Next generation sequencing in medicine

Najmeh Sadaat Alirezaie, MD, MSc

Department of Human Genetics
Faculty of Medicine
McGill University, Montreal, Quebec, Canada

August 2018

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Dedication

To my family, a constant source of support and encouragement during the challenges of graduate school and life.

Abstract

In recent decades, there has been immense progress in DNA sequencing technologies. One domain where these techniques are undoubtedly transformative is medicine. Understanding the underlying genomic and genetic factors in diseases is important in diagnosis, treatment, and identification of at-risk family members. In this regard, cancer genome analysis provides a better understanding of the underlying tumorigenic mechanisms and gives hope of finding genetic events that contribute to metastasis, recurrence, and therapeutic resistance, which are the most common causes of cancer fatality. In the first part of this thesis, we focus on uncovering the genetic predisposition factors in familial pancreatic cancer. By applying whole exome sequencing (WES) to 109 familial pancreatic cases and performing a filter-based candidate gene approach focused on DNA repair genes, we propose FAN1, NEK1 and RHNO1 as the strongest candidates. In the second part, we investigate the mechanisms underlying resistance to chemotherapy in Triple-negative breast cancer (TNBC). WES- and RNA-seq analysis on serial tumor biopsies during chemotherapy suggests that there are two major factors associated with chemotherapy: RAD21 gene amplification and presence of immune response. The results of this study suggest that RAD21 may be both a marker and a target to overcome drug resistance in TNBCs, and combination of chemotherapy and immunotherapy (anti-PD-1/PD-L1 monoclonal antibody therapies) would improve outcomes of TNBC patients, especially PD-L1-positive patients with low tumor-infiltrating lymphocytes (TILs) in tumors.

Furthermore, motivated by accurately identifying pathogenic variants from WES data, I developed an improved ensemble machine learning method, ClinPred, to predict in silico and select pathogenic variants in large-scale sequencing studies. Through rigorous testing, while avoiding problems common in machine learning, such as overfitting and circularity, I showed that

ClinPred outperforms all currently available prediction methods, achieving the highest Area Under the Curve (AUC) score and increasing both the specificity and sensitivity in different test datasets. It also obtained the best performance according to various other metrics.

Together, our work demonstrates the value of next generation sequencing techniques as powerful tools for understanding the mechanisms of various diseases and their subsequent implications for the clinical arena.

Résumé

Au cours des dernières décennies, d'immenses progrès ont été réalisés dans les technologies de séquençage de l'ADN. Un domaine où ces techniques sont sans aucun doute transformatrices est la médecine. Comprendre les facteurs génomiques et génétiques sous-jacents des maladies est important dans le diagnostic, le traitement et l'identification des membres de la famille du patient à risque. À cet égard, l'analyse de la génomique du cancer permet une meilleure compréhension des mécanismes tumorigéniques sous-jacents et nous donne espoir de trouver les événements génétiques qui mènent aux métastases, à la réapparition du cancer et à la résistance aux traitements thérapeutiques, qui sont les causes les plus fréquentes de décès par le cancer. Dans la première partie de cette thèse, nous nous concentrons sur la découverte des facteurs de prédisposition génétique au niveau du cancer du pancréas familial. En appliquant le séquençage de l'exome entier (WES) à 109 cas pancréatiques familiaux et en utilisant une approche, qui filtre les résultats du séquençage, basée sur des gènes candidats de la réparation de l'ADN, nous proposons FAN1, NEK1 et RHNO1 comme étant les candidats les plus forts. Dans la deuxième partie, nous étudions les mécanismes sous-jacents à la résistance à la chimiothérapie du cancer du sein triple négatif. L'analyse du séquençage de l'exome entier et séquençage haut débit d'ARN des biopsies sur la tumeur lors de plusieurs étapes de la chimiothérapie, suggère qu'il y a deux facteurs majeurs associés à la chimiothérapie : l'amplification du gène RAD21 et la présence d'une réponse immunitaire. Les résultats de cette étude suggèrent que RAD21 pourrait être à la fois un indicateur et une cible pour vaincre la pharmacorésistance du cancer du sein triple négatif, et qu'une combinaison de chimiothérapie et d'immunothérapie (anti-PD-1 / PD-L1) améliorerait les résultats des patients atteints du cancer du sein triple négatif, en particulier les patients atteints de tumeurs PD-L1-positifs / faibles enTILs.

De plus, motivée par l'identification précise des variantes pathogènes obtenus à partir des données du séquençage de l'exome entier, j'ai développé une méthode d'apprentissage automatique d'ensemble améliorée, ClinPred, pour prédire in silico et sélectionner des variantes pathogènes dans des études de séquençage à grande échelle. Grâce à des tests rigoureux dans lesquels sont évités les problèmes communs en apprentissage automatique, comme le surapprentissage ou la circularité, j'ai démontré que ClinPred surpasse toutes les méthodes de prédiction actuellement disponibles, en obtenant le score le plus élevé pour la superficie sous la courbe et en ayant plus de spécificité et sensibilité lors de différentes analyses. ClinPred a obtenu les meilleures performances en fonction de divers autres indicateurs.

Ensemble, nos travaux démontrent la valeur des techniques de séquençage de nouvelle génération en tant qu'outils puissants pour comprendre les mécanismes de diverses maladies et leurs implications subséquentes pour l'arène clinique.

Table of contents

List of abbreviations

ABC	ATP-binding cassette
ACMG	American College of Medical Genetics and Genomics
AF	Allele frequency
AUC	Area under the curve
BWA	Burrows-Wheeler Aligner
CADD	Combined Annotation-Dependent Depletion
CAROL	Combined Annotation Scoring Tool
Condel	CONsensus DELeteriousness score of missense mutations
COSMIC	Catalogue of Somatic Mutations in Cancer
dbSNP	The Single Nucleotide Polymorphism Database
DDR	DNA damage response
EVS	Exome Variant Server
ER	Estrogen receptor
EXAC	Exome Aggregation Consortium
FDR	First degree relative
FPC	Familial pancreatic cancer
FFPE	Formalin-fixed, paraffin-embedded
GATK	Genome Analyzer Tool Kit
GERP	Genomic Evolutionary Rate Profiling
gnomAD	Genome Aggregation Database
HER2	Human epidermal growth factor receptor 2

HGMD	Human Gene Mutation Database
IGV	Integrated Genomics Viewer
IHC	Immunohistochemistry
IM	Immunomodulatory
INDEL	Insertions or deletions
LOF	Loss of function
LOH	Loss of heterozygosity
LRT	Likelihood ratio test
M-CAP	Mendelian Clinically Applicable Pathogenicity
NGS	Next generation sequencing
OMIM	Online Mendelian Inheritance in Man
PC	Pancreatic cancer
PDA	Pancreatic ductal adenocarcinoma
PolyPhen-2	Polymorphism Phenotyping v2
PR	Progesterone receptor
PROVEAN	Protein Variation Effect Analyzer
PTV	Protein-truncating variant
QC	Quality control
REVEL	Rare exome variant ensemble learner
SIFT	Sorting Intolerant from Tolerant
SNV	Single nucleotide variants
TNBC	Triple negative breast TNBC
TSG	Tumor suppressor gene

UniProt	The Universal Protein Resource
WES	Whole exome sequencing
WGS	Whole genome sequencing

List of Figures

List of Tables

Acknowledgment

Scientific publications are almost never written by just one person. I very much appreciate the chance to work with so many bright and supportive people. First and foremost, I would like to express my sincere gratitude to my supervisor Dr. Jacek Majewski, for his continuous help, encouragement, and supportive attitude throughout my research. His guidance and understanding helped to elevate and promote my work to a level that would not have been otherwise possible. I am deeply grateful for his patience and willingness to devote his time and expertise throughout this dissertation, and my graduate education.

I also would like to extend my appreciation to Dr. George Zogopoulos, my first collaborator, and members of his research group including Alyssa Smith, who provided me with her samples as well as her continuous support. I am also indebted to Dr. Mark Basik, a member of the advisory committee and a clinical collaborator. I benefited a lot from his insight, valuable comments and suggestions throughout the years. I would like to address a special thanks to Dr. Basik research group, without whom the objectives of this study would not exist. I especially want to thank Dr. Yasser Riazalhosseini and Dr. Hamed Najafabadi, who provided me with helpful guidance and constructive suggestions to improve my thesis.

I thank the former and current members of the Majewski lab, Claudia Kleinman, Jeremy Schwartzenbruber, Amandine Bemmo, Martine Tetreault, Simon Papillon, Somayyeh Fahiminiya, Javad Nadaf, Rui Li, Octavia Dancu, Hamid Nikbakht, Eric Bareke, Jian Carrot-Zhang, Anissa Djedid, Haifen Chen, Nargess Farhang, Cynthia Horth, Frank Hu, and Matt Osmon. Thank you for not only being my coworkers, but my friends, and making it a pleasure to come to the lab every day. I thank Hamid, Octavia, Eric and Simon who contributed to so many aspects of my PhD

experience. I am also very grateful to Dr. Aimee Ryan and Ross MacKay for their help and support throughout my journey as a graduate student in the department of human genetics.

Last but not least, I owe a special appreciation to my dearest family, to my parents who have never stopped believing in me and have always encouraged me to keep going; my husband, for joining me, supporting me in this adventure and reminding me every day that I could accomplish this goal when often I would forget. I thank my children, Hamed and Niki, who are the pride and joy of my life. I love you and I appreciate all your patience and support during mommy's PhD studies.

Originality and Significance

The work presented in this thesis represents an original contribution to medicine, especially cancer genomics. In Chapter 3, the genetic predisposition factors in familial pancreatic cancer is studied and we propose FAN1, NEK1 and RHNO1 as novel genes that may have a role in hereditary PC. The work in Chapter 4 innovatively examines the mechanism of resistance in triple negative breast cancer. Our findings suggest that there are two major driving factors in chemotherapy response in the neoadjuvant setting in TNBCs, which provides new opportunities for the development of novel therapeutic strategies. The work in Chapter 5 develops a novel algorithm to improve variant pathogenicity prediction. This method adds value over pre-existing variant scoring algorithms and is useful for those who perform clinical variant classification for diagnosis of genetic disorders from WES data and will fit into current ACMG-CAP guidelines.

