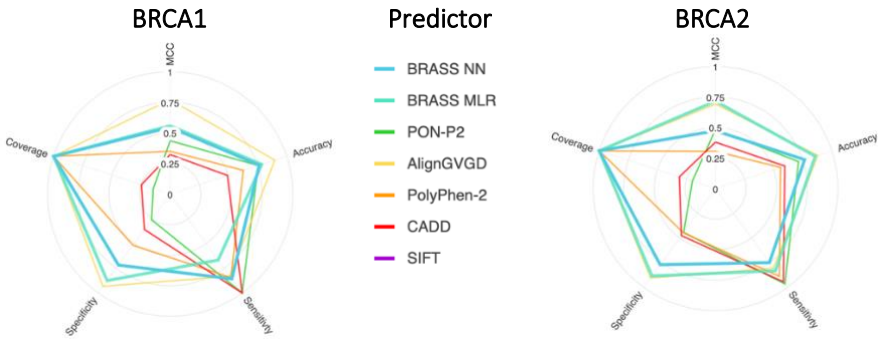


Figure 4.4 ***Performance metrics (MCC, Accuracy, Sensitivity, Specificity and Coverage) of in-house predictors BRASS NN and BRASS MLR and external predictors PON-P2, Align-GVGD, PolyPhen-2, CADD and SIFT.***

4.3.2. Interpreting the pathogenicity prediction

To provide a better understanding of our predictions, we add a section contextualizing the predicted score and the explanatory features used by the predictor, relative to the values of a curated dataset of pathogenic and neutral variants (Padilla et al., 2019).

In this section of the web, the user can compare graphically, the score for his/her variant relative to those of the variants in the reference dataset (Figure 4.5).

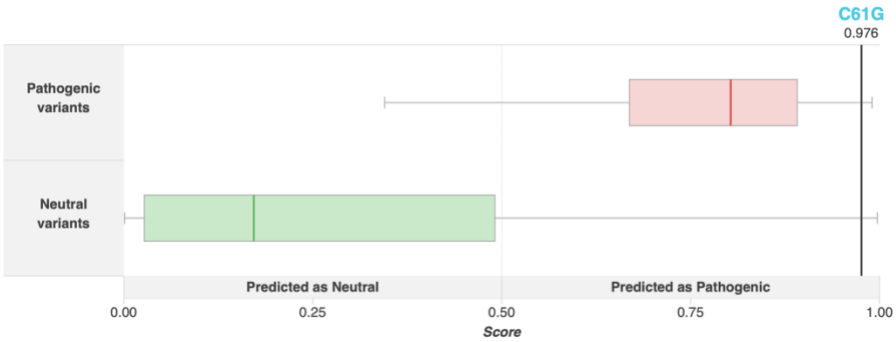


Figure 4.5 ***Horizontal boxplot of NN predicted scores of pathogenic (red boxplot) and neutral (green boxplot) variants in the curated dataset. The score of the predicted variant BRCA1 p.Cys61Gly is shown with a vertical line.***

Underneath, the user can find a set of boxplots showing a break down of the prediction score into its components. This allows to identify which feature has the largest contribution to the prediction (Figure 4.6).

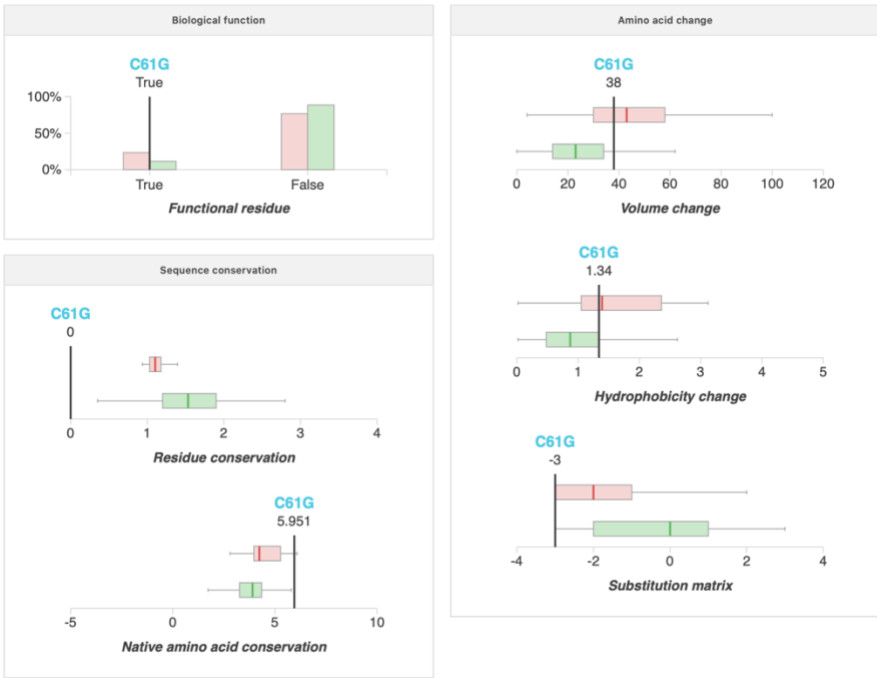


Figure 4.6 ***Boxplots of the values of the explanatory features used by the NN predictor of the curated pathogenic (red) and neutral (green) variants along with the variant of interest BRCA1 p.Cys61Gly.***

These plots allow the user to get a fair idea of whether the pathogenicity signal was or was not strong, and whether it was the result of a clear signal for one/several features or just the sum of small trends. This may help him/her to reach an informed decision on whether the prediction can be trusted and how much.

4.3.3. Additional information of the variant's impact

We provide a table (Figure 4.7) summarizing additional information on the impact of the variant retrieved from several databases: the clinical significance of the variant from ClinVar (Landrum et al., 2016) and UniProt (Bateman et al., 2017a) databases; the functional relevance of the mutated residue i.e. active site, DNA binding domain, etc. from UniProt (Bateman et al., 2017a); the population allele frequency from ExAC (Lek et al., 2016) and gnomAD (Karczewski et al., 2020) database; and variant information from dbSNP (Sherry et al., 2001) and Ensembl (Hunt et al., 2018) databases.

Clinical Evidence	ClinVar	Pathogenic (reviewed by expert panel)
	UniProt	Disease
Biological Relevance	Functional residue	zinc finger region: RING-type
Variant Information	dbSNP	-
	Ensembl	variant
Population Allele Frequency	ExAC	6.722e-05
	gnomAD	3.255e-05

Figure 4.7 ***Table summarizing additional information on the impact of the variant of interest BRCA1 p.Cys61Gly from several databases.*** *Adapted from https://www.biotoclin.org/predictor/BRCA1/NN-P38398-C61G.*

This information is also valuable to assess the reliability of the prediction. For example, if we collect the ClinVar's clinical significance of the BRCA1 and BRCA2 variants in the reference dataset and compare them with the results of the NN or MLR predictors, we see that the majority of variants have a prediction concordant with their clinical significance. The situation is less clear for BRCA2 ClinVar's pathogenic variants, which have pathogenic as well as neutral predictions (Figure 4.8).

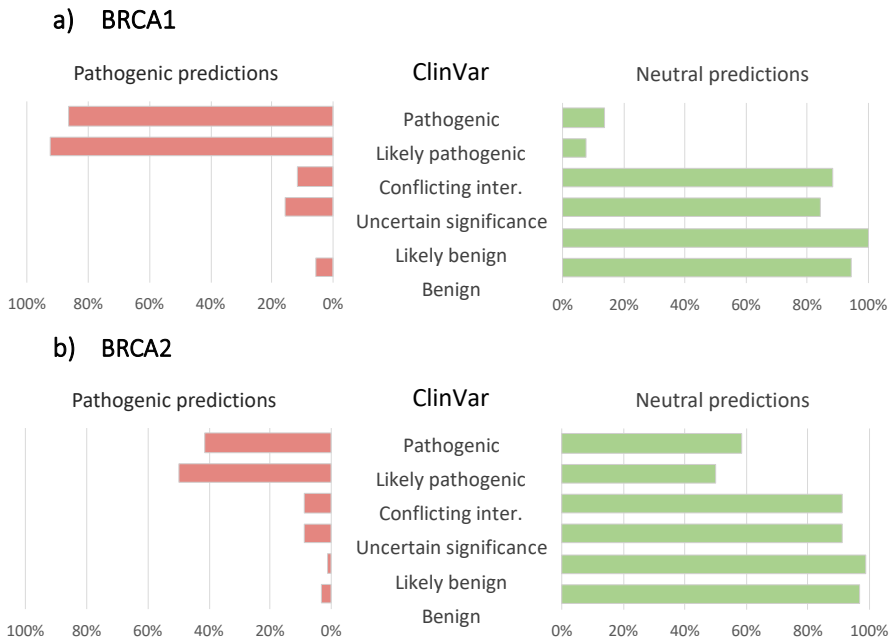


Figure 4.8 ***Concordance between pathogenicity predictions of NN method in BRCA1 (a) and BRCA2 (b) with ClinVar's clinical significance of variants.***

Moreover, to understand completely the context of the variant within the protein, we add a plot of the protein sequence in which additional information is displayed. This plot can be navigated using a zoom tool (Figure 4.9) and can be utilized to see the distribution of the exons of BRCA1/2, its protein domains, the functional residues, and reference neutral and pathogenic variants along the protein.

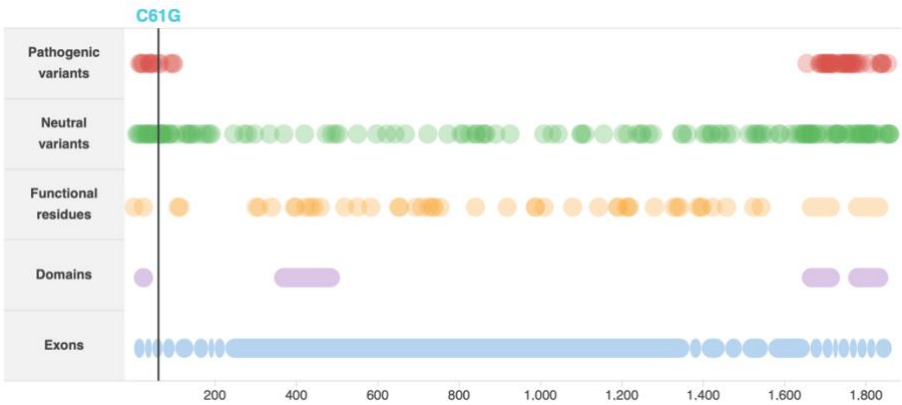


Figure 4.9 ***Plot of the variant C61G localized within BRCA1 protein with curated pathogenic (red) and neutral (green) variants, functional residues (orange), protein domains (lilac) and exons (blue).*** *In the website, the user can use a zoom tool to enlarge the region surrounding the variant and visualize it with more detail. Moreover, when the user hovers over one of these elements, information regarding the name, function or localization of the element appears.*

4.3.4. Downloading the pathogenicity predictions

If the user is interested in a more frequent use of the NN and MLR predictions, other approaches more user-friendly are available for him/her. For instance, the user can download a file with the predictions of all the missense variants of BRCA1/2, or can download programmatically a batch or a single variant throughout the RESTful API (Figure 4.10).

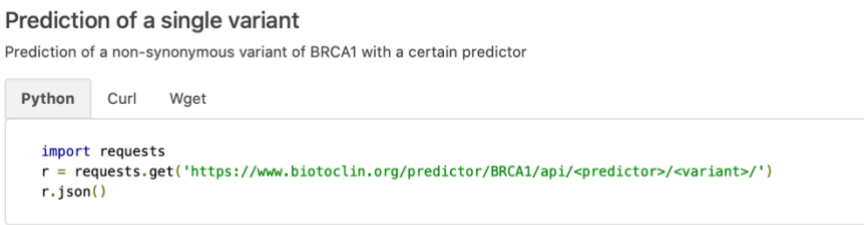


Figure 4.10 ***Detail of the RESTful API to download the pathogenicity prediction of a variant.*** *Adapted from https://www.biotoclin.org/predictor/BRCA1/download/.*

4.4. Conclusions

BRASS (https://www.biotoclin.org/BRASS) is a user-friendly website, available to the scientific community for the use of a novel family of protein specific pathogenicity predictors of BRCA1 and BRCA2 missense variants (Padilla et al., 2019). This family of *in silico* tools score the molecular impact of variants on protein function estimating the value of the HDR assay (MLR predictor) and predicting the clinical significance of a variant (NN predictor). In addition to provide predictions and their reliability, the website gives supplementary information on the impact of the variant from several databases (ClinVar, UniProt, gnomAD, etc.), as well as, a chart of the variant localized within the protein. This last feature includes additional data regarding the localization of reference pathogenic and neutral variants or the functional domains of the protein. Thereby, we aim to give the user a thorough view of the pathogenicity prediction of the variant and additional resources for a better interpretation.

5. Application of pathogenicity predictors in clinical research: characterization of a novel pediatric neurologic disorder caused by variants in H3.3

The results presented here are part of a manuscript (Bhoj et al., 2020) that is presently under review in Science Advances.

5.1. Introduction

Our group has recently participated in a large international effort (135 researchers involved) led by the Children's Hospital of Philadelphia (CHOP), USA, for the characterization of a novel pediatric neurologic disorder caused by variants in the *H3-3A* and *H3-3B* genes, which encode for an identical protein: histone H3.3, a type of histone 3 (H3). In this chapter, after going through the main characteristics of this project, I will focus on the structural bioinformatics analyses that I carried out and constitute the contribution of our group to this project.

5.1.1. Histone H3.3

Histones are nuclear proteins of eukaryotic cells that pack and order the DNA into structural units called nucleosomes. Nucleosomes are composed of two H2A-H2B dimers and a H3-H4 tetramer, wrapped 1.7 times by 146 bp of DNA around the histone octamer (Luger, Mäder, Richmond, Sargent, & Richmond, 1997). Histones are relatively similar in structure and are highly conserved through evolution, all featuring a helix-turn-helix motif binding DNA and a long tail on one end of the protein sequence where post-translational modifications (PTMs) occur dynamically. PTMs regulate several processes such as DNA repair, gene expression, mitosis, and meiosis. Of the four core histones, histone H3 is the most heavily modified, totaling at least 26 potentially modified amino acids (Young, DiMaggio, & Garcia, 2010).

Behind the H3 denomination there is a family of very similar proteins. The most prevalent members of this family are H3.1 and H3.2, canonical histones that are replication-dependent and, thereby, added to chromatin during DNA replication in S phase (Frank, Doenecke, & Albig, 2003). H3.3 however, is a histone variant assembled along with H4 into the nucleosome in a replication-

independent manner, with the aid of histone chaperones such as HIRA, DAXX and DEK (Burgess & Zhang, 2013). At the protein sequence level, H3.3 which has 135 amino acids after the cleavage of the first methionine, differs from the canonical H3.1 and H3.2 by five and four amino acids respectively; its three-dimensional structure is composed of a histone tail, four α-helices and two loop domains (Ederveen, Mandemaker, & Logie, 2011).

H3.3 is expressed ubiquitously during development and throughout life, with different expression patterns and levels of *H3-3A* and *H3-3B* (Ederveen et al., 2011). H3.3 has been associated to the maintenance of epigenetic memory, heterochromatin and telomeric integrity (Ng. & Gurdon, 2008; Udugama et al., 2015). Somatic variants in *H3-3A* and *H3-3B* have been strongly associated with pediatric glia and other tumors (G. Wu et al., 2012), but no germline variants in humans have been reported.

5.1.2. Identification of the causative variants

Within the project led by the CHOP, the collaborative consortium has characterized a cohort of 42 patients with core phenotypes of progressive neurologic dysfunction and congenital anomalies, but no malignancies. Notably, nine of the 42 patients (21%) have demonstrated clinical neurologic degeneration, which suggests that this may be a progressive disorder. Multiple patients (26% of the cohort) have cortical atrophy on brain MRI, even without intractable epilepsy.

Patients were sequenced through exome or genome sequencing and 36 *de novo* germline variants were identified on *H3-3A* or *H3-3B* genes. Five of these variants were detected in two or more unrelated patients. At the protein level, the variants identified corresponded to 33 amino acid replacements in the H3.3 protein sequence.

From these variants, only one (*H3-3A* c.362T>A p.M121K, 1 incidence) was found in the large database of controls: gnomAD (Karczewski et al., 2020), which contains the genetic variants of more than 140,000 exomes and genomes. As expected, in the general population both genes have a very low rate of missense variants. The gnomAD missense Z score for H3-3A is 3.16 and for H3-3B is 2.88 (>2 is significant) (Karczewski et al., 2020).

5.1.3. Understanding the effect of causative variants

The pathogenicity of these variants is likely to result from different mechanisms, as they are found throughout the entire H3.3 coding sequence.

Of particular interest, two of the variants in the patient cohort (S32F and G91R) are located in the residues that differentiate H3.3 from the canonical H3. Both S32 and G91 residues are essential for proper recognition of H3.3 by other proteins. Mutagenic analysis in yeast shows that mutations at G91 prevents H3.3 specific chaperones DAXX and UBN1 from binding (Elsässer et al., 2012; C. P. Liu et al., 2012; Ricketts et al., 2015). Interestingly, S32 is required for recognition of H3.3 by ZMYND11 (R. Guo et al., 2014; Wen et al., 2014). Mutations in ZMYND11 cause an autosomal dominant neurodevelopmental phenotype similar to that seen in our patient cohort (Coe et al., 2014; Moskowitz et al., 2016), and in fact, variant G35V of our dataset, has been shown to disrupt ZMYND11 binding (Wen et al., 2014).

Other variants may disrupt histone octamer formation, nucleosome sliding, chaperone binding based on mutagenic analysis of both H3 and H3.3 in model organisms (Johnson et al., 2015; Matsubara, Sano, Umehara, & Horikoshi, 2007; Norris, Bianchet, & Boeke, 2008).

Even though these are the first germline variants associated with H3, germline variants in H1 and H4 with similar features have been reported. The overgrowth and neurodevelopmental delay associated with Rahman syndrome are caused by truncating variants in H1.4 encoded by *HIST1H1E* (Tatton-Brown et al., 2017). Recently two specific germline variants in H4, which caused delayed growth and neurodevelopment have been described in two families (Tessadori et al., 2017). In addition, there are many neurodevelopmental disorders associated with the histone lysine methylases and demethylases (Faundes et al., 2018).