Format of the Thesis

This thesis has been written in the traditional style and is composed of a general introduction and literature review, material and methods, three original research chapters, general discussion, conclusions and future directions. Chapter 3 contains portions of a manuscript that was published. Permissions were obtained for reproducing published materials. Version of Chapter 4 is under preparation for publication. Chapter 5 is accepted in American Journal of Human Genetics pending formatting verification. The format of the thesis respects the McGill University Guidelines for Thesis Preparation.

Contribution of the Authors

All parts of chapters 3, 4, and 5 are original research that have made a distinct contribution to scientific knowledge.

Chapter 3: Candidate DNA repair susceptibility genes in familial pancreatic cancer

Najmeh Alirezaie performed the bioinformatics analysis, interpreted the data and contributed to the novel findings. Alyssa smith collected samples and performed the Sanger sequencing analysis and segregation analysis. Ashton Connor, Robert Grant, Iris Selander, Claire Bascuñana, Ayelet Borgida, Anita Hall, Thomas Whelan, Spring Holter, Treasa McPherson,Sean Cleary, Gloria M. Petersen, Atilla Omeroglu, Emmanouil Saloustros, John McPherson, and Lincoln D. Stein provided samples and critical input. Jacek Majewski oversaw all bioinformatics analysis. William Foulkes and Steven Gallinger oversaw interpretation of data. George Zogopoulos designed the project and oversaw all aspects of the project.

Chapter 4: The Genomic landscape of triple negative breast cancer in neoadjuvant chemotherapy

Najmeh Alirezaie performed the WES data analysis of the triple negative breast cancer study, helped for the RNA-seq analysis and contributed to the novel findings. Eric Bareke performed RNA-seq analysis. Ewa Przybytkowski and Sheida Nabavi performed the copy number changes analysis. Luca Cavallone performed ddPCR analysis. Josiane Lafleur, Cathy Lan ,Federico Discepola, Manuela Pelmus, , Olga Aleynikova, André Robidoux, Catalin Mihalcioiu , Josée Anne Roy, Elizabeth Marcus contributed to data collection, sample genotyping and provided study subjects. Jacek Majewski oversaw all bioinformatics analysis. Mark Basik and Adriana Aguilar-Mahecha designed the project and oversaw all aspects of the project.

Chapter 5: ClinPred - A prediction method to identify clinically relevant nonsynonymous single nucleotide variants.

Najmeh Alirezaie designed the project, developed and implemented the algorithm. Kristin D. Kernohan and Taila Hartley helped for clinical patients' data collection. Jacek Majewski designed the project and oversaw the algorithm development. Toby Dylan Hocking designed and oversaw the algorithm development.

Chapter 1. General Introduction and Literature Reviews

Genetics has advanced our understanding of many disease mechanisms and has had a large impact on medicine. Knowledge of the genetic and molecular basis of disorders is helping in diagnosis, guiding therapy and developing new drugs. By providing the order of nucleotides within a DNA molecule, sequencing technologies have had a great impact in genetics and have enabled us to better understand the molecular basis of disorders.

Sequencing techniques have a long history. The first success in sequencing was the protein sequence of insulin in the early 1950s by Sanger (Sanger 1958). Later in the 1960s Holley et al. for the first time determined the complete nucleotide sequence of an alanine transfer RNA (Holley, et al. 1965). However, DNA sequencing was much harder and more complicated than expected. Maxam and Gilbert developed a chemical degradation method for DNA sequencing in 1977 (Gilbert and Maxam 1977). The same year, Sanger and colleagues developed a chain-termination technique, which has been since optimized and used to obtain the first draft of the human genome (Human Genome Project) (Sanger, et al. 1977).

The main limitations of these two classical DNA sequencing techniques (also known as first generation sequencing) are low throughput, problems in detecting low frequency variants (e.g. somatic mutation in cancer and mosaic mutation) and cost. Therefore, newer approaches were developed to sequence reads in parallel for a faster and more cost-efficient way (Morey, et al. 2013). Although the first next generation sequencing (NGS) equipment became available in 2004 after the human genome project, these technologies have rapidly changed genetics, genomics and medicine (Morey, et al. 2013). Comparison between an individual sequence and a normal reference genome makes it feasible to identify variants in an individual's DNA; therefore, likely

disease-causing variants can be spotted. Next-generation high-throughput sequencing technologies have been widely used in different type of diseases to identify germ-line mutations underlying Mendelian disorders, complex diseases and somatic mutations in various cancers.

Whole genome sequencing (WGS) currently represents the most complete, comprehensive strategy for variant detection. However, there is significant variability in sequencing efficiency across the genome when we use a genome sequencing approach. Different regions can have different coverage and many regions of interest can be missed (Majewski, et al. 2011; Ng, et al. 2010). Moreover, it is costly when a large sample size is needed. In contrast, exome-sequencing targets only protein coding sequences, which enables the analysis of the coding regions of more than 20,000 genes. The human genome comprises approximately 180,000 exons, which is less than 2% of the entire genome. Although a whole exome is a small fraction of a genome, it harbors most of the known diseases causing DNA changes that lead to genetic disorders (Majewski, et al. 2011). Whole exome sequencing (WES) produces considerably fewer sequencing reads and a smaller, more manageable data set (4–5 GB of sequencing per exome compared to ~90 GB per whole genome). However, it produces higher sequence coverage. The cost of WES is also lower than WGS. Therefore, through a faster and easier analysis, WES is highly suitable for discovery of mutations (Harding and Robertson 2014; Yang, et al. 2014).

1.1. Application of whole-exome sequencing data in medicine

One of the main common goals of the clinical genetics field is to discover the causes of diseases and find the best way to address them both in diagnosis and therapy. In a clinical setting, targeted sequencing, which applies a gene panel relevant to the patient's disease, was the first next-generation sequencing (NGS) technology to be used. Over the past years, WES has been increasingly applied to identify disease-causing genes in medical research areas. The use of exome sequencing has been successful in the characterization of many rare diseases (Biesecker 2010). In addition, WES has been useful in cancer-genomics and has been helpful in understanding the mechanisms underlying specific cancers and identifying new biomarkers and/or drug targets (Grossmann, et al. 2011).

After the significant success of whole exome sequencing in the research area, diagnostic and clinical genetics fields started to use this technique to improve medical care and explore challenges available in this area in a cost-effective, highly efficient way. This technology is now widely accepted in clinical laboratories, and in the near future will change the landscape of clinical testing and diagnostics.

1.1.1. Application of WES in Mendelian diseases

Exome sequencing has been applied to find the molecular basis of rare Mendelian disorders in many studies. Today clinical WES is implemented in either finding disease genes or diagnosis for these patients in some countries. Mendelian diseases are rare diseases that occur at a rate of 40 to 82 per 1000 live births. Epidemiologic studies suggest approximately 8% of people are identified as having a genetic disorder before adulthood when considering all congenital anomalies as part

of the genetic load (Yang, et al. 2013). Therefore, rare genetic disorders affect substantial numbers of people. This attracts considerable interest in rare diseases in both the research and clinical diagnostic fields.

To date, researchers have identified around 7000 rare inherited disorders, although they were successful in characterizing the molecular basis in only almost half of them (Frebourg 2014). In recent years, application of WES has made it possible to identify novel disease-causing variants and genes for rare diseases. In addition, it has been successful in expanding knowledge about the phenotype of known genes.

One of the problems remaining in rare diseases is the correct diagnosis. Diagnosis is based on clinical symptoms, radiographic features, biopsy findings, analysis of metabolites and genomic tests such as karyotyping (Yang, et al. 2013), though genetic conditions usually have a wide range of clinical features common to different diagnoses. As a result, the majority of patients remain undiagnosed or misdiagnosed (Williams and Hegde 2013).

Undiagnosed patients present various signs and symptoms that are unclear or atypical, and consequently no definite diagnosis is made for them (Pinxten and Howard 2014; Soden, et al. 2014). They usually undergo extensive evaluations, which are time-consuming and costly, before a diagnosis is made (Srivastava, et al. 2014). Moreover, some of the rare diseases are clinically unrecognizable or have many genes involved in them. For example, cardiomyopathy involves over 50 genes or polyneuropathies associated with over 70 genes. In some cases, different syndromes can explain patients' symptoms at a younger age. Traditional methods such as Sanger sequencing are hard to use for these patients, as it is necessary to know in advance the genes required for sequencing (Harding and Robertson 2014). Ultimately, WES has a clear advantage as a diagnostic tool in this area since no pre-selection of genes is required and it allows the screening of the genes

not suspected to be associated with the disorder and to manage patients at lower cost (Harding and Robertson 2014; Williams and Hegde 2013). The lack of a diagnosis can have considerable adverse effects, such as failure to identify potential treatments, and failure to recognize the risk of recurrence in subsequent pregnancies. Therefore, the NIH Undiagnosed Diseases Program (UDP) suggests next generation sequencing as a helpful technique in the diagnosis of complicated and challenging medical conditions (Gahl, et al. 2012).

The first paper that emphasized exact diagnosis by using WES was published in 2009. A patient was misdiagnosed as having Bartter syndrome. However, after performing WES, the researchers found homozygous missense mutation in SLC26A3, the known congenital chloride diarrhea locus. The doctors re-evaluated the patient and found the symptoms could be related to this disease. Therefore, the treatment plan was changed to the proper one (Choi, et al. 2009). In another study, exome sequencing helped to identify and treat an unknown intestinal disorder in a patient who went through prolonged medical examinations and several surgeries, without specific disease diagnosis. After exome sequencing and analysis, the physicians were able to develop a treatment plan based on the correct diagnosis (X-linked inhibitor of apoptosis deficiency), which improved the overall outcome (Worthey, et al. 2011).

WES has even been able to detect disease variants that were not revealed by previous genetic tests—probably because of their lower sensitivity or poor design. For instance, in a study published by Landouré et al., WES was applied on a sample from a patient diagnosed as Charcot-Marie-Tooth type 2 (CMT2). The analysis revealed a mutation in the TRPV4 gene. This mutation was missed initially by Sanger sequencing as the SNP was located in the primer (Landoure, et al. 2012). This study again emphasized applicability of the WES technique in a clinical setting.

1.1.2. Application of WES in cancer

Cancers arise as the result of genomic alterations such as point mutations, copy number alterations and structural rearrangements in cancer cells (Mardis and Wilson 2009). The identification of the Philadelphia chromosome by cytogenetic techniques in chronic myeloid leukemia was the first success in finding genetic abnormalities causing cancer (Nowell and Hungerford 1960). In the 1980s and 1990s, linkage analysis helped to find many cancer susceptibility genes, especially tumour suppressors (Foulkes 2008).

Since cancer is a disease of genes, advancements in DNA sequencing technologies have a significant impact on detection, understanding the mechanism underlying cancer pathogenesis, as well as the management and treatment of the patients. By sequencing an individual's genome, we are no longer limited to mutations detectable by traditional linkage studies and targeted re-sequencing of candidate genes, and are able to efficiently detect somatic changes including single nucleotide substitutions, deletions, insertions, copy number variants, chromosomal rearrangements and microbial infections (Meyerson, et al. 2010). These technologies have enabled national and international projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) to systematically catalogue mutations in a wide variety of cancers and continue to uncover genomic aberrations in cancer (Chmielecki and Meyerson 2014; Wang and Wheeler 2014). Whole exome sequencing has led to the understanding of mechanisms underlying specific cancers and identifying new biomarkers and/or drug targets, and it provides unprecedented opportunities to study inherited and acquired genetic variants in cancer (Wang and Wheeler 2014).

Numerous cancer susceptibility genes associated with high risks of cancer among carriers have been identified, for example, BRCA1 and BRCA2 in breast cancer (Foulkes 2008). More

recent WES studies have identified new cancer susceptibility genes. The successful identification of germline mutations in PALB2, a susceptibility gene for familial pancreatic cancer, has been followed by other studies to uncover many other cancer susceptibility genes (Jones, et al. 2009). As examples, new studies are leading to the discovery of germ-line mutations in GATA2 in acute myeloid leukemia with Emberger syndrome (Ostergaard, et al. 2011), recurrent germline PAX5 mutations in pre-B cell acute lymphoblastic leukemia (Shah, et al. 2013), NPAT mutation in Hodgkin lymphoma (Saarinen, et al. 2011), BAP1 mutations in malignant mesothelioma (Testa, et al. 2011) as well as XRCC2 in familial breast cancer (Park, et al. 2012). Exome sequencing in familial pheochromocytoma (PCC) identified mutations in MAX. This gene was then added to the panel of other genes (RET, VHL, SDHA, SDHB, SDHC, SDHD, SDHAF2, NF1, and TMEM127) linked to familial PCC, to be simultaneously tested in familial PCC (Comino-Mendez, et al. 2011; Stadler, et al. 2014).

Moreover, cancer genome analysis provides a better understanding of the underlying tumorigenic mechanisms (Zhang, et al. 2013). The lower cost of exome sequencing makes it possible to carry on studies with a large sample size. This opportunity helps to have greater power for detecting recurrent somatic mutations. WES studies on gastric cancer identified recurrent somatic mutations in the chromatin-remodeling gene ARID1A and alterations in the cell adhesion gene FAT4, a member of the cadherin gene family (Wang, et al. 2011; Zang, et al. 2012). A recent study by our lab on small cell carcinoma of the ovary hypercalcemic type (SCCOHT) succeeded in identifying a mutation in the chromatin-remodeling gene SMARCA4. This finding suggests that developing drugs to target this gene might have widespread benefits because this gene has been implicated in various primary cancers, such as kidney and pediatric brain tumors (Witkowski, et al. 2014). Other examples of these successful somatic mutation discoveries are mutations in

isocitrate dehydrogenase 1 in glioblastoma (Parsons, et al. 2008), frequent mutations of the SWI/SNF complex gene PBRM1 in primary clear cell renal carcinoma (ccRCC) (Varela, et al. 2011), somatic mutations in histone H3.3 and chromatin remodelling genes in pediatric glioblastoma (Schwartzentruber, et al. 2012).

One of the biggest potential applications of NGS is in the area of personalized genomic medicine. WES is not only helping to find driver mutations, but also gives hope to find genetic events leading to metastasis, recurrence and therapeutic resistance, which are the most important reasons of the death toll due to cancer (Majewski, et al. 2011). WES can also be used as complementary to histopathological analysis to identify subpopulations that may benefit from targeting specific mutations, leading to better prognosis (Chmielecki and Meyerson 2014). For instance, exome sequencing revealed SF3B1 mutation is associated with better prognosis in myelodysplastic syndrome (MDS) with ring sideroblasts (Rabbani, et al. 2014). Another study on endometrial cancer patients revealed POLE as a prognostic marker in this type of cancer (Wang and Wheeler 2014). The genomic information provided by applying WES on cancer samples helps to identify not only cancer-associated genes and risks, but also pharmacogenomics markers. These data provide a guide to better treatment choices and designing new therapeutic protocols based on the genomic traits of tumours (Ciriello, et al. 2013; Stadler, et al. 2014).

The next parts of this introduction will first review pancreatic and breast cancers as the main two objectives of this thesis, and then I will discuss exome sequencing analysis and its challenges.

1.2. Pancreatic Cancer

The pancreas is a glandular organ in the digestive system with both endocrine and exocrine functions. Pancreatic cancer accounts for 2% of all cancer diagnoses, and it affected 367,000 people in 2015 worldwide. More than 50% of these patients were diagnosed in high-income countries (Ferlay, et al. 2015). Although the incidence rate is the tenth most common cancer worldwide, it is the fourth ranking cause of death due to cancer (Cotterell 2014; Klein, et al. 2002). More recently, it surpassed breast cancer as the third leading cause of death in the United States and is projected to become the second in North America by 2020 (Rahib, et al. 2014; Yabar and Winter 2016). Pancreatic cancer is one of the most lethal malignancies, with nearly as many deaths per year as incidence cases. An important factor in poor prognosis is late diagnosis when the cancer is in the advanced stage and has metastasized to other organs. This leads to a poor survival rate, with a five-year survival rate of less than 5% in pancreatic cancer, which has not changed in almost 50 years (Jemal, et al. 2010). Even though surgery is the best treatment option, almost 80% of the patients have metastasis, especially to distant organs, at the time of diagnosis, and consequently are not candidates for surgery at the time of diagnosis. Unfortunately, palliative therapies such as chemotherapy do not offer much increase in survival time for these patients (Cotterell 2014).

1.2.1 Classification of pancreatic cancer

A broad range of pathologically distinct types of neoplasms can originate in the pancreas. Recognizing different types of the neoplasms is important because they may have different clinical features and prognoses. Pancreatic tumors can be solid or cystic. Although cystic tumors are common and mostly benign, many will progress to invasive carcinomas without treatment. The

four main cystic neoplasms in the pancreas are intraductal papillary mucinous neoplasm (IPMNs), solid-psudopapillary tumour, musinous cystic neoplasm (MCNs) and serous cystic neoplasm (Wolfgang, et al. 2013).

Among the solid tumors, pancreatic ductal adenocarcinomas (PDA) are the most common and account for >90% of pancreatic cancers (Rustgi 2014). This tumor, as the name suggests, forms glands and infiltrates to an intensely desmoplastic stroma. The invasive ductal adenocarcinomas invade nerves, small veins nearby and the lymphatic system. They spread beyond the pancreas, preventing the patient from meeting criteria for surgical resection (Kleeff, et al. 2016; Wolfgang, et al. 2013).

Pancreatic neuroendocrine, the second most common solid type, originates in islet cells. It is characterized as a slow growing tumor with extensive neuroendocrine differentiation and can be treated by surgery with a 10-year survival rate of 45% (Kleeff, et al. 2016; Wolfgang, et al. 2013). Colloid carcinomas, another solid type, comprise 2% of pancreatic cancers and always arise in association with IPMNs. This mucin-producing neoplasm has a better prognosis than PDA as it is diagnosed at an earlier stage.

Rare pancreatic tumors include acinar cell carcinoma, pancreatoblastomas, adenosquamous, hepatoid, medullary, lypmphomas, sarcomas, giant cell tumors and undifferentiated carcinomas. Pathological diagnosis of medullary carcinomas is important for the treatment of the patient. Although this type of tumor is poorly differentiated, it has a good prognosis and could be sensitive to some immunotherapies (Kleeff, et al. 2016; Wolfgang, et al. 2013).

1.2.2. Hereditary pancreatic cancer

Pancreatic cancer usually affects elderly people. Most patients are more than 50 years old, with median age of 71 at diagnosis (Yabar and Winter 2016). Family history is the greatest risk factor for pancreatic ductal adenocarcinoma (PDA). Around 10 percent of PDAs occur in families. In a family affected by pancreatic cancer (FPC), at least one pair of first-degree relatives is affected. Although some environmental factors or stochastic effects may be the underlying cause in some FPCs, many of them are thought to be due to underlying genetic susceptibilities (Shi, et al. 2009; Wang, et al. 2007). An autosomal dominant pattern of inheritance was reported in 50-80% of families with FPC (Bartsch, et al. 2012). The estimated risk ratio of a person with a positive family history is 2.3 to 32 depending on the number and relatedness of affected relatives in a family (Hruban, et al. 2010). Estimated lifetime risk of developing pancreatic cancer in a person with one affected first degree relative (FDR) is 6%. This risk significantly increases if there are two, three or more affected FDR (10% and 40% respectively) (Klein, et al. 2004; Yabar and Winter 2016). Familial pancreatic cancer can occur alone (40%) or in association with other tumor spectrums in families (60%). In the German national collection of FPC, the first three associated tumor types with PC were breast (30%), colon (21%) and lung cancer (12%) (Schneider, et al. 2011).

1.2.2.1. Genetic risk factors

Several hereditary conditions have been associated with increased risk of pancreatic cancer such as hereditary pancreatitis, Peutz–Jeghers syndrome, Familial atypical multiple mole melanoma, Lynch syndrome, Cystic fibrosis, Breast and ovarian cancer syndrome, Ataxia telangiectasia, and Li–Fraumeni syndrome (Table 1.1). Besides these high lifetime risk factor conditions, there

is increasing evidence of the involvement of ABO blood group with different levels of risk (Wolfgang, et al. 2013).