The impact of variants on histone PTMs

It was hypothesized that other missense variants in the cohort could induce epigenetic dysregulation of histone PTMs. To quantify this dysregulation, cells from patients were obtained along with matched controls to extract their histones and analyze them by nanoLC-MS/MS as previously described (Sidoli & Garcia, 2015). This experiment showed significantly altered histone PTMs in patients (Figure 5.1).

Altered PTMs may affect one or more of the multiple functions they carry out in the nucleosome, including: chromatin state, mitotic initiation, protein-chromatin interactions and gene expression; or may impact on the recognition of H3.3 by histone chaperones and its incorporation into the nucleosome (Burgess & Zhang, 2013; Chang et al., 2015; Crosio et al., 2002; Hake & Allis, 2006; Hake et al., 2005; Lau & Cheung, 2011; Sawicka & Seiser, 2012; Schulmeister, Schmid, & Thompson, 2007; Van Hooser, Goodrich, David Allis, Brinkley, & Mancini, 1998).

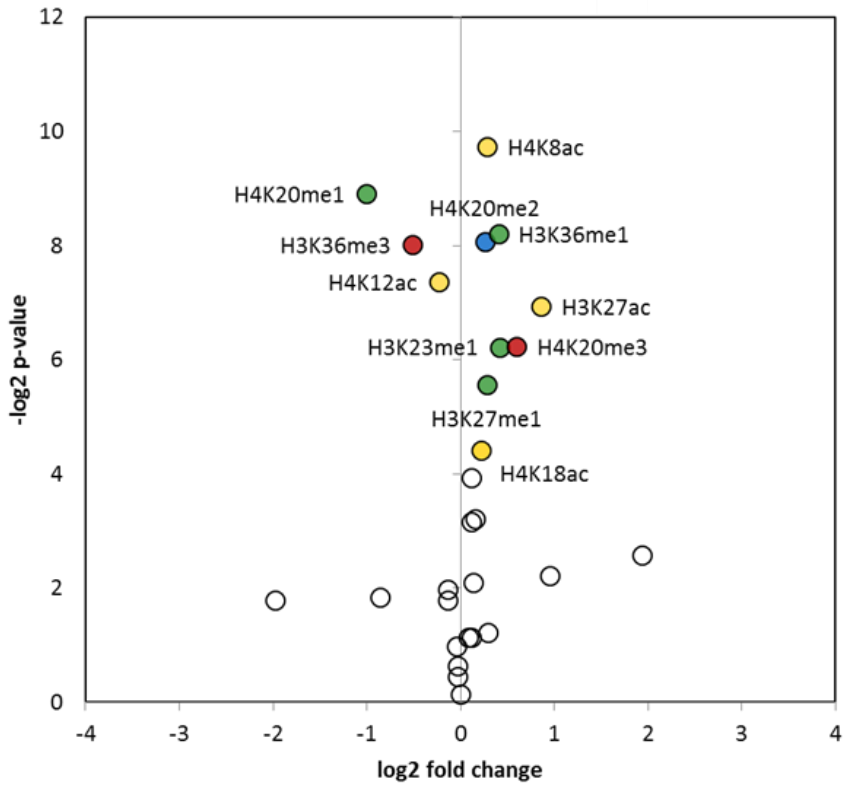


Figure 5.1 ***Volcano plot showing the significantly altered histone PTMs in patients versus controls.*** *Adapted from Bhoj et al., 2020.*

Impact of variants on early development

Although all patients share a common phenotype of developmental delay, only some of them developed major congenital malformations, like cardiac and cranial anomalies. It was decided that further study of the specific local dysregulation of development may lead to insights into the transcriptional control in the developmental processes underlying these anomalies.

This possibility was explored using a previously reported dominant zebrafish model with the equivalent variant of the human D124N and patent craniofacial abnormalities (Cox et al., 2012). Interestingly, this heterozygous variant replicates the dominant inheritance observed in humans and is only

two amino acids away from a variant identified in an affected patient (Q126R). Further investigation of this zebrafish model also revealed a defect in foxd3-positive neural crest-derived glia, as well as melanocytes and xanthophores. The loss of these cell types may relate to the hypomyelination phenotype that is noted on the brain MRIs of over one-third of the cohort.

Impact of variants on the H3.3 turnover rate

Another way that histone mutations might disrupt normal cell physiology is via altered histone turnover rates. Quantification of overall histone H3.3 protein levels did not reveal significant differences in patient versus control cells, just a slight increase in the ratio of H3.3 to H3 in patient cells compared to controls was found by Western blot. This observation may suggest that the mutant histones favor a dominant negative effect of the mutations rather than a loss-of-function effect.

Although the exact mechanism of the cellular pathology in these patients is unclear, H3.3 is vital for normal neurologic functioning. A recent study (Maze et al., 2015), showed that H3.3 begins to replace H3.1 and H3.2 in post-natal mouse and human brain in a time-dependent manner and displaces these canonical H3 almost completely in adulthood. The important role of H3.3 over time may explain the unique neurodegenerative phenotype. Mice with decreased H3.3 expression in the hippocampus have impaired long-term memory. Humans with major depressive disorder have increased percentages of H3.3 in the nucleus accumbens, which is modulated by antidepressant therapy (Lepack et al., 2016).

Impact of variants on gene expression

Together, these data suggest that mutant histones can be incorporated into the nucleosome, cause important local deregulation of chromatin state and

modestly alters the global control of PTMs. Thus, local chromatin changes may be induced by mutant histone deposition. H3.3 is known to have roles in diverse functions, including gene expression and repression, chromatin stability, DNA damage repair, and differentiation. These mutant proteins and their aberrant PTM states could significantly disrupt any of these processes to lead to the observed phenotype.

To evaluate which biological pathways were differentially perturbed in the patients, a RNA-Seq experiment was carried out on fibroblast cells derived from patients and matched controls. It was found a total of 323 genes to be differentially expressed with at least 2-fold change. Of these genes, 166 were upregulated and 157 were downregulated in cases. Differentially expressed genes were analyzed and it was found a significant enrichment for upregulated genes in mitotic cell cycle process, mitotic nuclear division, cell division and many other mitosis-related processes.

Impact of variants on cell proliferation

To assess if upregulation of mitosis-related genes alters cells proliferation, the cellular proliferation capacity of five patient fibroblast lines was quantified and compared to six matched control fibroblast lines. Patient lines had increased cell proliferation, notably at 72 and 96 hours. Furthermore, all five patient lines shared similar viability to the six control lines. Cell cycle analyses showed that *H3-3A* G91R and T46I had a similar cell cycle profile to the control lines, while *H3-3A* R18G showed a decrease in cells in G1 phase and an increase in cells in S phase compared to all three control lines. It is possible that this increased proliferation is tied to the oncogenic mechanism of the somatic mutations, which has not been fully elucidated (Bjerke et al., 2013).

Elucidating the impact of variants with in silico tools

In this project, our group was charged with the computational characterization of the novel missense H3.3 variants, in terms of both *in silico* predictions and structural/mechanistic analyses. These results constitute the bulk of this chapter and are divided into two blocks. First, the obtention of the *in silico* predictions for the H3.3 variants, followed by a description of the inconsistencies found between the bioinformatic and the clinical evidence, and our explanation of these disagreements. Second, the analysis of the mutational impact using the three-dimensional structures of H3.3 in different contexts (nucleosome, interaction with epigenetic regulators and histone chaperones) and biophysics computations.

5.2. Materials and Methods

5.2.1. Patient cohort and identified variants

The patient cohort is composed of 42 unrelated patients from the Children's Hospital of Philadelphia (USA) with core phenotypes of progressive neurologic dysfunction and congenital anomalies, but no malignancies yet.

Patients were characterized by exome or genome sequencing and *de novo* germline missense variants were identified in *H3-3A* and *H3-3B* genes. These two genes encode for the same protein H3.3. A set of 33 unique missense variants were identified in this protein: R9C, R9G, R9S, S11P, G14R, A16G, R18G, T23I, A30P, A30T, S32F, G35V, K37E, H40R, H40Y, T46I, L49R, L62R, D78N, D82H, R84C, G91R, N109S, I113L, I113V, V118L, M121I, M121K, M121V, P122L, P122R, Q126R and R129C.

5.2.2. Pathogenicity prediction of variants

We estimate the pathogenicity of these missense variants with 12 widely used predictors (Table 5.1). I obtained the predictions using either their respective webservers or dbNSFP (X. Liu, Wu, Li, & Boerwinkle, 2016), a database developed for functional prediction and annotation of all potential non-synonymous single-nucleotide variants in the human genome. Afterwards, we unified the predicted variant class of the different tools into two main types: pathogenic and neutral, as described in Table 5.1.

Pathogenicity predictor	Pathogenic classes	Neutral classes	Reference
CADD	deleterious	neutral	(Kircher et al., 2014)
FATHMM	damaging	tolerated	(Shihab et al., 2012)
MutationTaster2	disease	polymorphism	(Schwarz, Cooper, Schuelke, & Seelow, 2014)
MutPred2	pathogenic	benign	(B. Li et al., 2009)
PANTHER	probably damaging, possibly damaging	probably benign	(Thomas et al., 2006)
PolyPhen-2 (HumDiv)	probably damaging, possibly damaging	benign	(I. A. Adzhubei et al., 2010)
PON-P2	pathogenic	neutral	(Niroula et al., 2015)
PROVEAN	deleterious	neutral	(Choi, Sims, Murphy, Miller, & Chan, 2012)
REVEL	disease	neutral	(Ioannidis et al., 2016)
SIFT	deleterious	tolerated	(Kumar et al., 2009)
SNAP	effect	neutral	(Bromberg & Rost, 2007)
VEST3	pathogenic	neutral	(Carter, Douville, Stenson, Cooper, & Karchin, 2013)

Table 5.1 ***List of pathogenic predictors used for the estimation of the functional impact of variants.*** *The original predicted variant classes are categorized into two classes: pathogenic or neutral.*

5.2.3. Three-dimensional structures of H3.3

We retrieved the H3.3 structures used in this work from its UniProt record P84243. In all cases, the H3.3 protein appears incomplete. Depending on the interaction partners, the structures retrieved can be divided into three groups (Table 5.2): H3.3 in the nucleosome, H3.3 tails interacting with epigenetic regulators and H3.3 interacting with histone chaperones.

PDB	H3.3 sequence	Structure context (interaction partners)	Reference
3ASL	2-10	Epigenetic regulator UHRF1	(Arita et al., 2012)
3AV2	39-135	Nucleosome Histones, DNA	(Tachiwana et al., 2011)
3JVK	13-16	Epigenetic regulator BRD4	(Vollmuth, Blankenfeldt, & Geyer, 2009)
3MUK	22-28	Epigenetic regulator BRD4	(Vollmuth & Geyer, 2010)
4GNF	2-10	Epigenetic regulator NSD3	(He, Li, Zhang, Wu, & Shi, 2013)
4GUS	2-21	Epigenetic regulator KDM1B	(Fang et al., 2013)
4H9N	38-135	Chaperone DAXX	(Elsässer et al., 2012)
4N4I	30-40	Epigenetic regulator ZMYND11	(Wen et al., 2014)
4QQ4	2-10	Epigenetic regulator MORC3	(Y. Liu et al., 2016)
4TMP	4-12	Epigenetic regulator MLLT3	(Y. Li et al., 2014)
4U7T	2-11	Epigenetic regulator DNMT3A	(X. Guo et al., 2015)
5BNV	59-135	Chaperone MCM2	(Huang et al., 2015)
5BNX	59-135	Chaperone MCM2	(Huang et al., 2015)
5DWQ	13-22	Epigenetic regulator CARM1	(Boriack-Sjodin et al., 2016)

5JA4	58-135	Chaperone MCM2	(Saredi et al., 2016)
5JJY	30-43	Epigenetic regulator SETD2	(Yang et al., 2016)
5JLB	30-43	Epigenetic regulator SETD2	(Yang et al., 2016)

Table 5.2 ***H3.3 structures used in this work.*** *The numbering of the H3.3 sequence follows the human genetics standards and includes the first residue methionine.*

The retrieved H3.3 structures were used for two analyses. First (section 5.2.4), to study the pattern of interatomic contacts of the native amino acid either within the same H3.3 monomer or with other partners. These partners are diverse and include DNA, histones (H2A.1, H4), epigenetic regulators and chaperones. In parallel, (section 5.2.5), we computed the changes in monomer stability upon mutation and in binding affinity of the complexes of H3.3 with the DNA or other proteins.

5.2.4. Interatomic contacts at the native locus

For each residue, I computed the network of interatomic contacts of the native amino acid. These networks were obtained using the RING software (Piovesan, Minervini, & Tosatto, 2016). This software computes all the atom-atom interactions between the atoms in a protein molecule or between molecules. The program is executed online (http://protein.bio.unipd.it/ring/) and provides a complete result that can be downloaded, for local processing.

In our case, we applied the following protocol:

i. Retrieve the PDB(s) where the native amino acid appears.
ii. Execute RING with default parameters, except for the option Interaction type which was set to All.
iii. Extract the interactions of the native residue from RING's output.

iv. Organize these interactions into three groups: intra-monomer (within the same H3.3 monomer and separated by > 2 residues in sequence), inter-monomer (between H3.3 and other proteins), and H3.3-DNA (at a distance below 5.5 Å).

5.2.5. Protein stability and binding affinity change upon mutation

For each variant, I computed its impact on the H3.3 monomer stability and on the binding affinity of H3.3-protein (PPI) and H3.3-DNA (PDI) interactions using the package mCSM (Pires, Ascher, & Blundell, 2014).

These protein stability changes and binding affinity changes of PPI and PDI were obtained as follows:

i. Localization of the variant's residue in the PDB structures.
ii. Variant's impact on H3.3 monomer stability: extraction of the H3.3 monomer of the nucleosome structure and computation of the protein stability change upon mutation for each variant with mCSM.
iii. Variant's impact on H3.3 protein-protein interactions: for each PDB separately, extraction of the H3.3 monomer and all of the interacting proteins (histones, epigenetic regulators and chaperones), and computation of the binding affinity change of the PPI upon mutation.
iv. Variant's impact on H3.3-DNA interactions: extraction of the H3.3 monomer and DNA fragment of the nucleosome structure and computation of the binding affinity change of PDI upon mutation.

5.3. Results and Discussion

5.3.1. A sequence-based view of the variants

In this section, I analyse the distribution of the patients' variants along the protein sequence.

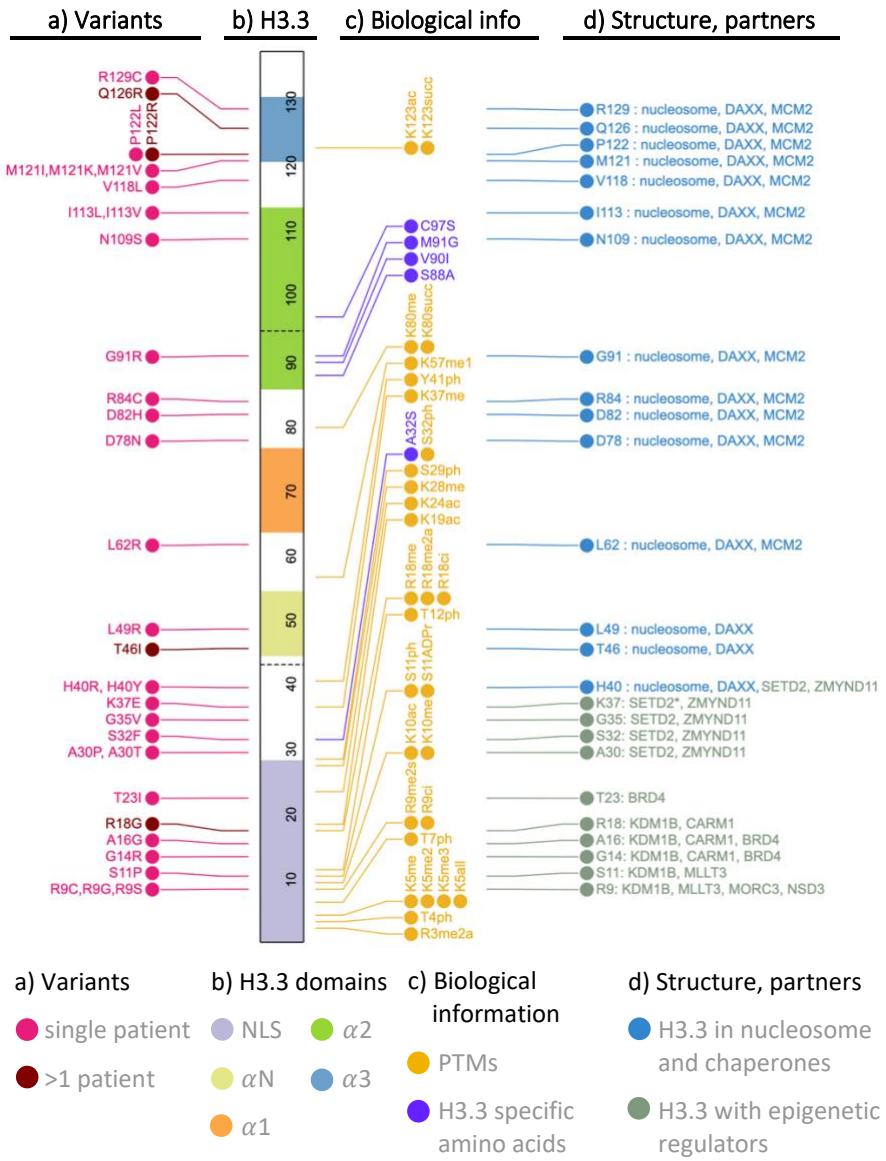


Figure 5.2 ***Dataset of H3.3 variants.*** *a) Variants used in this work colour coded according to the number of patients carrying each variant, pink: single patient, brown: multiple patients. b) H3.3 sequence showing its main structural domains: nuclear localization sequence (NLS), α-helix N, α-helix 1, α-helix 2, α-helix 3. c) Residues susceptible of PTMs are annotated in yellow along with their potential PTMs extracted from UniProt. Specific amino acids of H3.3 compared to the canonical H3 are displayed in blue. d) Native residues mapping to PDB structures are annotated in blue if the structure is a nucleosome or involves chaperones (DAXX, MCM2), in green if H3.3 interacts with an epigenetic regulator (SETD2, ZYMD11, BRD4, KDM1B, CARM1, MLLT3, MORC3, NSD3).*

The 33 identified variants mapped to a total of 25 residues distributed through the protein sequence (Figure 5.2a), affecting both the histone tail and core region of H3.3.