Table 1. 1. Genetic syndromes associated with PC

(Modified from Wolfgang et al., A Cancer Journal for Clinicians, 2013)

Syndrome	Affected genes	Risk Estimate
Peutz–Jeghers syndrome	STK11 (also known as LKB1)	RR=132
Hereditary pancreatitis	PRSS1	RR=58
Familial atypical multiple mole melanoma	CDKN2A	RR=38
Lynch syndrome	MSH2, MLH1, MSH6, PMS and PMS2	RR=9
Cystic fibrosis	CFTR	RR=5
Breast and ovarian cancer syndrome	BRCA1, BRCA2 and PALB2	RR=2–4
Ataxia telangiectasia	ATM	Unknown- Elevated
Li–Fraumeni syndrome	TP53	Unknown
ABO blood group		OR=1.2

RR=relative risk, OR=odds ratio

1.2.2.2. Pancreatic cancer susceptibility genes

Although the role of genetics in pancreatic cancer is well known, the genetic basis of most of the familial pancreatic cancer patients is still unknown. Linkage analysis on a family with multiple FPC with autosomal dominant pattern suggested chromosome 4q32-34 as a potential candidate. Eventually PALLD gene, located in that region, was recommended as a causative gene, though further studies on other FPC cancers did not confirm this gene as the causative one (Bartsch, et al. 2012).

Currently BRCA2 is the most frequent germline mutated gene in FPC cancers with variable prevalence in patients (6-12% in patients with 2 FDR and 16% with 3 FDR affected patients in one study). It is also reported in PC without presence of breast cancer. Although the role of BRCA2 is well recognized in FPC, there is still debate about BRCA1 role, even though 2.2 fold increase risk was estimated in some research (Wolfgang, et al. 2013).

PALB2, another DNA repair gene that binds to BRCA2, is also an important gene with recognized role in FPC. Even though several studies had confirmed PALB2 role in other cancer types, its role in PC was first reported by Jones et al. who applied exome-sequencing on pancreatic cancer samples (Jones, et al. 2009).

ATM, a serine/threonine kinase involved in repairing double strand breaks in DNA, is another gene detected by whole genome sequencing as a FPC susceptibility gene. Germline mutation and loss of heterozygosity of the wild type allele were reported in a family with PC. As ATM deleterious variants are seen to be common in the population, further study is needed to confirm this role (Roberts, et al. 2012).

CDKN2A, a tumor suppressor gene, has a significant role in familial multiple melanoma as well as lifetime increased risk of pancreatic cancer.

STK11/LKB1 mutation is involved in Peutz–Jeghers syndrome that is inherited with autosomal dominant pattern. Patients with this syndrome have increased risk of suffering from lung, pancreatic, breast, colon, gastric and ovarian cancer in their lifetimes.

Beside germline mutations that are mentioned in table 1.1, germline mutation in other BRCA2 pathway genes such as FANC-C and FANC-G was identified in early onset pancreatic cancers (Roberts, et al. 2012; Wolfgang, et al. 2013). It is believed BRCA2, BRCA1, PALB2, ATM and other DNA repair genes' mutation would interfere in repairing the defective DNA and result in accumulation of mutations. Consequently, this genomic instability will be involved in cancer development.

Since the genetic basis underlying FPC in 85-90 percent of patients is still unknown, I hypothesize that there are other genes involved in FPC. Therefore, in chapter three of this thesis, I will explore this hypothesis.

1.3. Breast Cancer

Breast cancer is the most common malignancy among women worldwide and accounts for about 25.9% of all new cancer diagnoses in women with 2.4 million incident cases in 2015 (Miller, et al. 2016). It is ranked as the second most common cancer in Canada with more than 25,000 new diagnoses every year. Unfortunately, breast cancer occurs at an earlier age than the other type of cancers. As examples, the median age in breast cancer is 61 years in comparison to 70 years and 68 years for lung cancer and colorectal cancer respectively. Unfortunately, about 19% of breast cancer patients were diagnosed in between 30 to 49 years old age.

Although breast cancer burden is increasing globally, the incidence rate varies in different parts of the world near 10 fold. For instance, the breast cancer incidence is 27 per 100,000 in middle Africa and eastern Asia and is 92 in northern America (Althuis, et al. 2005; Miller, et al. 2016; Torre, et al. 2016). Although higher incidence was reported in the more developed parts of the world (Torre, et al. 2016), this high incidence is partly because of better screening procedures, which helps early detection, rather than genetic background (Chlebowski, et al. 2010).

Breast cancer is the cause of death in many people and it is the fifth cause of cancer deaths for both women and men (523 000 women and 10 000 men death in 2015). It was reported as the number one killer in women in 2015 (Miller, et al. 2016). To give a better picture, there is a chance of one in 14 women and 1 in 603 men developing breast cancer globally between birth and age 79 years, and 1 in 30 women dies because of that (Miller, et al. 2016).

Recurrence is one of the important considerations in breast cancer. Although in many cancers risk of recurrence is low if it does not occur in the first five years, in breast cancer, recurrences were reported after twenty years or more in some subtypes. Therefore, the breast cancer prognosis correlates to primary breast tumor specifications (Global Burden of Disease

Cancer, et al. 2017). It is clear that there are still more room for future research in breast cancer to improve patient outcomes.

1.3.1. Breast Cancer classification

Breast cancer is a heterogeneous disease that can be categorized in different ways such as histopathological type, grade of the tumor, stage of tumor, and the expression of proteins and genes. These classifications are useful in predicting prognosis as well as choosing effective treatment in different patients (Cho 2016).

Based on the microscopic characteristic of the tumor, breast cancer can be categorized histhopatologically in different types. The main three subtypes are invasive ductal carcinoma, ductal carcinoma in situ and invasive lobular carcinoma (Eheman, et al. 2009). Further important microscopic information is grading of the tumor, which is assessed by tumor cells' similarity to normal cells. This classification grades cancer from low-grade to high-grade (well differentiated and poorly differentiated respectively). High-grade tumors have the worst prognoses.

As a clinical tool, staging categorization is developed to describe the location and extent of the primary tumour as well as the magnitude of body involvement. TNM—the most widely used cancer-staging system—was recommended by American committee on Cancer (AJCC) and the International Union Cancer (UICC) (Amin, et al. 2017). This system is based on three indicators: T represents tumor values based on the primary site of breast tumor (ranging from TX to T4). N represents lymph node involvement depending on number, size and location of involved lymph nodes (ranging from Nx to N3) and M represents metastases to other organs than breast and lymph nodes (Ranging from MX to M1)(Amin, et al. 2017). Lower stages have better outcomes than higher stages.

The most important classification that provides useful information in predicting prognosis and responsiveness to treatment is tumor receptor status. This classification is based on the presence or absence of estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2) amplification. This information is provided by immunohistochemistry (IHC). Tumor receptor status is very important as it determines targeted therapy available for that type of cancer. Seventy percent of breast cancer patients are ER+ that means breast cancer expresses estrogen receptor. Although ER+ breast cancer has various subtypes, these tumors need estrogen to grow. Therefore this type responds to endocrine therapy either by reducing the effect of estrogen (tamoxifen) or reducing estrogen levels (aromatase inhibitors) (Paik, et al. 2004; van 't Veer, et al. 2002). Another example of the role of tumor receptor in treatment is the success in treating patients with trastuzumb in conjugation with chemotherapy in HER2 positive patients (Romond, et al. 2005; Slamon, et al. 1987).

The other type of classification was introduced through expression array analysis. Based on expression data, breast cancer tumors are classified in four distinct intrinsic molecular subtypes: “luminal A,” “luminal B,” “HER2-enriched,” and “basal-like” (Table 1.2) (Cancer Genome Atlas 2012; Hon, et al. 2016).

Table 1. 2. Surrogate definitions of intrinsic subtypes of breast cancer classification from the St. Gallen Consensus 2013.

Modified from Cho et al. Ultrasonography, 2016.

Intrinsic subtype	Clinicopathologic surrogate definition					Type of therapy
		ER	PR	HER2	Recurrence	
Luminal A	Luminal A-like	+	+	-	Low	Endocrine therapy is often used alone Cytotoxic therapy may be added
Luminal B	Luminal B-like (HER2-negative)	+	- or low	-	High	Endocrine therapy for all patients, cytotoxic therapy for most
	Luminal B-like (HER2-positive)	+	Any	Over-expressed or amplified	NA	Cytotoxics+anti-HER2+endocrine therapy
ErbB-2 overexpression	HER2-positive (non-luminal)	Absent	Absent	Over-expressed or amplified	NA	Cytotoxics+anti-HER2
Basal-like	Triple negative (ductal)	-	-	-	NA	Cytotoxic

ER, estrogen receptor; PR, progesterone receptor; HER2, human epidermal growth factor receptor 2; NA, not applicable.

Combining information from different classifications that was discussed above assigns patients to a high or low risk groups, determines the prognosis and plans specific treatment (Sotiriou and Pusztai 2009).

1.3.2. Triple Negative Breast Cancer

Triple-negative breast cancer (TNBC) is an aggressive breast cancer subtype, characterized by minimal or no expression (less than one per cent expression) of estrogen receptors (ER) and progesterone receptors (PR), as well as absence of overexpression of human epidermal growth factor 2 (HER2). TNBC makes up 15-20% of breast cancers and affects almost 200,000 patients each year (Trivers, et al. 2009; Yao, et al. 2017).

Compared to hormone-positive breast cancer, these tumors tend to occur more in women who are young and/or African-American (Dietze, et al. 2015). Additionally, TNBC patients usually are diagnosed at later stages than other breast cancer subtypes and are associated with poorer prognosis. Due to its aggressive nature, and high rate of visceral and distant relapse, mortality rate is higher in TNBC compared with other breast cancer subtypes. Overall, 5-year survival in TNBC patients is 64% in comparison to 81% in non-TNBC patients (Liedtke, et al. 2008).

1.3.2.1. Genomic profile of Triple Negative Breast Cancer

Although it was argued TNBC and basal like are the same subtype of breast cancer, not all of TNBC tumors express basal like subtype markers and not all basal like tumors are TNBC (Cho 2016). About 80% of TNBC cases are basal subtype. These TNBCs are very aggressive and have

high rates of metastasis to the visceral and central nervous system (Carey, et al. 2010). Among genes, there is high incidence of BRCA1 and in lesser extent BRCA2 mutation in TNBC patients (Stevens, et al. 2013).

Researchers have tried to classify TNBC into different molecular subtypes. Lehmann et al classified TNBCs with the aim to better helping molecular-based therapies and improving prognosis (Lehmann, et al. 2011). They categorized TNBC into six subtypes: basal-like 1 (BL1), basal-like 2 (BL2), immunomodulatory (IM), mesenchymal (M), mesenchymal stem-like (MSL) and luminal androgen receptor (LAR), based on gene expression profile. TNBC, as reported by Lehmann et al, is a significantly heterogeneous cancer and subtypes can respond to different chemotherapies. Basal like TNBCs may respond to cisplatin, as they are DNA-repair deficient and LAR subtype can respond to AR antagonists due to its androgen receptor signaling characterization.

More recently, work by Burstein et al on TNBCs reviewed DNA and RNA profile of samples and re-categorized TNBCs into 4 categories: LAR, mesenchymal (MES), basal-like immunosuppressed (BLIS) and basal-like immune activated (BLIA) (Burstein, et al. 2015). Among these, BLIA has the best and BLIS the worst prognosis.

1.3.2.2. Treatment Strategies for TNBC

As mentioned above, TNBCs do not present receptors, thus these tumors do not respond to hormonal therapies. As a result these tumors must be treated by chemotherapeutic drugs, usually anthracycline- and taxane-based drug regimens. TNBCs are more sensitive to chemotherapy than ER+ tumors because of higher rates of proliferation. Despite better initial response, the prognosis is poorer than other BC types even in early stages (Dent, et al. 2007). One of the main factors in

the prognosis is the resistance of TNBCs to chemotherapy treatment. Around 30-55% of patients respond completely to powerful chemotherapeutic drugs in pre-operative cases (Carey, et al. 2007; Loibl, et al. 2015). However, if the TNBC becomes resistant to such treatment, the prognosis is very poor (Cortazar, et al. 2014). Approximately 80% of TNBC patients have residual disease and consequently are at risk of relapse and metastasis (Dent, et al. 2007). When metastasis happens, tumors may at first show some treatment response, but all of them rapidly become resistant to a wide variety of chemotherapeutic drugs. Therefore, it will become difficult to treat them due to limited therapeutic options, which leads to the high mortality observed in these cases. Tremendous research is in progress on recently discovered therapeutic targets for different pathways and many pathways are under investigation in clinical trials (Table 1.3).

Table 1. 3. Ongoing studies in the clinical trial stage to investigate therapeutic targets for TNBC.

Modified from Shao et al., Oncotarget. 2017

Therapeutic targets	Drug	Mechanism of action	Patient population
EGFR	Afatinib	Pan-ErbB dimers inhibitor	TNBC
	Gefitinib	EGFR TKI	TNBC with EGFR positive
	Cetuximab	EGFR-mAb	Breast Cancer contains TNBC
	MM 151	Oligoclonal anti- EGFR antibody	Advanced solid tumor contains TNBC
	Lapatinib	EGFR/HER2 TKI	Metastatic TNBC
VEGF/VEGFR	Bevacizumab	VEGF-A inhibitor	TNBC
	Cediranib	VEGFR inhibitor	Solid tumors contain TNBC
AR	GTx-024	Selective androgen receptor modulator	TNBC with AR positive
	Bicalutamide	AR inhibitor	TNBC with AR positive
PI3K/AKT/ mTOR	GSK2141795	AKT kinase inhibitor	Cancer contains TNBC
	BKM120	PI3K inhibitor	TNBC
	AZD5363	AKT kinase inhibitor	Cancer contains TNBC
PARP	Iniparib	PARP inhibitor	TNBC
	Olaparib	PARP inhibitor	Cancer contains TNBC
	Talazoparib	PARP inhibitor	Breast cancer patients with BRCA mutation
Notch pathway	PF-03084014	Gamma-Secretase inhibitor	TNBC
	RO4929097	Gamma-Secretase inhibitor	Breast cancer contains TNBC
Hedge-hog pathway	LDE225	Smo antagonist	TNBC
PD-1	JS001	anti-PD-1- mAbs	TNBC
	Pembrolizumab	anti-PD-1- mAbs	TNBC

1.3.3. Mechanisms of drug resistance

Resistance to chemotherapy is the underlying cause of most cancer fatalities. Cancer cells have the ability to escape from chemotherapy and become resistant to traditional anti-cancer drugs. There are different known mechanisms involved in resistance (Figure 1.1).

One of the ways cancer cells escape from chemotherapy is drug inactivation. Tumor cells decrease metabolic activation of the drugs in many ways such as alteration in CYP, and elevation of GST expression (Manolitsas, et al. 1997; Shen, et al. 2007). As a result, a drug cannot achieve its clinical efficacy or will be detoxified.

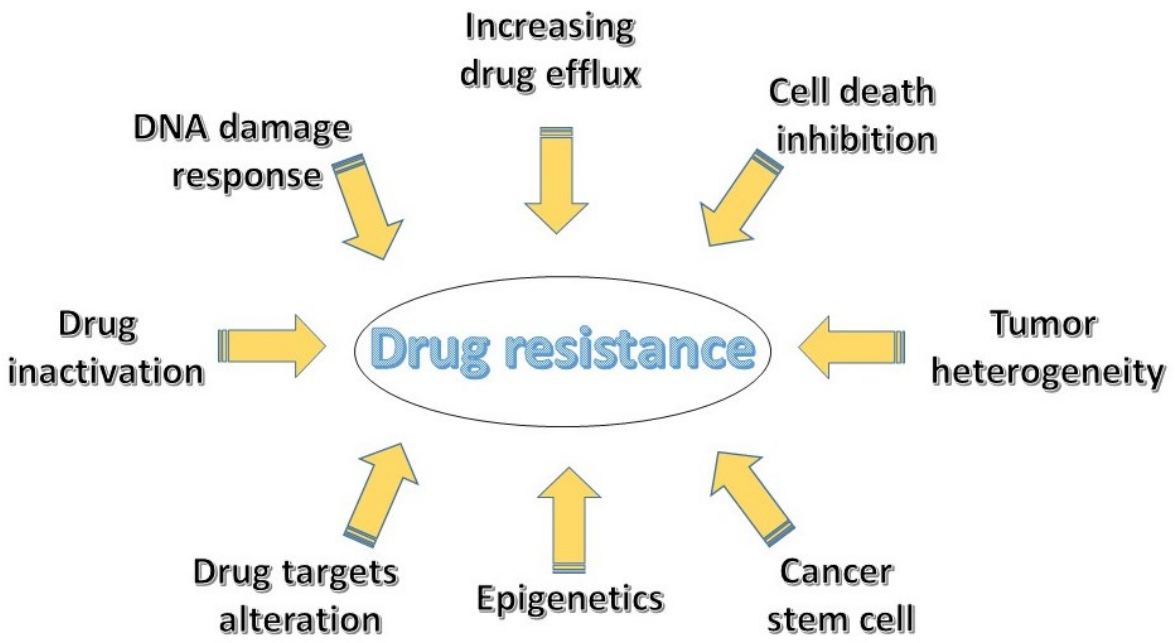


Figure 1. 1. Different mechanisms involved in drug resistance

A drug's efficacy has a close relationship with its molecular targets. Alteration of these drug targets such as modifying enzyme expression level or altering signal transduction process can lead to resistance. Examples of this type of resistance in breast cancer are resistance to trastuzumab in HER2 positive breast cancer and tamoxifen in ER+ type (Dieras, et al. 2007; Shou, et al. 2004). Mutation in the target can also develop resistance. As an example, mutation in topoisomerase II gene leads to resistance to drugs that target this enzyme (Holohan, et al. 2013; Stavrovskaya 2000).

Another main anti-cancer drug resistance mechanism is reducing drug accumulation by increasing drug efflux. ATP-binding cassette (ABC) transporter family proteins' over expression is involved in increasing efflux. Three main known transporters involved in drug resistance are multidrug resistance protein 1 (MDR1), multidrug resistance-associated protein 1 (MRP1), and breast cancer resistance protein (BCRP) (Gottesman, et al. 2002; Haber, et al. 2006; Yanase, et al. 2004).

In addition to the above mechanisms, DNA damage response (DDR) mechanisms can fix the damage induced by chemotherapy. As examples, DNA repair via O6-methylguanine DNA methyltransferase (MGMT) confer resistance to alkylating chemotherapy agents, and resistance to cisplatin occurs due to nucleotide excision repair and homologous recombination (Bonanno, et al. 2014; Esteller, et al. 2000).

In addition, cancer cells can inhibit cell death by antiapoptotic activity such as downregulating proapoptotic molecules Bax /Bad or highly expressing anti-apoptosis proteins such as BCL-2 family proteins and Akt.

Recent studies, which focused on involvement of cancer stem cells in relapse and metastasis, showed drug resistance could emerge during the signalling processes of differentiation (Byler, et al. 2014). For example increased expression of integrin αvβ1 serves as a survival signal

for cancer cells against drugs by increasing TGFβ expression (Bates and Mercurio 2005). In addition, there is a hypothesis that due to aberrant DNA repair mechanisms, cancer tumours are not homogenes. This hypothesis surveyed by Witz et al who revealed that breast cancer tumors may be either monogenomic or multiple genomic (Witz 2008). This heterogeneity can be another reason in drug resistance.

Importantly, recent studies suggest that epigenetic alterations can have further impact in the development of drug resistance (Worm, et al. 2001). Although the role of epigenetics in cancer development has been shown before, new findings suggest more studies should be done in this area.

1.3.3.1. Genetic basis of drug resistance in TNBC

The genomic characterization of primary TNBCs has revealed few clues about the factors underlying drug resistance. Indeed, primary TNBCs are characterized by frequent p53 mutations (>70%) and a high degree of genomic instability, but with few other recurrent actionable mutations (Cancer Genome Atlas 2012). Unfortunately, the TP53 gene has not been associated with therapeutic response in large retrospective analyses and the best biomarkers of response to chemotherapy are BRCA1/2 germline mutations (about 15%), which may indicate responsiveness to platinum-based chemotherapy and to PARP inhibitors (de Bono, et al. 2017; Fernandez-Cuesta, et al. 2012).

Resistance to taxanes especially paclitaxel was investigated in several studies. Drug target alterations such as mutations in beta-tubulin were reported as a cause of resistance to this medication in ovarian cancers (Giannakakou, et al. 1997). Increasing the efflux of the paclitaxel

by MDR1 and doxorubicin by both ABCG2 and ABCB1 were accounted as drug resistance factors to these drugs (Abdullah and Chow 2013; Litman, et al. 2000).