Analysis of the variants in the histone tail points to a possible effect on the addition of PTMs by epigenetic regulators. In fact, many of these variants may destroy or create targets for epigenetic regulators because either the native or mutated amino acids are classical targets of epigenetic modifications in histones. Histone PTMs are covalent attachments of methyl or acetyl groups to lysine and arginine amino acids, or phosphorylations of serine or threonine amino acids. Of the variants located in the histone tail (amino acids 1 to 89), 71% of them have one of these amino acids (lysine, arginine, serine or threonine) as the native or mutated amino acid. Furthermore, seven of these variants, R9C/G/S, S11P, R18G, S32F, K37E, occur in residues known to undergo PTMs (Figure 5.2c) which may induce epigenetic dysregulation of histone PTMs.

From looking at the sequence, we also learn that the impact of some variants may have a deep effect, because they affect the residues defining the identity of H3.3. This is the case of G91R and S32F, which affect amino acids differentiating H3.3 from the canonical H3 (Frank et al., 2003), and are

essential for the proper recognition of H3.3 by other proteins. For example, mutagenic analysis in yeast shows that mutations at G91 prevent the binding of H3.3 specific chaperones DAXX and UBN1 (Elsässer et al., 2012; C. P. Liu et al., 2012; Ricketts et al., 2015). Analogously, mutations in S32 are required for recognition of H3.3 by the chromatin reader ZMYND11 (R. Guo et al., 2014; Wen et al., 2014).

5.3.2. Beyond sequence-based features: an estimation of variant pathogenicity by bioinformatic predictors

We used 12 widely employed bioinformatics pathogenicity predictors (see Material and Methods section 5.2.2) to generate *in silico* evidence of the nature of H3.3 variants. Because these tools approach the prediction problem from slightly different ways, their terminology may differ. For our analyses, we unified the predicted classes into two categories: pathogenic and neutral (see Material and Methods section 5.2.2).

In Figure 5.3, I represent the whole set of predictions as a heatmap, with variants predicted as pathogenic and neutral shown in red and green, respectively. Variants are sorted vertically by the number of pathogenic predictions they received, and predictors are sorted horizontally by the number of pathogenic predictions, both in decreasing order. As we can see, the predictors can be divided into 3 groups: (i) those methods predicting all or almost all the variants as pathogenic (from PANTHER to MutPred2), (ii) those predicting from half to all of the variants as neutral (from PolyPhen2 to FATHMM), and (iii) those not giving predictions (PON-P2). The evidence provided by the first group of predictors is consistent with the experimental information. However, for the second group of predictors, the evidence provided is counter-intuitive: they estimate most or all of the variants as neutral. I address this contradiction in the next section 5.3.3.

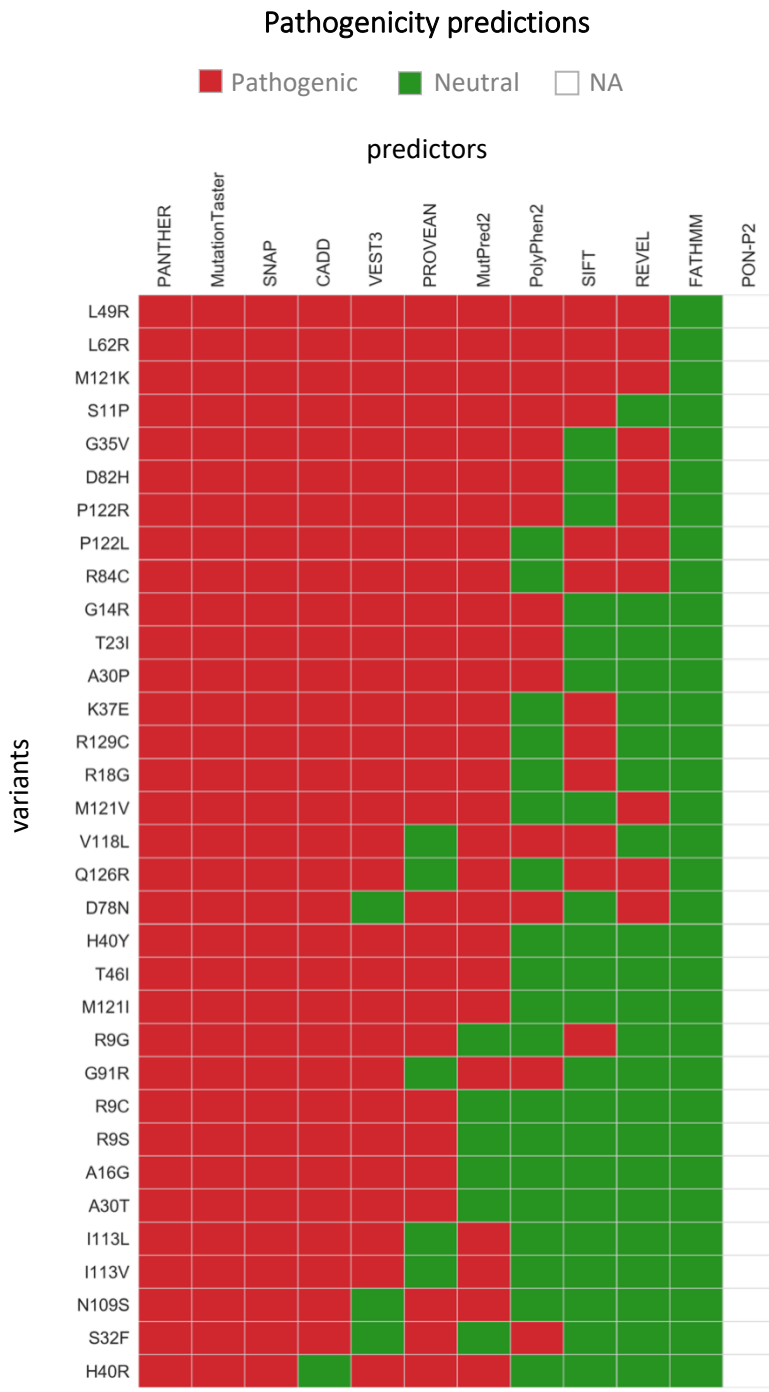


Figure 5.3 ***Pathogenicity predictions of variants***. *From left to right and top to bottom, predictors and variants are ordered by decreasing number of pathogenic predictions.*

5.3.3. Understanding the contradictory results of some relevant bioinformatic predictors

As we have seen, 4 out of the 12 pathogenicity predictors used (PolyPhen-2, SIFT, FATHMM, and REVEL) annotate the majority of the variants as neutral, in contrast with the experimental results. Here, I analyse these predictions, to see if we can understand this apparent contradiction.

PolyPhen-2

PolyPhen-2 (I. A. Adzhubei et al., 2010) combines sequence conservation, structural information and annotation of residues to predict the class of protein sequence variants. Although with an accuracy generally above 75% (de la Campa, Padilla, & de la Cruz, 2017), in our case it erroneously predicted 58% of the variants as neutral. Previous work from our group (Colobran et al., 2016), shows that the score of this predictor may be sometimes biased by one of the input features. On this basis, we decided to check whether this was the case for the H3.3 variants in this study.

No clear trend was found for the sequence-based predictive features; however, for structure-based features the situation was different. Our attention was immediately attracted by the fact that one of these key features, the normalized accessible surface area (acc_normed), was obtained from the DSSP database (Touw et al., 2015). Acc_normed has a significant discriminant power between pathogenic and neutral variants (Figure 5.4a) (Ancien, Pucci, Godfroid, & Rooman, 2018); in fact, it is one of the 11, out of 32, features selected to build PolyPhen-2. For this reason, biases affecting acc_normed may result in systematic predictions errors. The fact that PolyPhen-2 uses acc_normed values retrieved from DSSP points to the possibility of this type of error, because DSSP does not always take into account the full 3D environment of a protein as presented in its PDB.

We tested this possibility for variants affecting H40. Visual inspection of the nucleosome structure (Figure 5.4b) showed that residue H40 is buried in the H3.3-DNA interface, with almost no access to the solvent. In our dataset, two variants are located in this position: H40R and H40Y, both of them predicted as neutral by PolyPhen-2. The acc_normed value used for these predictions is 0.84 (indicating high accessibility to the solvent), which is more frequent for neutral than for pathogenic variants (Figure 5.4a), and is coherent with the neutral predictions of PolyPhen-2. However, the large value of acc_normed is in contradiction with the visual analysis.

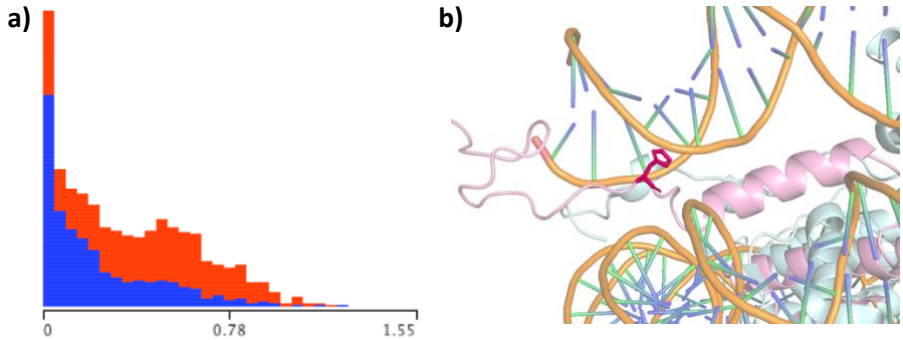


Figure 5.4 ***a) Distribution of normalized accessibility areas (acc_normed) of neutral (red) and pathogenic (blue) variants used in PolyPhen-2 training.*** *Adapted from (I. A. Adzhubei et al., 2010).* ***b) Localization of H40 (magenta) in H3 (pink), along with other histones (cyan) and DNA (coloured double helix) in the nucleosome.***

To clarify this problem, we computed acc_normed using the program dr_sasa (Ribeiro, Ríos-Vera, Melo, Schüller, & Valencia, 2019), obtaining a value of 0.21, more coherent with the visual analysis and more frequently observed for pathogenic than for neutral variants (Figure 5.4a). This result suggests that the PolyPhen-2 neutral predictions for H40 may be a systematic error resulting from the use of DSSP acc_normed values which do not take into account the DNA molecule for the calculations of accessibility. Therefore, this will affect any H3.3 variant in contact with the DNA.

SIFT

SIFT (Kumar et al., 2009) is a pathogenicity predictor based on the use of MSAs of homologous proteins to obtain position-specific scoring matrices to estimate the class of a variant. In our case it predicts 36% of the variants as neutral. To clarify this paradox, we centered our analysis on the MSA utilized by the SIFT server for its predictions, finding that SIFT discards proteins with a sequence identity above 90% by default. This filter, conceived to eliminate database noise, ignores a crucial feature of histones: they are highly conserved. We postulated that this misrepresentation of sequence conservation was partly responsible of the incorrect SIFT predictions. To test this idea, we forced the SIFT predictor to include in the MSA 121 sequences from H3, H3.1, H3.2 and H3.3 proteins with sequence identities above the 90% threshold. As a result, the SIFT scores of the variants shifted towards more pathogenic values and 6 variants previously predicted as neutral became pathogenic, increasing the total number of correct predictions from 36% to 54% (Figure 5.5).

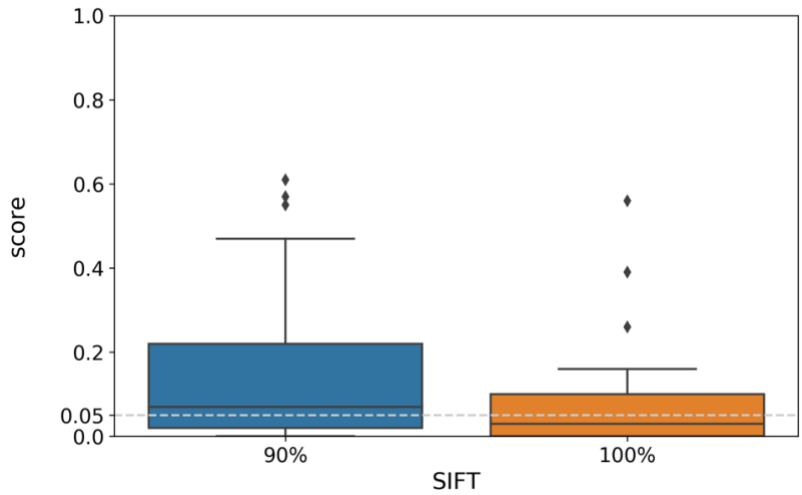


Figure 5.5 ***Boxplot of SIFT scores calculated with default parameters (blue) and after allowing the addition of proteins with a sequence identity >90% in the MSA (orange).*** *The cutoff of SIFT is marked as a dotted grey line.*

FATHMM

FATHMM predicts 100% of the H3.3 variants as neutral (Figure 5.3). This type of error strongly indicates the possibility of a bias in FATHMM predictions. In fact, FATHMM results have been very controversial, as explained in a recent article (Grimm et al., 2015), because of its tendency to learn the compositional population of pathogenic and neutral variants of the training dataset rather than the features discriminating them. Given the absence of H3.3 germline pathogenic variants in current databases, we suspect that this may be the reason why FATHMM predicts all the variants as neutral. In fact, if we predict these variants with an unweighted version of FATHMM that lacks this compositional weight, 11 variants are predicted as pathogenic.

REVEL

REVEL (Ioannidis et al., 2016) is a metapredictor that among other features, uses PolyPhen-2, SIFT and FATHMM scores as part of its input. Since these predictors are biased towards neutral predictions, we hypothesize that this may be the cause behind REVEL's neutral predictions.

5.3.4. Back to basics: *in silico* biophysics estimation of the functional impact of variants

As we have seen in previous sections, bioinformatics pathogenicity predictors display mixed success rates. This is worrisome in the case of PolyPhen-2, SIFT and REVEL, because they are amongst the most broadly utilized tools in the annotation of variants. In this situation, we decided to go beyond the use of pathogenicity predictors and utilize a more fundamental approach, comprising (i) structure analysis and (ii) biophysics-based models of the impact of sequence variants on protein stability and disruption of protein-protein and protein-DNA interactions.

Structure analysis of the effect of variants

Here, I present a characterization of the 3D environment of the native residues mutated in our patient cohort. This study provides a direct view of the amount of atomic interactions affected by the mutation, introducing a mechanistic component in the interpretation of the variants' impact.

The 3D structures used (section 5.2.3) are: (i) the nucleosome for the intra-monomer, H3.3-DNA and H3.3-histone interactions, (ii) chaperones for the H3.3-chaperone interactions and (iii) epigenetic regulators for the H3.3-epigenetic regulator interactions.

For each native residue, we collected its pattern of contacts (section 5.2.4) and classified them into the following groups: (i) intra-monomer contacts, (ii) contacts with DNA, (iii) contacts with other histones, (iv) contacts with chaperones and (v) contacts with epigenetic regulators.

As we can see in Table 5.3, the number of interatomic contacts varies substantially, but we can establish two main groups (Figure 5.2d). One is formed by those variants with native residues between 40 and 129, which are located in the centre and C-terminal end of H3.3. They mostly contact with DNA and other histones in the nucleosome, or histone chaperones. The second is formed by residues 9-40, which are located in the histone N-tail and mostly contact with epigenetic regulators.

	Intra-monomer	DNA	Histone		Chaperone		Epigenetic regulators							
residue	**H3.3**	**DNA**	**H2A.1**	**H4**	**DAXX**	**MCM2**	**BRD4**	**CARM1**	**KDM1B**	**MLLT3**	**MORC3**	**NSD3**	**SETD2**	**ZMYND11**
9									11	4	5	10		
11									4	1				
14								30	2					
16							1	3						
18														
23							1							
30														1
32													3	9
35													3	1
37													37	
40		96			2								6	4
46	3	26			2									
49	3	1	5		1									
62	4			7										
78	1													
82		5												
84		105		3	2	7								
91	3			2										
109	5		1	4	7									
113	4				9									
118		24	2	1										
121	6	15		2	2									
122	5			4	2									
126	8			1	4									
129	8			2										

Table 5.3 ***Summary of interatomic contacts between H3.3 variants' residues and intra-monomer, DNA, histones, chaperones and epigenetic regulators.***

For the first group, the pattern of interatomic contacts is summarized in Figure 5.6, focusing on the nucleosome structure. This suggests two possible scenarios for the impact of the variants. One is the disruption of the H3.3-DNA interaction, because the native residue is involved in a large number of contacts with the DNA. For instance, this would be the case for variant R84C, where the arginine residue penetrates the DNA minor groove. In the second scenario, variants are more likely to disrupt the histone octamer, either because they affect the intra-monomer contacts of H3.3 (e.g. Q126R) or because they alter the interaction with other histones (e.g. L49R).

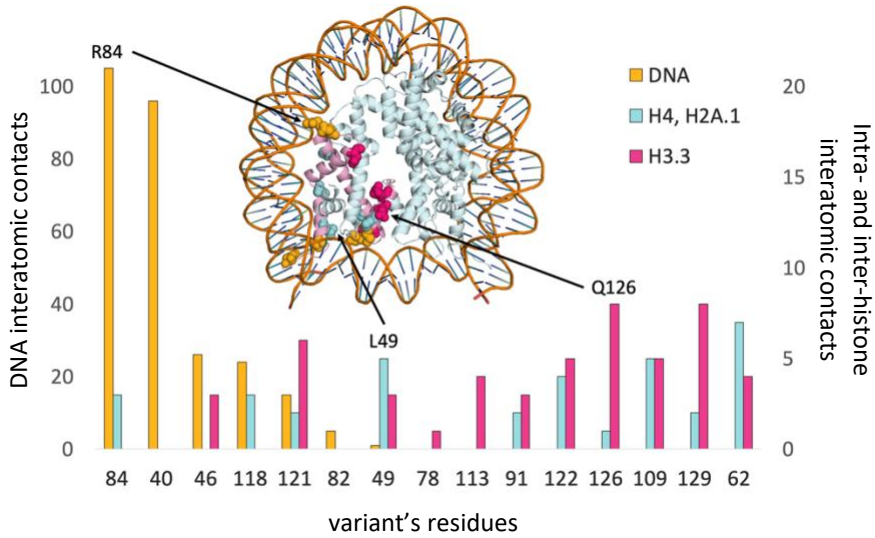


Figure 5.6 ***Bar plot of interatomic contacts between variant's residues and nucleosome atoms.*** *On top, I show the nucleosome structure with only one monomer of H3.3 colored (pink) for clarity, with the variants' native residues mapped and shown with spheres. Residues are colored according to the highest number of type of contacts they have. At the bottom, a bar plot with the number of contacts between variant's residues and DNA (yellow), histones (cyan) and intra-monomer H3.3 (magenta).*

In summary, the variants in this group are likely to affect through different mechanisms, either the formation or the stability of the nucleosomes containing H3.3.