In reality, pre-clinical studies showed that the causes of drug resistance in TNBC are potentially many. In a specific example, overexpression of ABCC3 drug transporter gene was proposed for in vitro resistance to paclitaxel and MMAE (O'Brien, et al. 2008). Moreover Lu et al. suggested loss of E-cadherin expression to be the reason of invasion/metastasis in 7/Adr cells resistant to drugs in comparison to their parental control (Lu, et al. 2012). Moitra et al. conferred ABCC1 and ABCC6 transporter genes are upregulated in MCF7VP cells significantly which leads to decreased cellular uptake (Balaji, et al. 2016; Moitra, et al. 2012). Altered signaling pathways resulting in dysregulated apoptosis, DNA repair and autophagy, paracrine effects, and changes in microRNA expression are other proposed reasons for drug resistance in TNBCs (Amornsupak, et al. 2014; Balaji, et al. 2016; Chen, et al. 2014; Guay, et al. 2008; Moitra, et al. 2012; Nair, et al. 2016; Tan, et al. 2015; Yao, et al. 2015).

Nevertheless, none of these molecular factors has been validated in clinical studies in drug-resistant triple negative breast tumors. Indeed, because of the great difficulty of obtaining clinical samples from advanced tumors in patients, there is scant information from actual drug resistant tumors. Balko et al provided the first look into the molecular and genomic features of drug-resistant TNBCs by using targeted next generation sequencing. They investigated residual tumors remaining or persisting after neoadjuvant chemotherapy treatment. By looking at a limited panel of genes, they reported several potentially actionable mutations enriched for in the residual tumors, including alterations involving the JAK2 and MLL3 genes (Balko, et al. 2014).

Although many researchers have tried to investigate drug resistance in Triple Negative Breast Cancer, still, the mechanism underlying resistance in this subtype, which accounts for much

higher proportion of all breast cancer mortality, is unknown. We hypothesize that the treatment of TNBC tumors with chemotherapy would lead to the enrichment and/or selection of genomic alterations that are associated with resistance to the chemotherapy. Therefore, in chapter four, I will explore this hypothesis.

1.4. Whole-Exome Sequencing analysis and its challenges

Whole exome sequencing was the primary investigating tool used for analyzing both familial pancreatic cancer and triple negative breast cancer samples. A typical WES analysis involves several steps such as library preparation, base calling, quality control, mapping, variant calling, annotation, filtration and prioritization (Figure 1.2). It is important to know how these steps function, as any limitation in any phase impacts downstream analyses.

1.4.1. Exome sequencing data analysis

1.4.1.1. Library preparation

The first step in WES is library preparation. Although there is a wide range of protocols, they all have common steps:

1- DNA samples are randomly fragmented to construct the library

2- Platform specific adaptors are attached to both ends of the fragments

3- Biotinylated DNA or RNA (baits) probes are used to selectively hybridize exonic sequences in solution-based WES (in array-base WES, probes are attached to high-density microarray).

4- The rest of DNA that was not targeted is washed and exonic regions are amplified using PCR.

1.4.1.2. Base Calling

The base calling step, the identification of each nucleotide in sequence, is an important step in accurate and powerful variants detection. Sequencing platforms vendors (Illumina, Roche 454, ABI SOLiD), provide image analyses and base calling software. Therefore, base calling is remained largely a closed subject. This step suffers from systematic errors due to either nucleotide misincorporation in sequencing process or error in reading during image processing. There are more base-calling errors at the end of the reads, in the high-GC content, and in inverted repeat regions in Illumina platforms. (Ledergerber and Dessimoz 2011; Nakamura, et al. 2011) Contamination of foreign DNA with the sample also causes some artifact (Flickinger, et al. 2015). In the base-calling field, tremendous efforts have been done to develop high performing base calling algorithms. As a result, it is currently possible to reach 99.5% accuracy in base-calling (Massingham and Goldman 2012).

1.4.1.3. Quality Control

A critical step in NGS data analysis is quality control (QC). Since base-calling errors can affect the biological interpretation of the results, it is necessary to assess and eventually correct the raw data from sequencers before proceeding to further analysis steps. Most of the tasks carried out at the quality control level involve adaptor clipping, trimming and filtering. Tools to perform visualization and statistical summary on the raw data are also useful in this step.

Quality control (QC) has two steps:

1) Step one is performed before alignment. The intrinsic quality of the raw reads is obtained from metrics generated by the sequencing platform. These data include error percent, per-

base quality scores and duplication percent. Then, reads with insufficient quality would be discarded or trimmed. Next, to prevent PCR-generated errors, duplicate reads would be set aside.

2) Step two is performed after mapping. This step uses calculation metrics from the alignment process. Generally, FastQC tool(http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and Picard toolkit (http://picard.sourceforge.net/command-line-overview.shtml) are used for QC assessment. Their generated metrics such as proportion of reads aligned to reference genome, average depth of coverage, over-represented sequences, error rate, and duplication rate is used for QC. If a sequencing run does not pass QC with acceptable result, it should be reported in order to be repeated.

1.4.1.4. Sequence alignment

After raw WES data has passed QC stages, the next step is reconstructing the exome. In this process, millions of short reads generated from sequencing are aligned to a human reference genome. Then mapping quality score will be calculated to measure how confident the read is placed to the correct spot. Currently, many alignment programs have been developed to efficiently process these short reads. The most commonly used program is Burrows-Wheeler Aligner (BWA)(Li, et al. 2009) . However, this step still suffers from some limitations due to the fact that the current human reference genome is still incomplete. Sequencing platform vendors also makes assumption on what regions of the genome are exomic.

At the end of the WES alignment, each base will be represented by approximately 100 independent sequencing reads. This amount of coverage is needed to filter random sequencing errors that may occur in individual reads.

1.4.1.5. Variant identification

After appropriate QC and sequence alignment are complete, the next step is to determine the variation between the sample and the reference genome. These variations include single nucleotide variants (SNVs), multi-nucleotide variants (MNVs), small insertions or deletions (INDELs; generally less than 50bp, larger structural variants (SVs) and copy number variants (CNVs).

With current tools and within the context of clinical applications, single nucleotide variations (SNVs) and small insertion/deletion (sINDELs) are currently easy to detect, but other types of variations (multi-nucleotide variants (MNVs) and combination of SNVs and INDELs at a single genomic locus) are at times tricky to accurately identify. Many studies are in progress to identify other complex class of variants such as structural variants (SVs), large INDELs, inversions and translocations. Yet, the success is not sufficient to be considered in clinical setting.

The Naïve variant detection approach believes that at any given locus, the number of occurrences of each distinct nucleotide among the reads aligned at a position can be counted, and that the true genotype would be obvious. This assumption will not work as other factors are involved in variant detection. Importantly read quality cutoffs, coverage, and errors should be considered when identifying true variants for a specific experiment.

To determine whether a detected variation in an aligned sequence is a true genetic variation, various variant calling (heuristic and probabilistic) methods have been developed. During NGS technologies evolvement, the probabilistic (based on Bayesian statistical models) methods, such as SAMtools (Li, et al. 2009) and Genome Analyzer Tool Kit (GATK)(McKenna, et al. 2010) have proved to be much more robust than heuristic approaches. It might be largely due to their statistical assumptions on various sequencing errors. It should be noted that the accuracy of these tools differs and depends on the average depth of sequencing and variant type of interest.

While SAMtools and GATK HaplotypeCaller (HC) results are similar in calling SNVs with a True Positive rate of 97.1% (O'Rawe, et al. 2013), their function differs in calling INDELs. This discordance (26.8%) brought the question: which one can be more effective? some studies showed that GATK HC could be robust in calling INDELs. It is now however proved that none of the above tools is unequivocally best and researchers should be cautious when calling INDEL from WES data in a clinical setting (O'Rawe, et al. 2013).

The sensitivity of variant identification (true positive rate) depends more on coverage depth than the software. For example, in a typical WES with 100-150X mean exome coverage, 93% of all known coding exons are covered at a depth of 20X or more. This would be almost equal to a sensitivity of around 93%, exome-wide. However, the specificity (the probability that a given identified variant is real) is dependent on the software that has been applied. Recent studies show slightly better result using GATK (O'Rawe, et al. 2013). Although many improvements have been achieved and there are more to come, there are still false positives. These errors can be detected partially through visual inspection of the data by experienced analysts using visualization tools such as Integrated Genomics Viewer (IGV)(Robinson, et al. 2011).