The second group of variant locations (residues 9-40) belong to small H3.3 fragments (pink) that are found in complex with epigenetic regulators (green) (Figure 5.7, left). The number of inter-protein contacts at the variant locus are summarized in a Sankey diagram (Figure 5.7, right). They vary substantially even for the same residue. For example, residue 40 has 33 contacts with SETD2 and only 4 contacts with ZMYND11 epigenetic regulator.

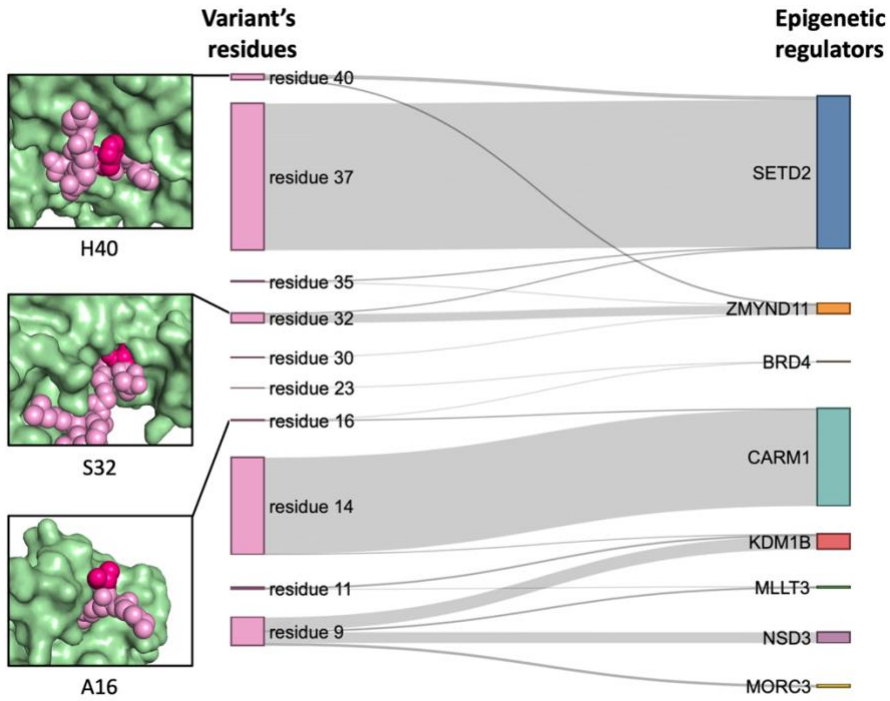


Figure 5.7 ***Sankey plot of interatomic contacts between variant's residues and epigenetic regulators.*** *To the left, structural detail of three native residues A16, S32 and H40 (magenta) within the H3.3 N-tail (pink) and the epigenetic regulator BRD4, ZMYND11 and SETD2 (green), respectively. The thickness of the grey bands represents the amount of contacts between native residues and epigenetic regulators.*

We find that residues R9, G14, S32, K37 and H40 are involved in more than 10 contacts across epigenetic regulators, suggesting that their mutation may disrupt one or more biologically relevant interactions. For the remaining residues, the number of inter-protein contacts decreases rapidly, limiting our

ability to interpret the mutation impact. For example, visual analysis of the H3.3-BRD4 complex shows that A16 barely participates in the complex between both proteins. In fact, the contact analysis of A16 shows that it has only one interatomic contact with BRD4. Consequently, destabilization of the H3.3-BRD4 complex is a less likely explanation for the impact of variants in residue A16.

Apart from disrupting H3.3-epigenetic interactions, there is another mechanism for the impact of variants in the histone tail. These variant may affect the inter-nucleosome packing, an important interaction in which H3.3 tails are involved (Pepenella, Murphy, & Hayes, 2014).

In summary, most of the variants in the second group are likely to affect the interaction between H3.3 and epigenetic regulators with consequences that will depend on the biological role of each complex. Or, they may loosen chromatin structure by disrupting inter-nucleosome packing.

Protein stability change of H3.3 monomer upon mutation and disruption of PPI and PDI interactions upon mutation

Our previous analysis uses interatomic contact networks to provide an intuitive, mechanistic view of the disruptive effect of the variants in our dataset. However, this view is limited in the sense that it does not provide a biophysical quantification of this effect, which gives the ultimate explanation of the molecular impact of mutations.

In this section, we follow a biophysically based approach to quantify this molecular impact, computing the change in protein stability upon mutation, as well as, the change in binding affinity between protein-protein and protein-DNA interactions (section 5.2.5). We follow the convention of the mCSM package in which values under zero indicate a reduction in protein stability or binding affinity.

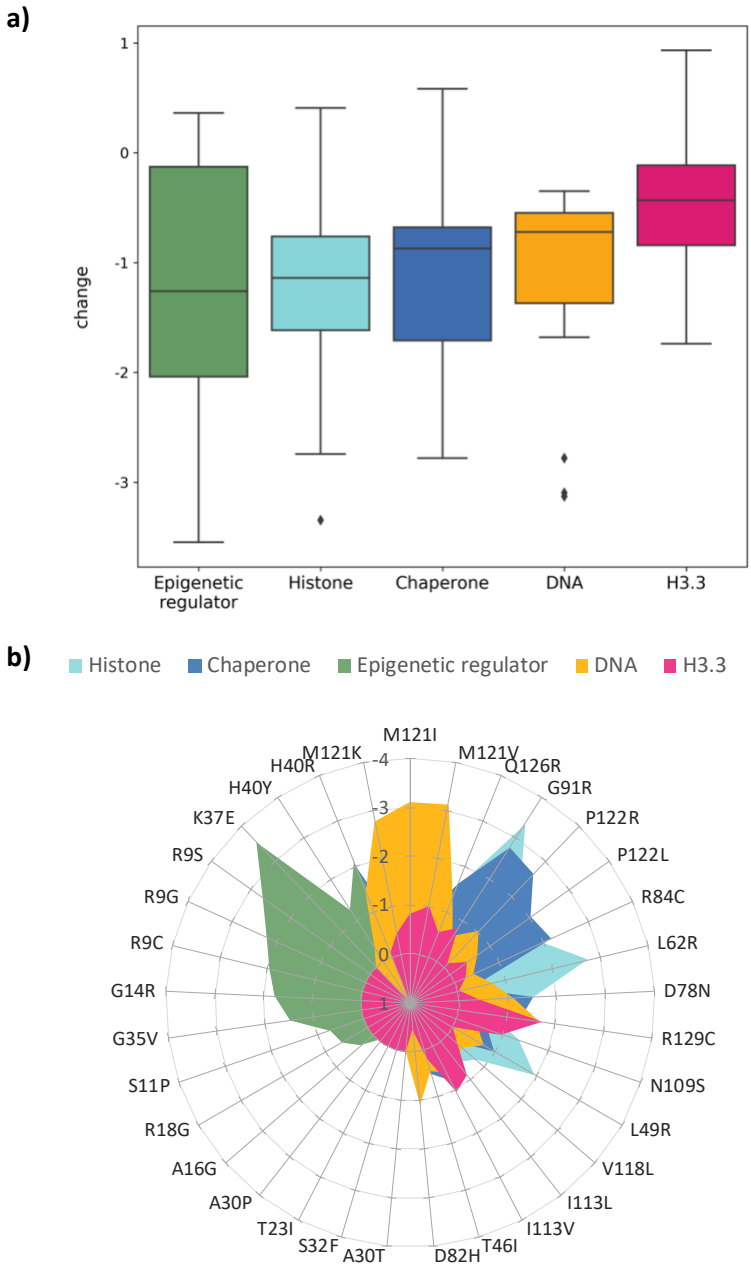


Figure 5.8 ***a) Box plot of changes in binding affinity upon mutation in the interactions between H3.3 and epigenetic regulators, other histones, chaperones and DNA; changes in protein stability upon mutation in the H3.3 monomer are also included (magenta boxplot to the right). b) Radar plot of these changes at the variant level.*** *Note: negative values correspond to disruptive effects of the variants.*

In Figure 5.8, I show the results for the following situations: (i) change in stability for the H3.3 monomer; change in binding affinity for the (ii) H3.3-DNA, (iii) H3.3-chaperones, (iv) H3.3-H4/H2 histones, and (iv) H3.3-epigenetic regulators interactions.

In Figure 5.8a, we can see that regardless of the situation considered (monomer stability, H3.3-DNA interaction, etc.), the H3.3 variants populate negative energy changes, indicating a general disruptive trend, in accordance with their pathogenic nature. In Figure 5.8b, we confirm this result, with a variant-level view of the results.

Moreover, we distinguish a set of variants highly disruptive (< -2) of different complexes: M121I, M121K and M121V disrupt H3.3–DNA interactions; L62R, G91R and P122R disrupt H3.3-histone interactions; R84C, G91R, P122L and P122R disrupt H3.3-chaperones interactions; and R9G, R9S, K37E and H40R disrupt H3.3-epigenetic regulators interactions.

5.4. Conclusions

In this chapter, I address the *in silico* characterization of H3.3 variants associated to a novel neuropediatric disorder using both bioinformatics and biophysics computations.

Our results show that, for this case, bioinformatic pathogenicity predictors may have an incorrect behavior due to the uncommon characteristics of histones. In particular, their large conservation degree which reflects their important functional role, is ignored by a renowned predictor such as SIFT. Moreover, PolyPhen-2, which bases its predictions on the use of structural information, can be misled by solvent accessibility values that ignore the presence of DNA. We find that once these technical problems are addressed, the results of bioinformatic predictors agree better with the results of the functional experiments.

In this context, the use of biophysics methods becomes very useful. Apart from providing independent evidence, structure-based biophysic tools illustrate the functional impact of variants in a more concrete way, pointing to possible mechanisms of action. In our case, they show how variants can affect different biological processes by disrupting different molecular complexes involving H3.3.

6. Conclusions

The conclusions of the present thesis are the following:

- Three sequence-based properties, Entropy, PSSM and Blosum62, can be combined to estimate the molecular impact of missense variants on the HDR function of BRCA1 and BRCA2, as measured in the homonym assay.
- The BRCA1- and BRCA2-protein specific predictors developed can be used, with moderate success, to identify variants with increased risk of HBOC, by predicting the variants' molecular impact on the HDR function of these proteins.
- The BRCA1- and BRCA2-protein specific predictors have a balanced sensitivity and specificity around 0.8, and an accuracy of 0.75 and 0.857 respectively, competitive with that of widely used predictors in the field.
- In relation with the ENIGMA challenge that took place in the 5th CAGI experiment, our predictors have an accuracy comparable or better than that of the standard predictors in the field, being able to predict the biased composition of the CAGI dataset, which is enriched in neutral variants.
- The BRASS website makes available to the scientific community a user-friendly site to access a novel family of pathogenicity predictors for missense BRCA1 and BRCA2 variants.
- Regarding the novel pediatric neurologic disorder caused by pathogenic variants in histone H3.3, it was found that some methods predict them as neutral. The study of this inconsistency revealed that this is mostly due to technical issues related to the automatization of these tools.

- Structural analysis of the nucleosome suggests that variants falling in this region have a molecular impact on the H3.3's function throughout various mechanisms including the disruption of (i) the H3.3 intra-monomer contacts, (ii) the contacts with other histones shaping the nucleosome, and (iii) the contacts with the DNA wrapping the histone octamer.
- Structural analysis of the N-tail of H3.3 in complex with epigenetic regulators suggests that some of the variants falling there may disrupt this interaction, thus affecting the associated biological process.
- Analysis of (i) the change upon mutation in protein stability of the H3.3 monomer, (ii) the change in binding affinity of the H3.3-DNA interactions, and the change in binding affinity of the H3.3-protein interactions, shows that the majority of variants have a negative impact in one or more of these properties.

7. Appendix

Appendix 1

Table 7.1 Predictions submitted in the ENIGMA challenge of the CAGI experiment. Here, we provide a list of the four predictions we submitted for the BRCA1 and BRCA2 variants. For each variants, the following information is provided: gene, DNA variant, protein variant, current IARC 5-tier class according to the ENIGMA consortium, predicted IARC 5 class by MLR protocol, predicted IARC 5 class by MLR + AS protocol, predicted IARC 5 class by NN protocol and predicted IARC 5 class by NN + AS protocol. In the last column, the effect of the variant (protein or splicing) is stated, as well as, the no missense variants, such as deletions, are marked with the arbitrary label.

Gene	DNA	Protein	ENIGMA	MLR	MLR+AS	NN	NN+AS	Comments
BRCA1	c.1036C>T	p.Pro346Ser	2	1	1	3	3	protein
BRCA1	c.1075C>T	p.Pro359Ser	2	1	1	3	3	protein
BRCA1	c.1081T>C	p.Ser361Pro	2	1	1	3	3	protein
BRCA1	c.1310A>T	p.His437Leu	2	3	3	3	3	protein
BRCA1	c.131G>T	p.Cys44Phe	5	4	4	3	3	protein
BRCA1	c.1342C>T	p.His448Tyr	2	2	2	2	2	protein
BRCA1	c.134A>C	p.Lys45Thr	2	2	2	3	3	protein
BRCA1	c.1361G>A	p.Ser454Asn	2	1	1	2	2	protein
BRCA1	c.1383T>A	p.Phe461Leu	1	3	3	4	4	protein
BRCA1	c.1396C>T	p.Arg466Trp	2	2	2	3	3	protein
BRCA1	c.140G>A	p.Cys47Tyr	5	4	4	3	3	protein
BRCA1	c.1418A>T	p.Asn473Ile	1	2	2	3	3	protein
BRCA1	c.1423A>T	p.Ser475Cys	2	1	1	3	3	protein
BRCA1	c.1508A>G	p.Lys503Arg	2	1	1	3	3	protein
BRCA1	c.1514A>T	p.Lys505Ile	2	3	3	4	4	protein
BRCA1	c.1534C>T	p.Leu512Phe	1	3	3	4	4	protein
BRCA1	c.154C>A	p.Leu52Ile	2	3	3	3	3	protein
BRCA1	c.1601A>G	p.Gln534Arg	2	1	1	3	3	protein
BRCA1	c.1703C>G	p.Pro568Arg	2	2	2	3	3	protein
BRCA1	c.172C>G	p.Pro58Ala	2	2	2	3	3	protein

Table 7.1 continuation

Gene	DNA	Protein	ENIGMA	MLR	MLR+AS	NN	NN+AS	Comments
BRCA1	c.1756C>T	p.Pro586Ser	2	2	2	3	3	protein
BRCA1	c.1772T>C	p.Ile591Thr	2	2	2	3	3	protein
BRCA1	c.1834A>G	p.Arg612Gly	1	2	2	4	4	protein
BRCA1	c.1846_1848del	p.Ser616del	1	3	3	3	3	arbitary
BRCA1	c.1879G>A	p.Val627Ile	2	1	1	3	3	protein
BRCA1	c.1903A>G	p.Asn635Asp	2	1	1	2	2	protein
BRCA1	c.1927A>G	p.Ser643Gly	1	2	2	3	3	protein
BRCA1	c.2006T>C	p.Met669Thr	2	3	3	2	2	protein
BRCA1	c.2042G>T	p.Ser681Ile	2	1	1	3	3	protein
BRCA1	c.2050C>T	p.Pro684Ser	2	2	2	3	3	protein
BRCA1	c.2060A>C	p.Gln687Pro	2	2	2	3	3	protein
BRCA1	c.2083G>T	p.Asp695Tyr	1	2	2	3	3	protein
BRCA1	c.211A>G	p.Arg71Gly	5	3	5	3	5	splicing
BRCA1	c.2180C>T	p.Pro727Leu	1	1	1	3	3	protein
BRCA1	c.2183G>A	p.Arg728Lys	2	2	2	3	3	protein
BRCA1	c.2245G>T	p.Asp749Tyr	2	2	2	3	3	protein
BRCA1	c.2338C>A	p.Gln780Lys	2	2	2	3	3	protein
BRCA1	c.2346T>A	p.Ser782Arg	2	1	1	5	5	protein
BRCA1	c.2351C>T	p.Ser784Leu	1	2	2	5	5	protein
BRCA1	c.2447A>G	p.His816Arg	2	2	2	2	2	protein
BRCA1	c.2452T>G	p.Cys818Gly	2	3	3	3	3	protein
BRCA1	c.2482G>A	p.Gly828Ser	2	2	2	3	3	protein
BRCA1	c.2503C>T	p.His835Tyr	2	2	2	2	2	protein
BRCA1	c.2518A>T	p.Ser840Cys	3	1	1	2	2	protein
BRCA1	c.2522G>A	p.Arg841Gln	2	1	1	3	3	protein
BRCA1	c.2597G>A	p.Arg866His	1	3	3	4	4	protein
BRCA1	c.2650A>G	p.Thr884Ala	2	1	1	2	2	protein
BRCA1	c.2662C>T	p.His888Tyr	3	2	2	2	2	protein
BRCA1	c.2668G>C	p.Gly890Arg	2	1	1	2	2	protein
BRCA1	c.2692A>G	p.Lys898Glu	2	1	1	2	2	protein
BRCA1	c.2728C>G	p.Gln910Glu	2	1	1	2	2	protein
BRCA1	c.2758G>A	p.Val920Ile	1	1	1	2	2	protein

Table 7.1 continuation

Gene	DNA	Protein	ENIGMA	MLR	MLR+AS	NN	NN+AS	Comments
BRCA1	c.2765C>G	p.Thr922Arg	2	1	1	3	3	protein
BRCA1	c.2783G>A	p.Gly928Asp	2	2	2	3	3	protein
BRCA1	c.2798G>C	p.Gly933Ala	1	1	1	2	2	protein
BRCA1	c.2857T>C	p.Cys953Arg	2	3	3	3	3	protein
BRCA1	c.2884G>A	p.Glu962Lys	1	1	1	3	3	protein
BRCA1	c.2912A>G	p.His971Arg	1	2	2	2	2	protein
BRCA1	c.2917C>G	p.Leu973Val	2	1	1	3	3	protein
BRCA1	c.2935C>T	p.Arg979Cys	2	1	1	2	2	protein
BRCA1	c.2963C>T	p.Ser988Leu	2	2	2	3	3	protein
BRCA1	c.2998_ 3003del	p.Glu1000_ Glu1001del	2	3	3	3	3	arbitary
BRCA1	c.2998G>A	p.Glu1000Lys	2	1	1	3	3	protein
BRCA1	c.3040A>T	p.Met1014Leu	2	2	2	2	2	protein
BRCA1	c.305C>G	p.Ala102Gly	1	2	2	3	3	protein
BRCA1	c.3082C>T	p.Arg1028Cys	1	2	2	2	2	protein
BRCA1	c.3143G>A	p.Gly1048Asp	2	1	1	3	3	protein
BRCA1	c.3143G>T	p.Gly1048Val	1	2	2	3	3	protein
BRCA1	c.3211G>A	p.Glu1071Lys	2	1	1	3	3	protein
BRCA1	c.3220A>G	p.Arg1074Gly	2	3	3	3	3	protein
BRCA1	c.3267G>T	p.Leu1089Phe	2	1	1	3	3	protein
BRCA1	c.3280T>G	p.Tyr1094Asp	2	3	3	3	3	protein
BRCA1	c.3305A>G	p.Asn1102Ser	2	2	2	3	3	protein
BRCA1	c.3416G>T	p.Ser1139Ile	1	1	1	3	3	protein
BRCA1	c.3424G>C	p.Ala1142Pro	2	1	1	3	3	protein
BRCA1	c.3425C>T	p.Ala1142Val	2	1	1	3	3	protein
BRCA1	c.3454G>A	p.Asp1152Asn	2	1	1	3	3	protein
BRCA1	c.3541G>A	p.Val1181Ile	1	1	1	3	3	protein
BRCA1	c.3581C>T	p.Thr1194Ile	2	1	1	2	2	protein
BRCA1	c.3596C>T	p.Ala1199Val	2	2	2	3	3	protein
BRCA1	c.3622A>G	p.Lys1208Glu	2	2	2	3	3	protein
BRCA1	c.3655G>A	p.Glu1219Lys	2	2	2	3	3	protein
BRCA1	c.3657G>C	p.Glu1219Asp	1	1	1	3	3	protein
BRCA1	c.3667C>T	p.Leu1223Phe	2	2	2	3	3	protein

Table 7.1 continuation

Gene	DNA	Protein	ENIGMA	MLR	MLR+AS	NN	NN+AS	Comments
BRCA1	c.3708T>G	p.Asn1236Lys	1	2	2	2	2	protein
BRCA1	c.3724A>G	p.Thr1242Ala	2	3	3	3	3	protein
BRCA1	c.3848A>G	p.His1283Arg	2	3	3	3	3	protein
BRCA1	c.3902G>A	p.Ser1301Asn	2	1	1	3	3	protein
BRCA1	c.397C>T	p.Arg133Cys	2	3	3	3	3	protein
BRCA1	c.3988A>T	p.Ser1330Cys	2	1	1	2	2	protein
BRCA1	c.398G>A	p.Arg133His	2	3	3	3	3	protein
BRCA1	c.4006A>T	p.Ser1336Cys	2	1	1	3	3	protein
BRCA1	c.4031A>G	p.Asp1344Gly	2	2	2	3	3	protein
BRCA1	c.4036G>A	p.Glu1346Lys	1	1	1	3	3	protein
BRCA1	c.4046C>G	p.Thr1349Arg	2	2	2	2	2	protein
BRCA1	c.4081A>T	p.Met1361Leu	1	2	2	2	2	protein
BRCA1	c.4103C>T	p.Ala1368Val	2	2	2	3	3	protein
BRCA1	c.4184A>G	p.Gln1395Arg	2	3	5	3	5	splicing
BRCA1	c.4213A>G	p.Ile1405Val	2	1	1	2	2	protein
BRCA1	c.4262A>G	p.His1421Arg	2	3	3	3	3	protein
BRCA1	c.4288C>T	p.Pro1430Ser	2	2	2	2	2	protein
BRCA1	c.4342A>G	p.Ser1448Gly	2	1	1	2	2	protein
BRCA1	c.43A>C	p.Ile15Leu	2	2	2	2	2	protein
BRCA1	c.4484G>C	p.Arg1495Thr	5	2	5	3	5	splicing
BRCA1	c.4520G>C	p.Arg1507Thr	1	2	2	3	3	protein
BRCA1	c.455T>C	p.Leu152Pro	2	2	2	3	3	protein
BRCA1	c.4585A>G	p.Ile1529Val	2	1	1	2	2	protein
BRCA1	c.4657T>A	p.Leu1553Met	2	1	1	3	3	protein
BRCA1	c.4675G>A	p.Glu1559Lys	5	1	5	3	5	splicing
BRCA1	c.469T>C	p.Asn1236Lys	2	1	1	3	3	protein
BRCA1	c.4726G>C	p.Thr1242Ala	2	1	1	2	2	protein
BRCA1	c.4733A>G	p.His1283Arg	2	2	2	3	3	protein
BRCA1	c.4766G>A	p.Ser1301Asn	2	1	1	2	2	protein
BRCA1	c.4776C>A	p.Arg133Cys	2	1	1	2	2	protein
BRCA1	c.478G>C	p.Ser1330Cys	2	3	3	3	3	protein
BRCA1	c.4814T>C	p.Arg133His	2	1	1	2	2	protein

Table 7.1 continuation

Gene	DNA	Protein	ENIGMA	MLR	MLR+AS	NN	NN+AS	Comments
BRCA1	c.4816A>G	p.Lys1606Glu	1	1	1	2	2	protein
BRCA1	c.5068A>C	p.Lys1690Gln	2	3	3	3	3	protein
BRCA1	c.5078_ 5080del	p.Ala1693del	3	3	3	3	3	arbitary
BRCA1	c.508C>T	p.Arg170Trp	1	2	2	3	3	protein
BRCA1	c.5144G>A	p.Ser1715Asn	5	3	3	3	3	protein
BRCA1	c.5189A>G	p.Asn1730Ser	2	2	2	2	2	protein
BRCA1	c.5189A>T	p.Asn1730Ile	2	2	2	3	3	protein
BRCA1	c.5198A>G	p.Asp1733Gly	1	3	3	3	3	protein
BRCA1	c.5207T>G	p.Val1736Gly	4	3	3	3	3	protein
BRCA1	c.5213G>A	p.Gly1738Glu	4	3	3	3	3	protein
BRCA1	c.5216A>T	p.Asp1739Val	4	3	3	3	3	protein
BRCA1	c.5243G>A	p.Gly1748Asp	4	3	3	3	3	protein
BRCA1	c.5312C>G	p.Pro1771Arg	2	3	3	3	3	protein
BRCA1	c.53T>C	p.Met18Thr	5	4	4	3	3	protein
BRCA1	c.5456A>G	p.Asn1819Ser	2	1	1	2	2	protein
BRCA1	c.5504G>A	p.Arg1835Gln	2	3	3	3	3	protein
BRCA1	c.5531T>G	p.Leu1844Arg	1	2	2	3	3	protein
BRCA1	c.5553C>G	p.Asp1851Glu	2	2	2	2	2	protein
BRCA1	c.716A>G	p.His239Arg	1	3	3	3	3	protein
BRCA1	c.722C>T	p.Pro241Leu	2	1	1	2	2	protein
BRCA1	c.792T>A	p.Ser264Arg	2	1	1	2	2	protein
BRCA1	c.823G>A	p.Gly275Ser	2	2	2	3	3	protein
BRCA1	c.824G>A	p.Gly275Asp	2	2	2	3	3	protein
BRCA1	c.827C>G	p.Thr276Arg	1	2	2	3	3	protein
BRCA1	c.891G>T	p.Met297Ile	1	3	3	3	3	protein
BRCA1	c.932C>T	p.Pro311Leu	2	2	2	3	3	protein
BRCA1	c.964G>A	p.Ala322Thr	2	1	1	3	3	protein
BRCA1	c.964G>C	p.Ala322Pro	2	2	2	3	3	protein
BRCA1	c.994C>T	p.Arg332Trp	2	1	1	3	3	protein
BRCA1	c.997A>G	p.Thr333Ala	1	1	1	3	3	protein
BRCA2	c.10070C>T	p.Thr3357Ile	2	1	1	2	2	protein
BRCA2	c.10120A>G	p.Thr3374Ala	2	1	1	2	2	protein

Table 7.1 continuation

Gene	DNA	Protein	ENIGMA	MLR	MLR+AS	NN	NN+AS	Comments
BRCA2	c.10121C>T	p.Thr3374Ile	2	1	1	2	2	protein
BRCA2	c.10204G>A	p.Glu3402Lys	2	1	1	2	2	protein
BRCA2	c.1040A>G	p.Gln347Arg	2	1	1	2	2	protein
BRCA2	c.1124C>T	p.Pro375Leu	2	1	1	3	3	protein
BRCA2	c.1127T>G	p.Phe376Cys	2	1	1	2	2	protein
BRCA2	c.1166C>A	p.Pro389Gln	1	1	1	2	2	protein
BRCA2	c.116C>T	p.Ala39Val	2	1	1	1	1	protein
BRCA2	c.1181A>C	p.Glu394Ala	1	1	1	2	2	protein
BRCA2	c.1225G>A	p.Glu409Lys	2	1	1	2	2	protein
BRCA2	c.1247T>G	p.Ile416Ser	2	1	1	3	3	protein
BRCA2	c.1447G>A	p.Ala483Thr	2	1	1	1	1	protein
BRCA2	c.1466C>G	p.Ser489Cys	1	1	1	2	2	protein
BRCA2	c.1514T>C	p.Ile505Thr	1	1	1	2	2	protein
BRCA2	c.1786G>C	p.Asp596His	1	1	1	3	3	protein
BRCA2	c.1792A>G	p.Thr598Ala	1	1	1	2	2	protein
BRCA2	c.1796C>T	p.Ser599Phe	1	1	1	3	3	protein
BRCA2	c.1798T>C	p.Tyr600His	2	1	1	2	2	protein
BRCA2	c.1810A>G	p.Lys604Glu	1	1	1	2	2	protein
BRCA2	c.1814T>C	p.Ile605Thr	2	1	1	2	2	protein
BRCA2	c.1865C>T	p.Ala622Val	1	1	1	2	2	protein
BRCA2	c.1875T>A	p.Phe625Leu	2	1	1	2	2	protein
BRCA2	c.1885C>T	p.Leu629Phe	2	1	1	1	1	protein
BRCA2	c.1897A>G	p.Asn633Asp	2	1	1	2	2	protein
BRCA2	c.1938C>A	p.Ser646Arg	2	1	1	1	1	protein
BRCA2	c.2125C>G	p.Leu709Val	2	1	1	1	1	protein
BRCA2	c.2135T>C	p.Leu712Pro	2	1	1	1	1	protein
BRCA2	c.2213G>T	p.Cys738Phe	2	1	1	1	1	protein
BRCA2	c.2303C>T	p.Thr768Ile	2	1	1	1	1	protein
BRCA2	c.2330A>G	p.Asp777Gly	2	1	1	1	1	protein
BRCA2	c.2348T>G	p.Val783Gly	2	1	1	1	1	protein
BRCA2	c.241T>G	p.Phe81Val	2	3	3	3	3	protein
BRCA2	c.2429C>T	p.Thr810Ile	2	1	1	1	1	protein

Table 7.1 continuation

Gene	DNA	Protein	ENIGMA	MLR	MLR+AS	NN	NN+AS	Comments
BRCA2	c.2515T>C	p.Tyr839His	2	1	1	1	1	protein
BRCA2	c.2589T>A	p.Asn863Lys	2	1	1	1	1	protein
BRCA2	c.2632G>C	p.Asp878His	2	1	1	1	1	protein
BRCA2	c.2698A>G	p.Asn900Asp	1	1	1	1	1	protein
BRCA2	c.2803G>C	p.Asp935His	1	1	1	1	1	protein
BRCA2	c.2872A>G	p.Ser958Gly	2	1	1	1	1	protein
BRCA2	c.2920G>A	p.Asp974Asn	2	1	1	1	1	protein
BRCA2	c.2963A>C	p.Asp988Ala	2	1	1	1	1	protein
BRCA2	c.2987T>G	p.Leu996Arg	1	1	1	1	1	protein
BRCA2	c.3071_ 3073del	p.Ile1024del	2	3	3	3	3	arbitary
BRCA2	c.3088T>G	p.Phe1030Val	2	1	1	3	3	protein
BRCA2	c.3172A>C	p.Lys1058Gln	2	1	1	3	3	protein
BRCA2	c.3197A>G	p.Asn1066Ser	2	1	1	2	2	protein
BRCA2	c.322A>C	p.Asn108His	2	1	1	2	2	protein
BRCA2	c.3260C>T	p.Thr1087Ile	2	1	1	2	2	protein
BRCA2	c.3326C>T	p.Ala1109Val	2	1	1	4	4	protein
BRCA2	c.343A>G	p.Lys115Glu	2	1	1	2	2	protein
BRCA2	c.3445A>G	p.Met1149Val	2	1	1	1	1	protein
BRCA2	c.3503T>C	p.Met1168Thr	2	1	1	1	1	protein
BRCA2	c.3569G>A	p.Arg1190Gln	2	1	1	3	3	protein
BRCA2	c.3575T>G	p.Phe1192Cys	2	1	1	2	2	protein
BRCA2	c.3598T>A	p.Cys1200Ser	2	1	1	2	2	protein
BRCA2	c.3622T>A	p.Leu1208Ile	2	1	1	2	2	protein
BRCA2	c.3731T>C	p.Ile1244Thr	2	1	1	3	3	protein
BRCA2	c.3749A>G	p.Glu1250Gly	2	1	1	3	3	protein
BRCA2	c.3865A>G	p.Lys1289Glu	2	1	1	3	3	protein
BRCA2	c.3962A>G	p.Asp1321Gly	2	1	1	2	2	protein
BRCA2	c.3966C>G	p.Asn1322Lys	2	1	1	3	3	protein
BRCA2	c.4141_ 4143del	p.Lys1381del	2	3	3	3	3	arbitary
BRCA2	c.4146_ 4148del	p.Glu1382del	1	3	3	3	3	arbitary
BRCA2	c.4159T>A	p.Leu1387Ile	2	1	1	1	1	protein

Table 7.1 continuation

Gene	DNA	Protein	ENIGMA	MLR	MLR+AS	NN	NN+AS	Comments
BRCA2	c.4271C>G	p.Ser1424Cys	1	1	1	1	1	protein
BRCA2	c.4334A>T	p.Lys1445Ile	2	1	1	2	2	protein
BRCA2	c.4376A>G	p.Asn1459Ser	2	1	1	1	1	protein
BRCA2	c.437T>C	p.Leu146Pro	2	1	1	3	3	protein
BRCA2	c.440A>G	p.Gln147Arg	1	1	1	3	3	protein
BRCA2	c.4483G>A	p.Val1495Ile	2	1	1	1	1	protein
BRCA2	c.4558A>G	p.Thr1520Ala	2	1	1	1	1	protein
BRCA2	c.4718G>A	p.Cys1573Tyr	2	1	1	1	1	protein
BRCA2	c.4779A>C	p.Glu1593Asp	2	1	1	1	1	protein
BRCA2	c.4828G>A	p.Val1610Met	2	1	1	1	1	protein
BRCA2	c.4849A>C	p.Ser1617Arg	2	1	1	1	1	protein
BRCA2	c.4856A>G	p.Asn1619Ser	2	1	1	1	1	protein
BRCA2	c.4861T>G	p.Cys1621Gly	2	1	1	1	1	protein
BRCA2	c.4874A>G	p.Glu1625Gly	2	1	1	1	1	protein
BRCA2	c.4901T>C	p.Phe1634Ser	2	1	1	1	1	protein
BRCA2	c.4915G>A	p.Val1639Ile	2	1	1	1	1	protein
BRCA2	c.4987G>C	p.Val1663Leu	2	1	1	1	1	protein
BRCA2	c.5020A>G	p.Ser1674Gly	2	1	1	1	1	protein
BRCA2	c.506A>G	p.Lys169Arg	2	1	1	3	3	protein
BRCA2	c.5117A>C	p.Asn1706Thr	2	1	1	1	1	protein
BRCA2	c.5171T>C	p.Ile1724Thr	2	1	1	1	1	protein
BRCA2	c.5186A>G	p.Lys1729Arg	2	1	1	1	1	protein
BRCA2	c.5383A>G	p.Lys1795Glu	2	1	1	2	2	protein
BRCA2	c.5474C>T	p.Ala1825Val	2	1	1	2	2	protein
BRCA2	c.5507A>C	p.Asn1836Thr	2	1	1	2	2	protein
BRCA2	c.5552T>G	p.Ile1851Ser	2	1	1	2	2	protein
BRCA2	c.5554G>A	p.Val1852Ile	2	1	1	2	2	protein
BRCA2	c.5602G>T	p.Asp1868Tyr	2	1	1	3	3	protein
BRCA2	c.5634C>G	p.Asn1878Lys	1	1	1	3	3	protein
BRCA2	c.5635G>A	p.Glu1879Lys	2	1	1	3	3	protein
BRCA2	c.5640T>G	p.Asn1880Lys	2	1	1	2	2	protein
BRCA2	c.5649A>C	p.Lys1883Asn	2	1	1	2	2	protein