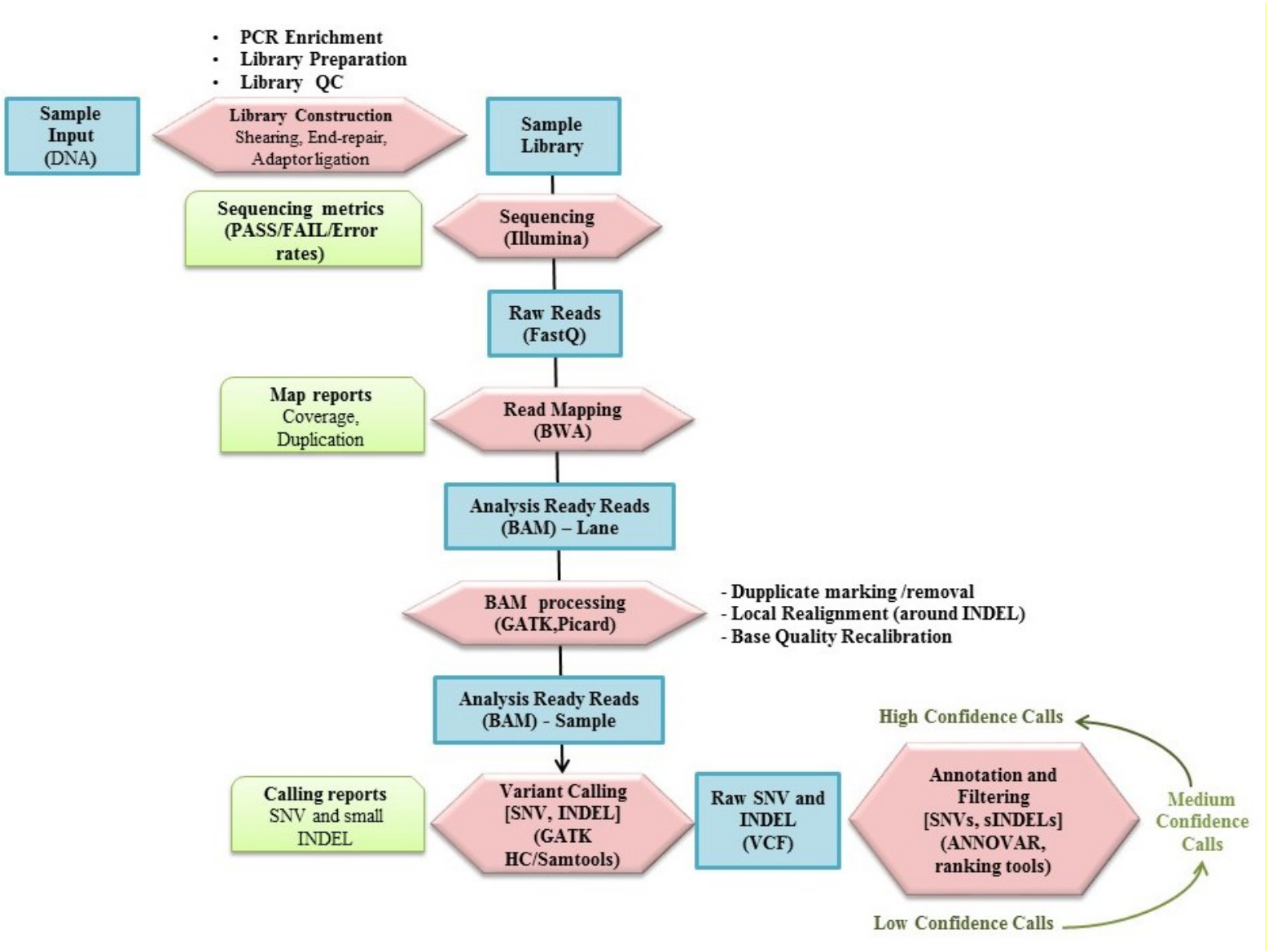


Figure 1. 2 . WES Analysis workflow in clinical setting for SNVs and small INDELs.

1.4.1.6. Variant annotation

One of the major steps in WES analysis is interpreting the data. Variant annotation is the main step for interpreting the data. At first, the variant is classified by its functional category (synonymous, missense, non-sense, splicing, frameshift, etc.). Any change that directly affects the amino acid sequence is considered as disease-related.

The other important information is obtained from the presence and frequency of a variant in public databases. Many laboratories obtain AF for the variants using Minor Allele Frequency (MAF) data from major population databases such as dbSNP (Sherry, et al. 2001), The 1000 Genomes Project Consortium (1000G), NHLBI Exome Sequencing Project (ESP) (Fu, et al. 2013), ExAC database (Karczewski, et al. 2017) and recently devolved database, gnomAD (http://gnomad.broadinstitute.org) . As allele frequency (AF) is one of the important factors for rare diseases, any variant present at a frequency of more than 5% in normal population in the aforementioned databases is unlikely to be of relevance to a rare genetic disease.

Next, information related to the association between human phenotype and causative genes will be added. These data can be obtained from databases such as the Online Mendelian Inheritance in Man (OMIM) (Amberger, et al. 2011), the Human Gene Mutation Database (HGMD) (Stenson, et al. 2014), and ClinVar (Landrum, et al. 2018).

Finally, information from in silico predictors such as measures of evolutionary sequence conservation (namely scores from PhastCons(Margulies, et al. 2003), GERP++ (Davydov, et al. 2010), Phylop (Pollard, et al. 2010) and SiPhy (Garber, et al. 2009)) or predictions about the variant's effect on the protein (for example CADD (Kircher, et al. 2014), SIFT (Ng and Henikoff 2003), and PolyPhen2 (Adzhubei, et al. 2010)) are commonly used to predict the deleteriousness of the variant. Currently, ANNOVAR (Yang and Wang 2015) and VEP

(http://bioinformatics.knowledgeblog.org) are the most applicable bioinformatics tools used for annotation.

1.4.1.7. Variant filtering and evaluation

The annotated variants will go through a filtration process. The final goal of this step is to reduce candidates to the acceptable number and highlight the most significant ones to validate.

There is a pre-filtering process independent to the disease of interest and its mode of inheritance. In this process, only variants in coding regions, which affect the protein sequence, will be retained for further processing. Then, low confidence variants such as common variants on known variant databases such as dbSNP and population frequency data sources such as ESP, 1000G and ExAC will be further filtered.

The next step in filtering depends highly on the type of the disease. For example, somatic mutations are more important in cancer than germline mutations are in rare diseases. The inheritance model (dominant, recessive, homozygous or compound heterozygous, or X-linked) and the number of samples involved are other factors that play major roles in this step. Family information, segregation in families, and the relationship of the gene to the phenotype will also be applied for further filtering. For example, in the case of the rare, fully penetrant Mendelian disease, combining information from different affected individuals and their family members helps in finding the causal mutations.

In the end, functional impact predictions and conservation scores such as SIFT, PolyPhen, GERP and CADD will be used in order to help ranking the remaining variants. However, the cutoff for AF and also prediction scores differ from one disease to another and should be set differently. For example, inherited SNVs and INDELs predisposing people to complex diseases are expected

to occur at a greater frequency in the population than rare diseases and fully penetrant autosomal dominant disease. Therefore, in the first scenario, a less stringent threshold might be employed in comparison to the second scenario.

As described above, even though next generation sequencing data are generated by common sequencing platforms, analyzing and interpreting the data often requires specialized methods.

1.4.2. Challenges in variant prioritization

High-throughput DNA sequencing has revolutionized the identification of variants in the human genome. Advances in these technologies reduced the price and made them affordable. Despite these technologies being very useful in finding new genes responsible for the diseases, the genetic causes of more than 60% of suspected mendelian phenotypes cannot be immediately determined with the current NGS analysis method (Yang, et al. 2013).

A serious challenge in using NGS technologies is interpreting the effect of discovered variants. Considering enormous amount of data is generated by next generation sequencing technologies these days, finding the real causative variant would be like finding a needle in the haystack. Distinguishing pathogenic amino acid changes from background polymorphisms is important for efficient use of these technologies in genetic discovery, personalized medicine and clinic. Accurate prediction of deleteriousness of genomic changes, especially nonsynonymous single nucleotide variants (nsSNVs) is a need in the NGS era.

To understand the scope of the challenge, each person may hold 24,000–40,000 single nucleotide variants, as compared with a reference genome, with many never having been seen

before. Although using common filtering criteria such as population allele frequency and mode of inheritance may narrow down the list, hundreds of variants may remain for further investigation (O'Fallon, et al. 2013). In case of finding a new gene responsible for a specific disease, having multiple affected individuals and focusing on genes mutated in all of them may be helpful. However, in case of diagnostics especially in rare diseases, multiple cases from the same family may not be available and mode of inheritance is not clear. As a result, common filtering yields a large variant list. Manual examination of these numbers of variants requires a lot of time. Moreover, experimental validation of the pathogenicity of large numbers of variants is not feasible as it is expensive and time consuming. Consequently, many algorithms have been developed to predict the potential impact of variants on protein structure. Those methods use various computational algorithms and different properties of the variant, such as relationship to local protein structure, evolutionary conservation and/or physiochemical and biochemical properties of amino acids to predict degree of pathogenicity.

1.4.2.1. Pathogenicity prediction methods

In order to assess the pathogenicity of the variants, different prediction methods have been developed. Traditional prediction tools are mainly based on sequence homology. They use multiple sequence alignments they each build independently to find the conservation information (Lopes, et al. 2012).Their estimate of the deleteriousness of a single nucleotide variant (SNV) is based on the assumption that highly conserved areas among living organisms are very important and these positions are those that have not been removed by natural selection. Therefore, individuals with variants in this area did not fit in the population and likely were removed from it. Many prediction methods such as likelihood ratio test (LRT), Sorting Intolerant from Tolerant

(SIFT), Genomic Evolutionary Rate Profiling (GERP), Protein Variation Effect Analyzer (PROVEAN), Siphy, PhastCons, and Mutation Assessor are from this category (Cooper, et al. 2005a; Garber, et al. 2009; Kumar, et al. 2009; Reva, et al. 2011; Siepel, et al. 2005).

LRT applies a likelihood ratio and uses the genomics data set of 32 vertebrate species. It can identify subsets of deleterious mutations that disrupt highly conserved amino acids within protein-coding sequences (Chun and Fay 2009). On the other hand SIFT applies sequence homology and performs alignment on the diverse set of homologs selected (Cline and Karchin 2011; Kumar, et al. 2009). PROVEAN algorithm computes a semi-global pairwise sequence alignment score between the given sequence and each of the set of its related protein sequences. GERP, PhastCons, phyloP and SiPhy use sequence information to detect functional elements conserved over relatively long evolutionary time periods (Cooper, et al. 2005a; Miosge, et al. 2015). Phastcons and phlyloP use phylogenetic methods. PhastCons uses hidden Markov model to identify evolutionarily conserved elements based on a phylogenetic tree of 46 different species (Siepel, et al. 2005). It not only considers conservation of individual columns, but also the effects of the neighbors. On the other hand, phyloP just measures the conservation in an individual nucleotide and predicts a fast evolving or conserved nucleotide (Cooper, et al. 2005a). GERP identifies constrained elements. It uses a likelihood tree-based method to calculate difference between estimated and expected evolution rate to find the slowly evolving regions that are more likely enriched for functional elements (Cooper, et al. 2005a). Mutation Assessor differs from other conservational methods in its algorithm, which uses evolutionary conservation in protein subfamilies (Reva, et al. 2011).

Although conservational algorithms have had success in predicting pathogenicity, not all the deleterious variants are in the constraint region or conserved among multiple sequence

alignment. To overcome this problem other methods were developed to consider protein structure and physiochemical properties besides sequence homology. These methods consider polarity, side change volume, secondary structures (alpha helix, beta sheet changes) and 3D structures to determine deleteriousness of amino acid change (Figure 1.3). As an example, Polymorphism Phenotyping v2 (PolyPhen-2) uses protein structure/function and evolutionary conservation information, and applies naïve bayes with entropy-based discretization approach to predict the impact of amino acid change on the stability and function of affected proteins (Adzhubei, et al. 2010). MutationTaster applies naïve bayes method as well but it uses sequence information and annotation data to predict the consequence of amino acid substitution (Schwarz, et al. 2010).