Table 7.1 continuation

Gene	DNA	Protein	ENIGMA	MLR	MLR+AS	NN	NN+AS	Comments
BRCA2	c.5651T>C	p.Ile1884Thr	2	1	1	2	2	protein
BRCA2	c.5702A>T	p.Glu1901Val	3	1	1	3	3	protein
BRCA2	c.5723T>C	p.Leu1908Pro	2	1	1	2	2	protein
BRCA2	c.5753A>G	p.His1918Arg	1	1	1	2	2	protein
BRCA2	c.5768A>C	p.Asp1923Ala	2	1	1	2	2	protein
BRCA2	c.5821A>C	p.Lys1941Gln	2	1	1	2	2	protein
BRCA2	c.5969A>G	p.Asp1990Gly	2	1	1	3	3	protein
BRCA2	c.6131G>C	p.Gly2044Ala	2	1	1	1	1	protein
BRCA2	c.6131G>T	p.Gly2044Val	2	1	1	2	2	protein
BRCA2	c.6188G>A	p.Gly2063Glu	2	3	3	4	4	protein
BRCA2	c.6196G>A	p.Val2066Ile	2	2	2	3	3	protein
BRCA2	c.6258C>G	p.Ile2086Met	2	1	1	2	2	protein
BRCA2	c.6441C>G	p.His2147Gln	2	1	1	1	1	protein
BRCA2	c.6443C>A	p.Ser2148Tyr	2	1	1	2	2	protein
BRCA2	c.6455C>A	p.Ser2152Tyr	1	1	1	3	3	protein
BRCA2	c.6532C>T	p.His2178Tyr	2	1	1	2	2	protein
BRCA2	c.6683T>C	p.Val2228Ala	2	1	1	3	3	protein
BRCA2	c.6698C>A	p.Ala2233Asp	2	3	3	3	3	protein
BRCA2	c.6737C>G	p.Pro2246Arg	2	1	1	3	3	protein
BRCA2	c.6746C>A	p.Ala2249Asp	2	1	1	2	2	protein
BRCA2	c.679G>A	p.Ala227Thr	2	1	1	2	2	protein
BRCA2	c.6871A>G	p.Asn2291Asp	2	1	1	2	2	protein
BRCA2	c.6953G>A	p.Arg2318Gln	1	2	2	3	3	protein
BRCA2	c.6991A>G	p.Thr2331Ala	2	1	1	3	3	protein
BRCA2	c.6995G>A	p.Cys2332Tyr	2	1	1	3	3	protein
BRCA2	c.7025A>C	p.Gln2342Pro	2	1	1	3	3	protein
BRCA2	c.7118G>C	p.Ser2373Thr	2	1	1	1	1	protein
BRCA2	c.7457A>G	p.Asn2486Ser	2	1	1	1	1	protein
BRCA2	c.7499G>C	p.Arg2500Thr	2	1	1	1	1	protein
BRCA2	c.7505G>A	p.Arg2502His	1	1	1	1	1	protein
BRCA2	c.7512T>G	p.Phe2504Leu	2	1	1	1	1	protein
BRCA2	c.7534C>T	p.Leu2512Phe	1	1	1	1	1	protein

Table 7.1 continuation

Gene	DNA	Protein	ENIGMA	MLR	MLR+AS	NN	NN+AS	Comments
BRCA2	c.7601C>T	p.Ala2534Val	2	1	1	1	1	protein
BRCA2	c.7633G>A	p.Val2545Ile	2	1	1	1	1	protein
BRCA2	c.7783G>T	p.Ala2595Ser	2	1	1	3	3	protein
BRCA2	c.7819A>C	p.Thr2607Pro	4	3	3	3	3	protein
BRCA2	c.7975A>G	p.Arg2659Gly	5	3	3	4	4	protein
BRCA2	c.7994A>G	p.Asp2665Gly	1	3	3	3	3	protein
BRCA2	c.8009C>T	p.Ser2670Leu	4	2	2	4	4	protein
BRCA2	c.800G>A	p.Gly267Glu	3	1	1	2	2	protein
BRCA2	c.8182G>A	p.Val2728Ile	1	1	1	2	2	protein
BRCA2	c.8254A>T	p.Ile2752Phe	2	1	1	2	2	protein
BRCA2	c.8308G>A	p.Ala2770Thr	1	1	1	3	3	protein
BRCA2	c.831T>G	p.Asn277Lys	1	1	1	3	3	protein
BRCA2	c.8324T>C	p.Met2775Thr	2	1	1	2	2	protein
BRCA2	c.8386C>T	p.Pro2796Ser	2	1	1	2	2	protein
BRCA2	c.841G>A	p.Asp281Asn	2	1	1	3	3	protein
BRCA2	c.8428A>G	p.Ser2810Gly	2	1	1	3	3	protein
BRCA2	c.8432A>G	p.Asp2811Gly	2	1	1	3	3	protein
BRCA2	c.8486A>G	p.Gln2829Arg	5	1	5	3	5	splicing
BRCA2	c.8503T>C	p.Ser2835Pro	2	1	1	2	2	protein
BRCA2	c.8572C>A	p.Gln2858Lys	2	1	1	3	3	protein
BRCA2	c.8599A>C	p.Thr2867Pro	2	1	1	2	2	protein
BRCA2	c.8651A>G	p.Tyr2884Cys	2	1	1	2	2	protein
BRCA2	c.8734G>A	p.Ala2912Thr	1	1	1	2	2	protein
BRCA2	c.8764A>G	p.Ser2922Gly	1	2	2	3	3	protein
BRCA2	c.8789A>C	p.Asn2930Thr	2	1	1	2	2	protein
BRCA2	c.8918G>A	p.Arg2973His	2	1	1	3	3	protein
BRCA2	c.8975_ 9100del	p.Pro2992_ Thr3033del	4	3	3	3	3	arbitary
BRCA2	c.9011A>G	p.Lys3004Arg	2	1	1	2	2	protein
BRCA2	c.9038C>T	p.Thr3013Ile	1	1	1	2	2	protein
BRCA2	c.9043A>G	p.Lys3015Glu	1	1	1	2	2	protein
BRCA2	c.9104A>C	p.Tyr3035Ser	1	1	1	3	3	protein
BRCA2	c.9175A>G	p.Lys3059Glu	1	1	1	2	2	protein

Table 7.1 continuation

Gene	DNA	Protein	ENIGMA	MLR	MLR+AS	NN	NN+AS	Comments
BRCA2	c.9199C>T	p.Pro3067Ser	2	1	1	3	3	protein
BRCA2	c.9242T>C	p.Val3081Ala	2	1	1	3	3	protein
BRCA2	c.9263C>T	p.Ala3088Val	2	1	1	3	3	protein
BRCA2	c.9286G>A	p.Glu3096Lys	2	1	1	3	3	protein
BRCA2	c.9350A>C	p.His3117Pro	2	1	1	2	2	protein
BRCA2	c.9371A>T	p.Asn3124Ile	5	3	3	3	3	protein
BRCA2	c.9434T>C	p.Val3145Ala	2	1	1	2	2	protein
BRCA2	c.9458G>C	p.Gly3153Ala	2	1	1	1	1	protein
BRCA2	c.9500A>C	p.Glu3167Ala	2	1	5	3	5	splicing
BRCA2	c.9513_ 9515del	p.Leu3172del	2	3	3	3	3	arbitary
BRCA2	c.955A>G	p.Asn319Asp	2	1	1	2	2	protein
BRCA2	c.956A>G	p.Asn319Ser	2	1	1	2	2	protein
BRCA2	c.9581C>A	p.Pro3194Gln	2	1	1	2	2	protein
BRCA2	c.964A>C	p.Lys322Gln	2	1	1	3	3	protein
BRCA2	c.9875C>T	p.Pro3292Leu	1	3	3	3	3	protein
BRCA2	c.9905G>A	p.Arg3302Lys	2	1	1	3	3	protein
BRCA2	c.9925G>A	p.Glu3309Lys	2	1	1	2	2	protein

8. Bibliography

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., … Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, *7*(4), 248–249. https://doi.org/10.1038/nmeth0410-248

Adzhubei, I., Schmidt, S., Peshkin, L., Ramensky, V., Gerasimova, A., Bork, P., … Sunyaev, S. (2010). PolyPhen-2 : prediction of functional effects of human nsSNPs. *Nat. Methods*, *7*(4), 248–249. https://doi.org/10.1017/CBO9781107415324.004

Altshuler, D. L., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., … Peterson, J. L. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–1073. https://doi.org/10.1038/nature09534

Ancien, F., Pucci, F., Godfroid, M., & Rooman, M. (2018). Prediction and interpretation of deleterious coding variants in terms of protein structural stability. *Scientific Reports*, *8*(1), 1–11. https://doi.org/10.1038/s41598-018-22531-2

Arita, K., Isogai, S., Oda, T., Unoki, M., Sugita, K., Sekiyama, N., … Shirakawa, M. (2012). Recognition of modification status on a histone H3 tail by linked histone reader modules of the epigenetic regulator UHRF1. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(32), 12950–12955. https://doi.org/10.1073/pnas.1203701109

Baldi, P., & Brunak, S. (2001). *Bioinformatics. 2nd ed.* Cambridge, Massachusetts: The MIT Pressle.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, *16*(5), 412–424.

https://doi.org/10.1093/bioinformatics/16.5.412

Baldi, Pierre, Brunak, S., Chauvin, Y., Andersen, C. A. F., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics*, *6*(5), 412–424. https://doi.org/10.1093/bioinformatics/16.5.412

Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., … Zhang, J. (2017a). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, *45*(D1), D158–D169. https://doi.org/10.1093/nar/gkw1099

Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., … Zhang, J. (2017b). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, *45*(Database issue), D158–D169. https://doi.org/10.1093/nar/gkw1099

Biswas, K., Das, R., Eggington, J. M., Qiao, H., North, S. L., Stauffer, S., … Sharan, S. K. (2012). Functional evaluation of BRCA2 variants mapping to the PALB2-binding and c-terminal dna-binding domains using a mouse es cell-based assay. *Human Molecular Genetics*, *21*(18), 3993–4006. https://doi.org/10.1093/hmg/dds222

Bjerke, L., Mackay, A., Nandhabalan, M., Burford, A., Jury, A., Popov, S., … Jones, C. (2013). Histone H3.3 mutations drive pediatric glioblastoma through upregulation of MYCN. *Cancer Discovery*, *3*(5), 512–519. https://doi.org/10.1158/2159-8290.CD-12-0426

Bondi, A. (1964). van der Waals Volumes and Radii - The Journal of Physical Chemistry (ACS Publications). *The Journal of Physical Chemistry*, *68*, 441–451. https://doi.org/10.1021/j100785a001

Boriack-Sjodin, P. A., Jin, L., Jacques, S. L., Drew, A., Sneeringer, C., Scott, M. P., … Copeland, R. A. (2016). Structural Insights into Ternary Complex

Formation of Human CARM1 with Various Substrates. *ACS Chemical Biology*, *11*(3), 763–771. https://doi.org/10.1021/acschembio.5b00773

Bouwman, P., van der Gulden, H., van der Heijden, I., Drost, R., Klijn, C. N., Prasetyanti, P., … Jonkers, J. (2013). A high-throughput functional complementation assay for classification of BRCA1 missense variants. *Cancer Discovery*, *3*(10), 1142–1155. https://doi.org/10.1158/2159-8290.CD-13-0094

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *68*(6), 394–424. https://doi.org/10.3322/caac.21492

Bromberg, Y., & Rost, B. (2007). SNAP: Predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Research*, *35*(11), 3823–3835. https://doi.org/10.1093/nar/gkm238

Brzovic, P. S., Rajagopal, P., Hoyt, D. W., King, M., & Klevit, R. E. (2001). *Structure of a BRCA1 – BARD1 heterodimeric RING – RING complex*. 833–837.

Burgess, R. J., & Zhang, Z. (2013). Histone chaperones in nucleosome assembly and human disease. *Nature Structural and Molecular Biology*, *20*(1), 14–22. https://doi.org/10.1038/nsmb.2461

Carter, H., Douville, C., Stenson, P. D., Cooper, D. N., & Karchin, R. (2013). Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics*, *14 Suppl 3*(Suppl 3). https://doi.org/10.1186/1471-2164-14-s3-s3

Castéra, L., Krieger, S., Rousselin, A., Legros, A., Baumann, J. J., Bruet, O., … Vaur, D. (2014). Next-generation sequencing for the diagnosis of

hereditary breast and ovarian cancer using genomic capture targeting multiple candidate genes. *European Journal of Human Genetics*, *22*(11), 1305–1313. https://doi.org/10.1038/ejhg.2014.16

Castilla, L. H., Couch, F. J., Erdos, M. R., Hoskins, K. F., Calzone, K., Garber, J. E., … Weber, B. L. (1994). Mutations in the BRCA1 gene in families with early-onset breast and ovarian cancer. *Nature*, *8*, 387–391.

Chang, F. T. M., Chan, F. L., McGhie, J. D. R., Udugama, M., Mayne, L., Collas, P., … Wong, L. H. (2015). CHK1-driven histone H3.3 serine 31 phosphorylation is important for chromatin maintenance and cell survival in human ALT cancer cells. *Nucleic Acids Research*, *43*(5), 2603–2614. https://doi.org/10.1093/nar/gkv104

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chen, S., Iversen, E. S., Friebel, T., Finkelstein, D., Weber, B. L., Eisen, A., … Parmigiani, G. (2006). Characterization of BRCA1 and BRCA2 mutations in a large United States sample. *Journal of Clinical Oncology*, *24*(6), 863–871. https://doi.org/10.1200/JCO.2005.03.6772

Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData Mining*, *10*(1), 1–17. https://doi.org/10.1186/s13040-017-0155-3

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., & Chan, A. P. (2012). Predicting the functional effect of amino Acid substitutions and indels. *PloS One*, *7*(10), e46688. https://doi.org/10.1371/journal.pone.0046688

Coe, B. P., Witherspoon, K., Rosenfeld, J. A., Van Bon, B. W. M., Vulto-Van Silfhout, A. T., Bosco, P., … Eichler, E. E. (2014). Refining analyses of copy number variation identifies specific genes associated with

developmental delay. *Nature Genetics*, *46*(10), 1063–1071. https://doi.org/10.1038/ng.3092

Colobran, R., Álvarez de la Campa, E., Soler-Palacín, P., Martín-Nalda, A., Pujol-Borrell, R., de la Cruz, X., & Martínez-Gallo, M. (2016). Clinical and structural impact of mutations affecting the residue Phe367 of FOXP3 in patients with IPEX syndrome. *Clinical Immunology*, *163*, 60–65. https://doi.org/10.1016/j.clim.2015.12.014

Couch, F. J., Nathanson, K. L., & Offit, K. (2014). Two Decades After BRCA: Setting Paradigms in Personalized Cancer Care and Prevention. *Science*, *343*(6178), 1466–1470. https://doi.org/10.1038/jid.2014.371

Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory*. USA: Wiley-Interscience.

Cox, S. G., Kim, H., Garnett, A. T., Medeiros, D. M., An, W., & Crump, J. G. (2012). An Essential Role of Variant Histone H3.3 for Ectomesenchyme Potential of the Cranial Neural Crest. *PLoS Genetics*, *8*(9). https://doi.org/10.1371/journal.pgen.1002938

Crockett, D. K., Lyon, E., Williams, M. S., Narus, S. P., Facelli, J. C., & Mitchell, J. A. (2012). Utility of gene-specific algorithms for predicting pathogenicity of uncertain gene variants. *Journal of the American Medical Informatics Association*, *19*, 207–211. https://doi.org/10.1136/amiajnl-2011-000309

Crosio, C., Fimia, G. M., Loury, R., Kimura, M., Okano, Y., Zhou, H., … Sassone-Corsi, P. (2002). Mitotic Phosphorylation of Histone H3: Spatio-Temporal Regulation by Mammalian Aurora Kinases. *Molecular and Cellular Biology*, *22*(3), 874–885. https://doi.org/10.1128/MCB.22.3.874

de la Campa, E. Á., Padilla, N., & de la Cruz, X. (2017). Development of

pathogenicity predictors specific for variants that do not comply with clinical guidelines for the use of computational evidence. *BMC Genomics*, *18*(S5), 569. https://doi.org/10.1186/s12864-017-3914-0

Díez, O., Osorio, A., Durán, M., Martinez-Ferrandis, J. I., De la Hoya, M., Salazar, R., … Baiget, M. (2003). Analysis of BRCA1 and BRCA2 genes in Spanish breast/ovarian cancer patients: A high proportion of mutations unique to Spain and evidence of founder effects. *Human Mutation*, *22*(4), 301–312. https://doi.org/10.1002/humu.10260

Dines, J. N., Shirts, B. H., Slavin, T. P., Walsh, T., King, M. C., Fowler, D. M., & Pritchard, C. C. (2020). Systematic misclassification of missense variants in BRCA1 and BRCA2 "coldspots." *Genetics in Medicine*, *0*(0), 1–6. https://doi.org/10.1038/s41436-019-0740-6

Easton, D. F., Ford, D., Bishop, D. T., Haites, N., Milner, B., Allan, L., … Egilsson, V. (1995). Breast and ovarian cancer incidence in BRCA1-mutation carriers. *American Journal of Human Genetics*, *56*(1), 265–271.

Eccles, E. B., Mitchell, G., Monteiro, A. N. A., Schmutzler, R., Couch, F. J., Spurdle, A. B., … Goldgar, D. (2015). BRCA1 and BRCA2 genetic testing-pitfalls and recommendations for managing variants of uncertain clinical significance. *Ann. Oncol.*, *26*, 2057–2065. https://doi.org/10.1093/annonc/mdv278

Ederveen, T. H. A., Mandemaker, I. K., & Logie, C. (2011). The human histone H3 complement anno 2011. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, *1809*(10), 577–586. https://doi.org/10.1016/j.bbagrm.2011.07.002

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Elsässer, S. J., Huang, H., Lewis, P. W., Chin, J. W., Allis, C. D., & Patel, D. J. (2012). DAXX envelops a histone H3.3-H4 dimer for H3.3-specific recognition. *Nature*, *491*(7425), 560–565. https://doi.org/10.1038/nature11608

Ernst, C., Hahnen, E., Engel, C., Nothnagel, M., Weber, J., Schmutzler, R. K., & Hauke, J. (2018). Performance of in silico prediction tools for the classification of rare BRCA1/2 missense variants in clinical diagnostics. *BMC Medical Genomics*, *11*(1), 1–10. https://doi.org/10.1186/s12920-018-0353-y

Fackenthal, J. D., & Olopade, O. I. (2007). Breast cancer risk associated with BRCA1 and BRCA2 in diverse populations. *Nature Reviews Cancer*, *7*(12), 937–948. https://doi.org/10.1038/nrc2054

Fang, R., Chen, F., Dong, Z., Hu, D., Barbera, A. J., Clark, E. A., … Shi, Y. G. (2013). LSD2/KDM1B and Its Cofactor NPAC/GLYR1 Endow a Structural and Molecular Model for Regulation of H3K4 Demethylation. *Molecular Cell*, *49*(3), 558–570. https://doi.org/10.1016/j.molcel.2012.11.019

Fauchere, J., & Pliska, V. (1983). Hydrophobic parameters of amino acid side-chains from the partitioning of N-acetyl-amino-acid amides. *Eur. J. Med. Chem.-Chim. Ther.*, *18*, 369–375.

Faundes, V., Newman, W. G., Bernardini, L., Canham, N., Clayton-Smith, J., Dallapiccola, B., … Banka, S. (2018). Histone Lysine Methylases and Demethylases in the Landscape of Human Developmental Disorders. *American Journal of Human Genetics*, *102*(1), 175–187. https://doi.org/10.1016/j.ajhg.2017.11.013

Ferlay, J., Colombet, M., Soerjomataram, I., Dyba, T., Randi, G., Bettio, M., … Bray, F. (2018). Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018. *European*

Journal of Cancer, *103*, 356–387. https://doi.org/10.1016/j.ejca.2018.07.005

Fernández-Recio, J. (2011). Prediction of protein binding sites and hot spots. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *1*(5), 680–698. https://doi.org/10.1002/wcms.45

Ferrer-Costa, C., Orozco, M., & De La Cruz, X. (2004). Sequence-based prediction of pathological mutations. *Proteins: Structure, Function and Genetics*, *57*(4), 811–819. https://doi.org/10.1002/prot.20252

Ferrer-Costa, Carles, Orozco, M., & de la Cruz, X. (2002). Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *Journal of Molecular Biology*, *315*(4), 771–786. https://doi.org/10.1006/jmbi.2001.5255\nS0022283601952556 [pii]

Findlay, G. M., Daza, R. M., Martin, B., Zhang, M. D., Leith, A. P., Gasperini, M., … Shendure, J. (2018). Accurate classification of BRCA1 variants with saturation genome editing. *Nature*, *562*, 217–222. https://doi.org/10.1038/s41586-018-0461-z

Ford, D., Easton, D. F., Stratton, M., Narod, S., Goldgar, D., Devilee, P., … Zelada-Hedman, M. (1998). Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. *American Journal of Human Genetics*, *62*(3), 676–689. https://doi.org/10.1086/301749

Frank, D., Doenecke, D., & Albig, W. (2003). Differential expression of human replacement and cell cycle dependent H3 histone genes. *Gene*, *312*(1–2), 135–143. https://doi.org/10.1016/S0378-1119(03)00609-7

Ghosh, R., Oak, N., & Plon, S. E. (2017). Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. *Genome*

Biology, *18*(1), 1–12. https://doi.org/10.1186/s13059-017-1353-5

Goldgar, D. E., Easton, D. F., Byrnes, G. B., Spurdle, A. B., Iversen, E. S., & Greenblatt, M. S. (2008). Genetic evidence and integration of various data sources for classifying uncertain variants into a single model. *Human Mutation*, *29*(11), 1265–1272. https://doi.org/10.1002/humu.20897

Gorodkin, J. (2004). Comparing two K-category assignments by a K-category correlation coefficient. *Computational Biology and Chemistry*, *28*, 367–374. https://doi.org/10.1016/j.compbiolchem.2004.09.006

Grimm, D. G., Azencott, C. A., Aicheler, F., Gieraths, U., Macarthur, D. G., Samocha, K. E., … Borgwardt, K. M. (2015). The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Human Mutation*, *36*(5), 513–523. https://doi.org/10.1002/humu.22768

Guerois, R., Nielsen, J. E., & Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *Journal of Molecular Biology*, *320*(2), 369–387.

Guidugli, L., Carreira, A., Caputo, S. M., Ehlen, A., Galli, A., Monteiro, A. N. A., … Vreeswijk, M. P. G. (2014). Functional assays for analysis of variants of uncertain significance in BRCA2. *Human Mutation*, *35*(2), 151–164. https://doi.org/10.1002/humu.22478

Guidugli, L., Pankratz, V. S., Singh, N., Thompson, J., Erding, C. A., Engel, C., … Couch, F. J. (2013). A classification model for BRCA2 DNA binding domain missense variants based on homology-directed repair activity. *Cancer Research*, *73*(1), 265–275. https://doi.org/10.1158/0008-5472.CAN-12-2081

Guidugli, L., Shimelis, H., Masica, D. L., Pankratz, V. S., Lipton, G. B., Singh, N.,

… Couch, F. J. (2018). Assessment of the Clinical Relevance of BRCA2 Missense Variants by Functional and Computational Approaches. *American Journal of Human Genetics*, *102*(2), 233–248. https://doi.org/10.1016/j.ajhg.2017.12.013

Guo, R., Zheng, L., Park, J. W., Lv, R., Chen, H., Jiao, F., … Shi, Y. (2014). BS69/ZMYND11 reads and connects histone H3.3 lysine 36 trimethylation-decorated chromatin to regulated pre-mRNA processing. *Molecular Cell*, *56*(2), 298–310. https://doi.org/10.1016/j.molcel.2014.08.022

Guo, X., Wang, L., Li, J., Ding, Z., Xiao, J., Yin, X., … Xu, Y. (2015). Structural insight into autoinhibition and histone H3-induced activation of DNMT3A. *Nature*, *517*(7536), 640–644. https://doi.org/10.1038/nature13899

Hake, S. B., & Allis, C. D. (2006). Histone H3 variants and their potential role in indexing mammalian genomes: The "H3 barcode hypothesis." *Proceedings of the National Academy of Sciences of the United States of America*, *103*(17), 6428–6435. https://doi.org/10.1073/pnas.0600803103

Hake, S. B., Garcia, B. A., Kauer, M., Baker, S. P., Shabanowitz, J., Hunt, D. F., & Allis, C. D. (2005). Serine 31 phosphorylation of histone variant H3.3 is specific to regions bordering centromeres in metaphase chromosomes. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(18), 6344–6349. https://doi.org/10.1073/pnas.0502413102

Hall, J. M., Lee, M. K., Newman, B., Morrow, J. E., Anderson, L. A., Huey, B., & King, M. (1990). Linkage of Early-Onset Familial Breast Cancer to Chromosome 17q21. *Science*, *250*, 17–22.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, *11*, 10–18. https://doi.org/10.1088/1751-8113/44/8/085201

He, C., Li, F., Zhang, J., Wu, J., & Shi, Y. (2013). The methyltransferase NSD3 has chromatin-binding motifs, PHD5-C5HCH, that are distinct from other NSD (nuclear receptor SET domain) family members in their histone H3 recognition. *Journal of Biological Chemistry*, *288*(7), 4692–4703. https://doi.org/10.1074/jbc.M112.426148

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, *89*(22), 10915–10919. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=1438297

Hoskins, R. A., Repo, S., Barsky, D., Andreoletti, G., Moult, J., & Brenner, S. E. (2017). Reports from CAGI: The Critical Assessment of Genome Interpretation. *Human Mutation*, *38*(9), 1039–1041. https://doi.org/10.1002/humu.23290

Huang, H., Strømme, C. B., Saredi, G., Hödl, M., Strandsby, A., González-Aguilera, C., … Patel, D. J. (2015). A unique binding mode enables MCM2 to chaperone histones H3-H4 at replication forks. *Nature Structural and Molecular Biology*, *22*(8), 618–626. https://doi.org/10.1038/nsmb.3055

Hunt, S. E., McLaren, W., Gil, L., Thormann, A., Schuilenburg, H., Sheppard, D., … Cunningham, F. (2018). Ensembl variation resources. *Database : The Journal of Biological Databases and Curation*, *2018*(8), 1–12. https://doi.org/10.1093/database/bay119

Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K.,

Baheti, S., … Sieh, W. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *American Journal of Human Genetics*, *99*(4), 877–885. https://doi.org/10.1016/j.ajhg.2016.08.016

Itan, Y., Shang, L., Boisson, B., Ciancanelli, M. J., Markle, J. G., Martinez-Barricarte, R., … Casanova, J. L. (2016). The mutation significance cutoff: gene-level thresholds for variant predictions. *Nature Methods*, *13*(2), 109–110. https://doi.org/10.1038/nmeth.3739

Johnson, P., Mitchell, V., McClure, K., Kellems, M., Marshall, S., Allison, M. K., … Duina, A. A. (2015). A systematic mutational analysis of a histone H3 residue in budding yeast provides insights into chromatin dynamics. *G3: Genes, Genomes, Genetics*, *5*(5), 741–749. https://doi.org/10.1534/g3.115.017376

Judkins, T., Rosenthal, E., Arnell, C., Burbidge, L. A., Geary, W., Barrus, T., … Roa, B. B. (2012). Clinical significance of large rearrangements in BRCA1 and BRCA2. *Cancer*, *118*(21), 5210–5216. https://doi.org/10.1002/cncr.27556

Jurman, G., Riccadonna, S., & Furlanello, C. (2012). A comparison of MCC and CEN error measures in multi-class prediction. *PLoS ONE*, *7*(8), e41882. https://doi.org/10.1371/journal.pone.0041882

Karbassi, I., Maston, G. A., Love, A., Divincenzo, C., Braastad, C. D., Elzinga, C. D., … Higgins, J. J. (2016). A Standardized DNA Variant Scoring System for Pathogenicity Assessments in Mendelian Disorders. *Human Mutation*, *37*(1), 127–134. https://doi.org/10.1002/humu.22918

Karchin, R., Agarwal, M., Sali, A., Couch, F., & Beattie, M. S. (2008). Classifying Variants of Undetermined Significance in BRCA2 with Protein Likelihood Ratios. *Cancer Informatics*, *6*, 203–216.

https://doi.org/10.4137/CIN.S618

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., … MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7

Kast, K., Rhiem, K., Wappenschmidt, B., Hahnen, E., Hauke, J., Bluemcke, B., … Engel, C. (2016). Prevalence of BRCA1/2 germline mutations in 21 401 families with breast and ovarian cancer. *Journal of Medical Genetics*, *53*(7), 465–471. https://doi.org/10.1136/jmedgenet-2015-103672

King, M. C., Marks, J. H., & Mandell, J. B. (2003). Breast and Ovarian Cancer Risks Due to Inherited Mutations in BRCA1 and BRCA2. *Science*, *302*(5645), 643–646. https://doi.org/10.1126/science.1088759

Kircher, M., Witten, D. M., Jain, P., O'roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–315. https://doi.org/10.1038/ng.2892

Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, *4*(7), 1073–1081. https://doi.org/10.1038/nprot.2009.86

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., … Maglott, D. R. (2016). ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, *44*(D1), D862–D868. https://doi.org/10.1093/nar/gkv1222

Lau, P. N. I., & Cheung, P. (2011). Histone code pathway involving H3 S28 phosphorylation and K27 acetylation activates transcription and antagonizes polycomb silencing. *Proceedings of the National Academy*

of Sciences of the United States of America, *108*(7), 2801–2806. https://doi.org/10.1073/pnas.1012798108

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., … Williams, A. L. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291. https://doi.org/10.1038/nature19057

Lepack, A. E., Bagot, R. C., Peña, C. J., Loh, Y. H. E., Farrelly, L. A., Lu, Y., … Maze, I. (2016). Aberrant H3.3 dynamics in NAc promote vulnerability to depressive-like behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(44), 12562–12567. https://doi.org/10.1073/pnas.1608270113

Li, B., Krishnan, V. G., Mort, M. E., Xin, F., Kamati, K. K., Cooper, D. N., … Radivojac, P. (2009). MutPred: Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics*, *25*(21), 2744–2750. https://doi.org/10.1093/bioinformatics/btp528

Li, Y., Wen, H., Xi, Y., Tanaka, K., Wang, H., Peng, D., … Shi, X. (2014). AF9 YEATS domain links histone acetylation to DOT1L-mediated H3K79 methylation. *Cell*, *159*(3), 558–571. https://doi.org/10.1016/j.cell.2014.09.049

Lindor, N. M., Guidugli, L., Wang, X., Vallée, M. P., Monteiro, A. N. A., Tavtigian, S., … Couch, F. J. (2012). A review of a multifactorial probability-based model for classification of BRCA1 and BRCA2 variants of uncertain significance (VUS). *Human Mutation*, *33*(1), 8–21. https://doi.org/10.1177/104398629200800406

Liu, C. P., Xiong, C., Wang, M., Yu, Z., Yang, N., Chen, P., … Xu, R. M. (2012). Structure of the variant histone H3.3-H4 heterodimer in complex with its chaperone DAXX. *Nature Structural and Molecular Biology*, *19*(12),

1287–1292. https://doi.org/10.1038/nsmb.2439

Liu, X., Wu, C., Li, C., & Boerwinkle, E. (2016). dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Human Mutation*, *37*(3), 235–241. https://doi.org/10.1002/humu.22932

Liu, Y., Tempel, W., Zhang, Q., Liang, X., Loppnau, P., Qin, S., & Min, J. (2016). Family-wide characterization of histone binding abilities of human CW domain-containing proteins. *Journal of Biological Chemistry*, *291*(17), 9000–9013. https://doi.org/10.1074/jbc.M116.718973

López-Ferrando, V., Gazzo, A., De La Cruz, X., Orozco, M., & Gelpí, J. L. (2017). PMut: A web-based tool for the annotation of pathological variants on proteins, 2017 update. *Nucleic Acids Research*, *45*(W1), W222–W228. https://doi.org/10.1093/nar/gkx313

Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F., & Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, *389*(6648), 251–260. https://doi.org/10.1038/38444

MacArthur, D. G., Manolio, T. A., Dimmock, D. P., Rehm, H. L., Shendure, J., Abecasis, G. R., … Gunter, C. (2014). Guidelines for investigating causality of sequence variants in human disease. *Nature*, *508*, 469–476. https://doi.org/10.1038/nature13127

Masica, D.L., & Karchin, R. (2016). Towards Increasing the Clinical Relevance of In Silico Methods to Predict Pathogenic Missense Variants. *PLoS Comput. Biol.*, *12*(5), e1004725.

Masica, David L., & Karchin, R. (2016). Towards Increasing the Clinical Relevance of In Silico Methods to Predict Pathogenic Missense Variants. *PLoS Computational Biology*, *12*(5), 1–16.

https://doi.org/10.1371/journal.pcbi.1004725

Matsubara, K., Sano, N., Umehara, T., & Horikoshi, M. (2007). Global analysis of functional surfaces of core histones with comprehensive point mutants. *Genes to Cells*, *12*(1), 13–33. https://doi.org/10.1111/j.1365-2443.2007.01031.x

Mavaddat, N., Peock, S., Frost, D., Ellis, S., Platte, R., Fineberg, E., … Easton, D. F. (2013). Cancer risks for BRCA1 and BRCA2 mutation carriers: Results from prospective analysis of EMBRACE. *Journal of the National Cancer Institute*, *105*(11), 812–822. https://doi.org/10.1093/jnci/djt095

Maze, I., Wenderski, W., Noh, K. M., Bagot, R. C., Tzavaras, N., Purushothaman, I., … Allis, C. D. (2015). Critical Role of Histone Turnover in Neuronal Transcription and Plasticity. *Neuron*, *87*(1), 77–94. https://doi.org/10.1016/j.neuron.2015.06.014

Miki, Y., Swensen, J., Shattuck-eidens, D., Futreal, P. A., Tavtigian, S., Liu, Q., … Skolnick, M. H. (1994). Candidate Ovarian and Breast Susceptibility Gene for the Cancer. *Science*, 1–7.

Millot, G. A., Carvalho, M. A., Caputo, S. M., Vreeswijk, M. P. G., Brown, M. A., Webb, M., … Monteiro, A. N. A. (2012). A guide for functional analysis of BRCA1 variants of uncertain significance. *Human Mutation*, *33*(11), 1526–1537. https://doi.org/10.1002/humu.22150

Moghadasi, S., Eccles, D. M., Devilee, P., Vreeswijk, M. P. G., & van Asperen, C. J. (2016). Classification and Clinical Management of Variants of Uncertain Significance in High Penetrance Cancer Predisposition Genes. *Human Mutation*, *37*(4), 331–336. https://doi.org/10.1002/humu.22956

Moles-Fernández, A., Duran-Lozano, L., Montalban, G., Bonache, S., López-Perolio, I., Menéndez, M., … Gutiérrez-Enríquez, S. (2018).

Computational tools for splicing defect prediction in breast/ovarian cancer genes: How efficient are they at predicting RNA alterations? *Frontiers in Genetics*, *9*, 366. https://doi.org/10.3389/fgene.2018.00366

Moran, A., O'Hara, C., Khan, S., Shack, L., Woodward, E., Maher, E. R., … Evans, D. G. R. (2012). Risk of cancer other than breast or ovarian in individuals with BRCA1 and BRCA2 mutations. *Familial Cancer*, *11*(2), 235–242. https://doi.org/10.1007/s10689-011-9506-2

Moreno, L., Linossi, C., Esteban, I., Gadea, N., Carrasco, E., Bonache, S., … Balmaña, J. (2016). Germline BRCA testing is moving from cancer risk assessment to a predictive biomarker for targeting cancer therapeutics. *Clin. Trans. Oncol.*, *18*, 981–987. https://doi.org/10.1007/s12094-015-1470-0

Moskowitz, A. M., Belnap, N., Siniard, A. L., Szelinger, S., Claasen, A. M., Richholt, R. F., … Schrauwen, I. (2016). A de novo missense mutation in ZMYND11 is associated with global developmental delay, seizures, and hypotonia . *Molecular Case Studies*, *2*(5), a000851. https://doi.org/10.1101/mcs.a000851

Mullan, P. B., Quinn, J. E., & Harkin, D. P. (2006). The role of BRCA1 in transcriptional regulation and cell cycle control. *Oncogene*, *25*(43), 5854–5863. https://doi.org/10.1038/sj.onc.1209872

Newman, B., Austin, M. A., Lee, M., & King, M. C. (1988). Inheritance of human breast cancer: Evidence for autosomal dominant transmission in high-risk families. *Proceedings of the National Academy of Sciences of the United States of America*, *85*(9), 3044–3048. https://doi.org/10.1073/pnas.85.9.3044

Ng., R. K., & Gurdon, J. B. (2008). Epigenetic memory of an active gene state depends on histone H3.3 incorporation into chromatin in the absence

of transcription. *Nature Cell Biology*, *10*(1), 102–109. https://doi.org/10.1038/ncb1674

Nielsen, F. C., Van Overeem Hansen, T., & Sørensen, C. S. (2016). Hereditary breast and ovarian cancer: New genes in confined pathways. *Nature Reviews Cancer*, *16*(9), 599–612. https://doi.org/10.1038/nrc.2016.72

Niroula, A., Urolagin, S., & Vihinen, M. (2015). PON-P2: Prediction Method for Fast and Reliable Identification of Harmful Variants. *Plos One*, *10*(2), e0117380. https://doi.org/10.1371/journal.pone.0117380

Niroula, A., & Vihinen, M. (2016). Variation Interpretation Predictors: Principles, Types, Performance, and Choice. *Human Mutation*, *37*(6), 579–597. https://doi.org/10.1002/humu.22987

Norris, A., Bianchet, M. A., & Boeke, J. D. (2008). Compensatory interactions between Sir3p and the nucleosomal LRS surface imply their direct interaction. *PLoS Genetics*, *4*(12), 18–20. https://doi.org/10.1371/journal.pgen.1000301

Oddoux, C., Struewing, J. P., Clayton, C. M., Neuhausen, S., Brody, L. C., Kaback, M., … Offit, K. (1996). The carrier frequency of the BRCA2 617delT mutation among Ashkenazi Jewish individuals is aproximately 1%. *Nat Genet*, *14*(2), 188–190.

Padilla, N., Moles-Fernández, A., Riera, C., Montalban, G., Özkan, S., Ootes, L., … de la Cruz, X. (2019). BRCA1- and BRCA2-specific in silico tools for variant interpretation in the CAGI 5 ENIGMA challenge. *Human Mutation*, *40*(9), 1593–1611. https://doi.org/10.1002/humu.23802

Paluch-Shimon, S., Cardoso, F., Sessa, C., Balmana, J., Cardoso, M. J., Gilbert, F., … on behalf of the ESMO Guidelines Committee. (2016). Prevention and screening in BRCA mutation carriers and other breast/ovarian hereditary cancer syndromes: ESMO clinical practice guidelines for

cancer prevention and screening. *Annals of Oncology*, *27*(5), v103–v110. https://doi.org/10.1093/annonc/mdw327

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn : Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://doi.org/10.1016/j.molcel.2012.08.019

Pepenella, S., Murphy, K. J., & Hayes, J. J. (2014). Intra- and inter-nucleosome interactions of the core histone tail domains in higher-order chromatin structure. *Chromosoma*, *123*(1–2), 3–13. https://doi.org/10.1007/s00412-013-0435-8

Piovesan, D., Minervini, G., & Tosatto, S. C. E. (2016). The RING 2.0 web server for high quality residue interaction networks. *Nucleic Acids Research*, *44*(W1), W367–W374. https://doi.org/10.1093/nar/gkw315

Pires, D. E. V., Ascher, D. B., & Blundell, T. L. (2014). MCSM: Predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, *30*(3), 335–342. https://doi.org/10.1093/bioinformatics/btt691

Plon, S. E., Eccles, D. M., Easton, D., Foulkes, W. D., Genuardi, M., Greenblatt, M. S., … Tavtigian, S. V. (2008). Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. *Human Mutation*, *29*(11), 1282–1291. https://doi.org/10.1002/humu.20880

Ponder, B., Pharoah, P. D. P., Ponder, B. A. J., Lipscombe, J. M., Basham, V., Gregory, J., … Dunning, A. (2000). Prevalence and penetrance of BRCA1 and BRCA2 mutations in a population-based series of breast cancer cases. *British Journal of Cancer*, *83*(10), 1301–1308. https://doi.org/10.1054/bjoc.2000.1407

Pons, T., Vazquez, M., Matey-Hernandez, M. L., Brunak, S., Valencia, A., & Izarzugaza, J. M. G. (2016). KinMutRF: A random forest classifier of sequence variants in the human protein kinase superfamily. *BMC Genomics*, *17*(Suppl. 2), 396. https://doi.org/10.1186/s12864-016-2723-1

Ransburgh, D. J. R., Chiba, N., Ishioka, C., Toland, A. E., & Parvin, J. D. (2011). *The effect of BRCA1 missense mutations on homology directed recombination*. *70*(3), 988–995. https://doi.org/10.1158/0008-5472.CAN-09-2850.The

Rebbeck, T. R., Friebel, T. M., Friedman, E., Hamann, U., Huo, D., Kwong, A., … Nathanson, K. L. (2018). Mutational spectrum in a worldwide study of 29,700 families with BRCA1 or BRCA2 mutations. *Human Mutation*, *39*(5), 593–620. https://doi.org/10.1002/humu.23406

Ribeiro, J., Ríos-Vera, C., Melo, F., Schüller, A., & Valencia, A. (2019). Calculation of accurate interatomic contact surface areas for the quantitative analysis of non-bonded molecular interactions. *Bioinformatics*, *35*(18), 3499–3501. https://doi.org/10.1093/bioinformatics/btz062

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., … Rehm, H. L. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, *17*(5), 405–423. https://doi.org/10.1038/gim.2015.30

Ricketts, M. D., Frederick, B., Hoff, H., Tang, Y., Schultz, D. C., Rai, T. S., … Marmorstein, R. (2015). Ubinuclein-1 confers histone H3.3-specific-binding by the HIRA histone chaperone complex. *Nature*

Communications, *6*, 1–11. https://doi.org/10.1038/ncomms8711

Riera, C., Lois, S., & De la Cruz, X. (2014). Prediction of pathological mutations in proteins: The challenge of integrating sequence conservation and structure stability principles. *Wiley Interdisciplinary Reviews: Computational Molecular Science*. https://doi.org/10.1002/wcms.1170

Riera, C., Lois, S., Dom Inguez, C., Fernandez-Cadenas, I., Montaner, J., Rodr Iguez-Sureda, V., & De La Cruz, X. (2015). Molecular damage in Fabry disease: Characterization and prediction of alpha-galactosidase A pathological mutations. *Proteins*, *83*, 91–104. https://doi.org/10.1002/prot.24708

Riera, C., Padilla, N., & de la Cruz, X. (2016). The Complementarity Between Protein-Specific and General Pathogenicity Predictors for Amino Acid Substitutions. *Human Mutation*, *37*(10), 1013–1024. https://doi.org/10.1002/humu.23048

Roy, R., Chun, J., & Powell, S. N. (2012). BRCA1 and BRCA2: Different roles in a common pathway of genome protection. *Nature Reviews Cancer*, *12*(1), 68–78. https://doi.org/10.1038/nrc3181

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Russo, A., Calò, V., Bruno, L., Rizzo, S., Bazan, V., & Di Fede, G. (2009). Hereditary ovarian cancer. *Critical Reviews in Oncology/Hematology*, *69*(1), 28–44. https://doi.org/10.1016/j.critrevonc.2008.06.003

Sankaran, S., Crone, D. E., Palazzo, R. E., & Parvin, J. D. (2007). Aurora-A kinase regulates breast cancer-associated gene 1 inhibition of centrosome-dependent microtubule nucleation. *Cancer Research*, *67*(23), 11186–

11194. https://doi.org/10.1158/0008-5472.CAN-07-2578

Saredi, G., Huang, H., Hammond, C. M., Alabert, C., Bekker-Jensen, S., Forne, I., … Groth, A. (2016). H4K20me0 marks post-replicative chromatin and recruits the TONSL-MMS22L DNA repair complex. *Nature*, *534*(7609), 714–718. https://doi.org/10.1038/nature18312

Sawicka, A., & Seiser, C. (2012). Histone H3 phosphorylation - A versatile chromatin modification for different occasions. *Biochimie*, *94*(11), 2193–2201. https://doi.org/10.1016/j.biochi.2012.04.018

Schulmeister, A., Schmid, M., & Thompson, E. M. (2007). Phosphorylation of the histone H3.3 variant in mitosis and meiosis of the urochordate Oikopleura dioica. *Chromosome Research*, *15*(2), 189–201. https://doi.org/10.1007/s10577-006-1112-z

Schwarz, J. M., Cooper, D. N., Schuelke, M., & Seelow, D. (2014). Mutationtaster2: Mutation prediction for the deep-sequencing age. *Nature Methods*, *11*(4), 361–362. https://doi.org/10.1038/nmeth.2890

Shah, T., & Guraya, S. (2017). Breast cancer screening programs: Review of merits, demerits, and recent recommendations practiced across the world. *Journal of Microscopy and Ultrastructure*, *5*(2), 59. https://doi.org/10.1016/j.jmau.2016.10.002

Shendure, J., Findlay, G. M., & Snyder, M. W. (2019). Genomic Medicine–Progress, Pitfalls, and Promise. *Cell*, *177*(1), 45–57. https://doi.org/10.1016/j.cell.2019.02.003

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, *29*(1), 308–311. https://doi.org/10.1093/nar/29.1.308

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L. A., Edwards, K. J., … Gaunt, T. R. (2012). Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. *Human Mutation*, *34*(1), n/a-n/a. https://doi.org/10.1002/humu.22225

Sidoli, S., & Garcia, B. A. (2015). Properly reading the histone code by MS-based proteomics. *Proteomics*, *15*(17), 2901–2902. https://doi.org/10.1002/pmic.201500298

Spurdle, A. B., Healey, S., Devereau, A., Hogervorst, F. B. L., Monteiro, A. N. A., Nathanson, K. L., … Goldgar, D. E. (2012). ENIGMA-evidence-based network for the interpretation of germline mutant alleles: An international initiative to evaluate risk and clinical significance associated with sequence variation in BRCA1 and BRCA2 genes. *Human Mutation*, *33*(1), 2–7. https://doi.org/10.1002/humu.21628

Starita, L. M., Ahituv, N., Dunham, M. J., Kitzman, J. O., Roth, F. P., Seelig, G., … Fowler, D. M. (2017). Variant Interpretation: Functional Assays to the Rescue. *American Journal of Human Genetics*, *101*(3), 315–325. https://doi.org/10.1016/j.ajhg.2017.07.014

Starita, L. M., Young, D. L., Islam, M., Kitzman, J. O., Gullingsrud, J., Hause, R. J., … Fields, S. (2015). Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics*, *200*(2), 413–422. https://doi.org/10.1534/genetics.115.175802

Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shaw, K., & Cooper, D. N. (2012). The human gene mutation database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. *Current Protocols in Bioinformatics*, *39*, 1.13.1-1.13.20. https://doi.org/10.1002/0471250953.bi0113s39

Sunyaev, S. R. (2012). Inferring causality and functional significance of human coding dna variants. *Human Molecular Genetics*, *21*(R1), 10–17. https://doi.org/10.1093/hmg/dds385

Sunyaev, S., Ramensky, V., Koch, I., Lathe, W., Kondrashov, A. S., & Bork, P. (2001). Prediction of deleterious human alleles. *Human Molecular Genetics*, *10*(6), 591–597. https://doi.org/10.1093/hmg/10.6.591

Tachiwana, H., Osakabe, A., Shiga, T., Miya, Y., Kimura, H., Kagawa, W., & Kurumizaka, H. (2011). Structures of human nucleosomes containing major histone H3 variants. *Acta Crystallographica Section D: Biological Crystallography*, *67*(6), 578–583. https://doi.org/10.1107/S0907444911014818

Tatton-Brown, K., Loveday, C., Yost, S., Clarke, M., Ramsay, E., Zachariou, A., … Rahman, N. (2017). Mutations in Epigenetic Regulation Genes Are a Major Cause of Overgrowth with Intellectual Disability. *American Journal of Human Genetics*, *100*(5), 725–736. https://doi.org/10.1016/j.ajhg.2017.03.010

Tavtigian, S. V., Deffenbaugh, A. M., Yin, L., Judkins, T., Scholl, T., Samollow, P. B., … Thomas, A. (2006). Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *Journal of Medical Genetics*, *43*(4), 295–305. https://doi.org/10.1136/jmg.2005.033878

Tavtigian, S. V., Greenblatt, M. S., Lesueur, F., & Byrnes, G. B. (2008). In silico analysis of missense substitutions using sequence-alignment based methods. *Human Mutation*, *29*, 1329–1336. https://doi.org/10.1002/humu.20892

Tessadori, F., Giltay, J. C., Hurst, J. A., Massink, M. P., Duran, K., Vos, H. R., … Van Haaften, G. (2017). Germline mutations affecting the histone H4

core cause a developmental syndrome by altering DNA damage response and cell cycle control. *Nature Genetics*, *49*(11), 1642–1646. https://doi.org/10.1038/ng.3956

Thomas, P. D., Kejariwal, A., Guo, N., Mi, H., Campbell, M. J., Muruganujan, A., & Lazareva-Ulitsky, B. (2006). Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. *Nucleic Acids Research*, *34*(WEB. SERV. ISS.), 645–650. https://doi.org/10.1093/nar/gkl229

Thompson, D., & Easton, D. (2001). Variation in cancer risks, by mutation position, in BRCA2 mutation carriers. *American Journal of Human Genetics*, *68*(2), 410–419. https://doi.org/10.1086/318181

Thompson, Deborah, & Easton, D. F. (2003). Cancer Incidence in BRCA1 Mutation Carriers. *Obstetrical & Gynecological Survey*, *58*(1), 27–28. https://doi.org/10.1097/00006254-200301000-00016

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, *22*(22), 4673–4680. https://doi.org/10.1093/nar/22.22.4673

Toland, A. E., & Andreassen, P. R. (2017). DNA repair-related functional assays for the classification of BRCA1 and BRCA2 variants: A critical review and needs assessment. *Journal of Medical Genetics*, *54*(11), 721–731. https://doi.org/10.1136/jmedgenet-2017-104707

Touw, W. G., Baakman, C., Black, J., Te Beek, T. A. H., Krieger, E., Joosten, R. P., & Vriend, G. (2015). A series of PDB-related databanks for everyday needs. *Nucleic Acids Research*, *43*(D1), D364–D368. https://doi.org/10.1093/nar/gku1028

Udugama, M., Chang, F. T. M., Chan, F. L., Tang, M. C., Pickett, H. A., McGhie, J. D. R., … Wong, L. H. (2015). Histone variant H3.3 provides the heterochromatic H3 lysine 9 tri-methylation mark at telomeres. *Nucleic Acids Research*, *43*(21), 10227–10237. https://doi.org/10.1093/nar/gkv847

Vallée, M. P., Di Sera, T. L., Nix, D. A., Paquette, A. M., Parsons, M. T., Bell, R., … Tavtigian, S. V. (2016). Adding In Silico Assessment of Potential Splice Aberration to the Integrated Evaluation of BRCA Gene Unclassified Variants. *Human Mutation*, *37*(7), 627–639. https://doi.org/10.1002/humu.22973

Van Hooser, A., Goodrich, D. W., David Allis, C., Brinkley, B. R., & Mancini, M. A. (1998). Histone H3 phosphorylation is required for the initiation, but not maintenance, of mammalian chromosome condensation. *Journal of Cell Science*, *111*(23), 3497–3506.

Venkitaraman, A. R. (2014). Cancer suppression by the chromosome custodians, BRCA1 and BRCA2. *Science*, *34*(6178), 1470–1475. https://doi.org/10.1126/science.1252230

Vihinen, M. (2012). How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. *BMC Genomics*, *13*, S2. https://doi.org/10.1186/1471-2164-13-S4-S2

Vihinen, M. (2013). Guidelines for Reporting and Using Prediction Tools for Genetic Variation Analysis. *Human Mutation*, *34*, 275–282. https://doi.org/10.1002/humu.22253

Vihinen, M. (2014). Majority vote and other problems when using computational tools. *Human Mutation*, *35*(8), 912–914. https://doi.org/10.1002/humu.22600

Vollmuth, F., Blankenfeldt, W., & Geyer, M. (2009). Structures of the dual

bromodomains of the P-TEFb-activating protein Brd4 at atomic resolution. *Journal of Biological Chemistry*, *284*(52), 36547–36556. https://doi.org/10.1074/jbc.M109.033712

Vollmuth, F., & Geyer, M. (2010). Interaction of propionylated and butyrylated histone H3 lysine marks with Brd4 bromodomains. *Angewandte Chemie - International Edition*, *49*(38), 6768–6772. https://doi.org/10.1002/anie.201002724

Wei, Q., & Dunbrack, R. L. (2013). The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics. *PLoS ONE*, *8*(7), e67863. https://doi.org/10.1371/journal.pone.0067863

Wen, H., Li, Y., Xi, Y., Jiang, S., Stratton, S., Peng, D., … Shi, X. (2014). ZMYND11 links histone H3.3K36me3 to transcription elongation and tumour suppression. *Nature*, *508*(7495), 263–268. https://doi.org/10.1038/nature13045

Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., … Stratton, M. R. (1995). Identification of the breast cancer susceptibility gene BRCA2. *Nature*, *378*, 667–668.

Wu, G., Broniscer, A., McEachron, T. A., Lu, C., Paugh, B. S., Becksfort, J., … Baker, S. J. (2012). Somatic histone H3 alterations in pediatric diffuse intrinsic pontine gliomas and non-brainstem glioblastomas. *Nature Genetics*, *44*(3), 251–253. https://doi.org/10.1038/ng.1102

Wu, K., Hinson, S. R., Ohashi, A., Farrugia, D., Wendt, P., Tavtigian, S. V., … Couch, F. J. (2005). Functional evaluation and cancer risk assessment of BRCA2 unclassified variants. *Cancer Research*, *65*(2), 417–426.

Yang, S., Zheng, X., Lu, C., Li, G. M., Allis, C. D., & Li, H. (2016). Molecular basis for oncohistone H3 recognition by SETD2 methyltransferase. *Genes and Development*, *30*(14), 1611–1616.

https://doi.org/10.1101/gad.284323.116

Yarden, R. I., Pardo-Reoyo, S., Sgagias, M., Cowan, K. H., & Brody, L. C. (2002). BRCA1 regulates the G2/M checkpoint by activating Chk1 kinase upon DNA damage. *Nature Genetics*, *30*(3), 285–289. https://doi.org/10.1038/ng837

Young, N. L., DiMaggio, P. A., & Garcia, B. A. (2010). The significance, development and progress of high-throughput combinatorial histone code analysis. *Cellular and Molecular Life Sciences*, *67*(23), 3983–4000. https://doi.org/10.1007/s00018-010-0475-7

Yue, P., Li, Z., & Moult, J. (2005). Loss of protein structure stability as a major causative factor in monogenic disease. *Journal of Molecular Biology*, *353*, 459–473. https://doi.org/10.1016/j.jmb.2005.08.020