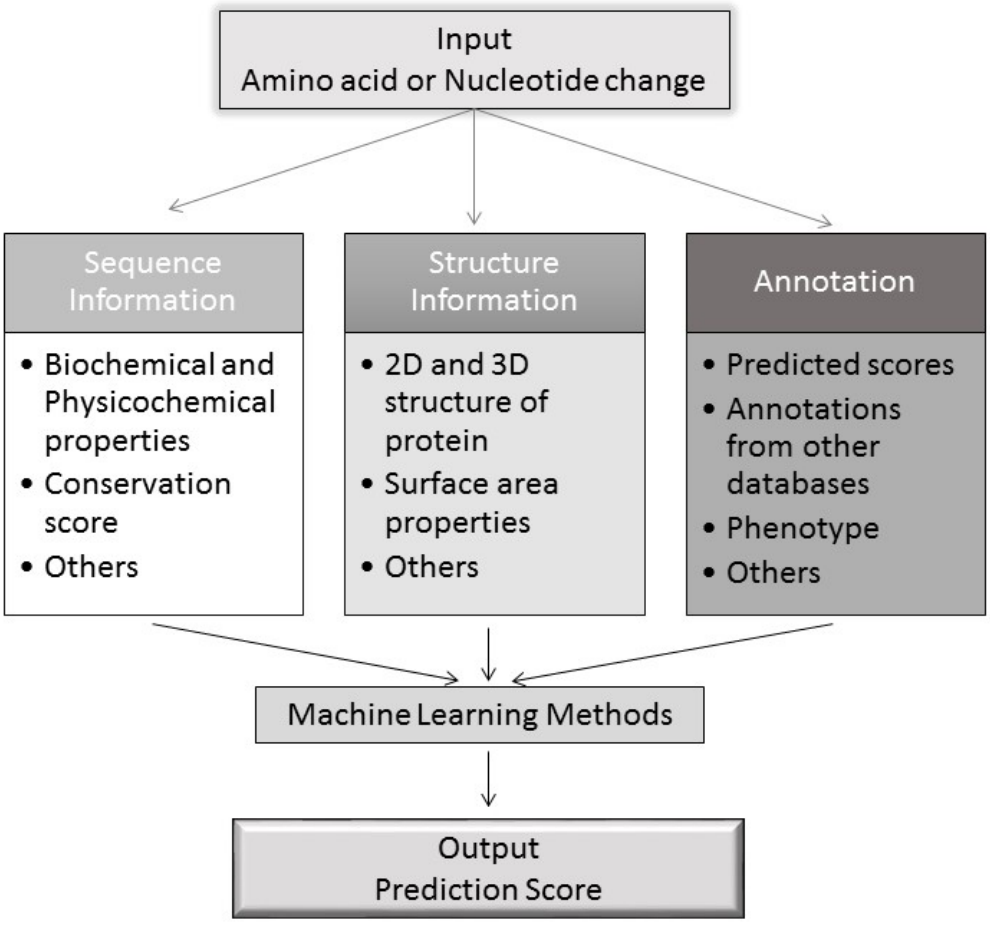


Figure 1. 3. Basics of prediction methods algorithms to estimate deleteriousness of nonsynonymous single-nucleotide polymorphism for human diseases.

Although such methods undoubtedly provide positive predictive power, they have their own limitations due to limited available experimental data. Another problem is the prevalence of missing data. That means even the most used prediction methods such as PolyPhen-2 and SIFT suffer from high rate of missing data (no prediction) when they were unable to find sufficient related protein sequences at the position of interest in their respective multiple sequence alignment pipelines (Lopes, et al. 2012). Importantly, when applied on real NGS data, pathogenicity scores are often in disagreement with each other (Ioannidis, et al. 2016; Li, et al. 2014). This problem makes it difficult to form a list of probable causative candidate variants for validation as it is difficult to choose which prediction method is the most reliable and generalizes better.

It is believed that individual methods have complementary strengths, depending on their specific features and algorithms. Hence, recently, new ensemble predictors have combined individual scores in order to achieve higher classification accuracy. These methods were trained on different training databases and used different machine learning algorithms (Table 1.4).

The Combined Annotation Scoring Tool (CAROL) and CONsensus DELeteriousness score of missense mutations (Condel) were the first ensemble prediction tools. CAROL used weighted Z method to combine the predictions of PolyPhen-2 and SIFT and was trained on data from dbSNP, HGMD-PUBLIC and 1000 genome (Lopes, et al. 2012). Condel additionally used the score from Mutation Assessor and PANTHER and was trained on humvar (Gonzalez-Perez and Lopez-Bigas 2011). One of the most successful methods in this category is Combined Annotation-Dependent Depletion (CADD). It integrated a large number of sequence based, structure based and genomic attributes data and applied support vector machine algorithm on the training data consisted of simulated variants (Kircher, et al. 2014).

Most recent ensemble methods such as M-CAP (Mendelian Clinically Applicable Pathogenicity) and REVEL (rare exome variant ensemble learner) tried to apply new algorithms and train on human variants to increase the accuracy. These tools applied supervised machine learning algorithms and trained on disease-causing mutations from Human Gene Mutation Database (HGMD) in combination with a large number of neutral variants from ExAC and EVS (Ioannidis, et al. 2016; Jagadeesh, et al. 2016) (Table 1.4).

In contrast to the above mentioned ensemble tools, Eigen and GenoCanyon were developed using unsupervised machine learning algorithm to overcome the pitfall in using labeled data as the training dataset. The authors believed the reliance on the labels of the variants in the training set affects the accuracy and imprecise training data affect the prediction score significantly. (Ionita-Laza, et al. 2016; Lu, et al. 2015).

Table 1. 4. Description of the most used prediction methods as well as recently developed ones.

Method	Method description	Features
SIFT	Statistical method using PSSM with Dirichlet priors	Sequence based (Evolutionary conservation)
PolyPhen-2	Naïve Bayes approach coupled with entropy-based discretization trained on Uniprot	Sequence based, structure based (Protein structure/function and evolutionary conservation)
LRT	Log ratio test	Sequence based
GERP	A "Rejected Substitutions" score computation to infer the constrained region	Sequence based (Nucleotide conservation prediction)
CADD	Support Vector Machines trained on simulated and observed substitutions	Conservation, protein function, Encode, DNA structure
PhastCons		Sequence based (Nucleotide conservation prediction)
Mutation assessor	evolutionary conservation patterns in protein family multiple sequence alignments	
fitCons	Probabilistic model and trained on in-house data	Conservation, Encode
DANN	Artificial Neural Network, trained on simulated and observed substitutions	Conservation, Encode, PolyPhen
SiPhy		Sequence based (Nucleotide conservation prediction)
PROVEAN		Multiple sequent alignment
phyloP	Measures p-values for conservation or acceleration based on an alignment and a model of neutral evolution	
GAVIN	pathogenic impact distribution was calculated as the relative proportion of the generalized effect impact categories	AF in Exac, SnpEff ,CADD
Condel	computes a weighted approach of missense mutations from the complementary cumulative distributions of scores of	Pfam, MAPP, SIFT, PolyPhen, Mutation Assesor

Method	Method description	Features
	deleterious and neutral mutations , trained on Uniprot	
CAROL	WeightedZ method and trained on HGMD, dbSNP, 1000G	SIFT, PolyPhen
REVEL	Random forest trained on HGMD, ESP	MutPred, FATHMM, VEST, PolyPhen, SIFT, PROVEAN, MutationAssessor, MutationTaster, LRT, GERP, SiPhy, phyloP, and phastCons. REVEL
FATHMM	Hidden Markov model	sequence-based method that associates evolutionary conservation in homologous sequences with disease-specific weights
MetaSVM	Support Vector Machines trained on Uniprot	SIFT score, PolyPhen-2 HDIV, PolyPhen-2, LRT, Mutation Assessor, FATHMM, GERP++, PhyloP, SiPhy and MMAF
MetaLR	Likelihood Ratio , trained on Uniprot	SIFT score, PolyPhen-2 HDIV, PolyPhen-2, LRT, Mutation Assessor, FATHMM, GERP++, PhyloP, SiPhy and MMAF
VEST	Random Forests, trained on HGMD, ESP	86 predictors from SNVBox
MutationTaster	Naïve bayes model trained on OMIM, HGMD and dbSNP	Sequence based, annotation (Protein structure/function and evolutionary conservation)
M-CAP	Gradient boosting trees	SIFT, PolyPhen-2, FATHMM, CADD, MutationTaster, MutationAssessor, LRT, MetaLR, MetaSVM, RVIS PhyloP, PhastCons, PAM250, BLOSUM62, SIPHY, GERP and 298 features derived from multiple-sequence alignment
Eigen	Unsupervised spectral approach trained on dbSNP and 1000G	Conservation, protein function, Encode, AF
GenoCanyon	Unsupervised statistical learning	22 features (2 conservation, 8 histon modification and 10 TFBS peaks) based on ENCODE

1.4.2.1.1. Data sources used as training data

As discussed, many prediction tools were trained or tested on available mutation databases. The most commonly used databases are The Universal Protein Resource (UniProt), The Single Nucleotide Polymorphism Database (dbsnp), Human Gene Mutation Database (HGMD), and ClinVar.

Uniprot is the database created by the European Bioinformatics Institute (EBI) in collaboration with the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR). Biological information from the literature has been gathered and computational analyses have been performed to annotate protein sequences in this database. Part of this database was used as a training data for polyphen, Condel, MetaSVM and MetaLR (Apweiler, et al. 2004)

.

The Single Nucleotide Polymorphism Database (dbsnp) is a freely available database that archives genetic variation within and across different species and contains variants for 55 organisms. The dbSNP contains both neutral polymorphisms and polymorphisms corresponding to known phenotypes (Sherry, et al. 2001). Mutation taster and Eigen used this database for training their methods.

HGMD is a collection of germ-line mutations in nuclear genes. This database is built based on the variants published in peer-reviewed publications. The developers' aim is to collect variants underlie human inherited disease or variant that is closely associated with it. This database has a professional version that comes with a fee and a free public version, which is permanently outdate for a minimum of 3.5 years. Disease-causing mutations (DM) label in this database is based on the claims of corresponded publication's authors. This database does not contain any somatic mutations (Stenson, et al. 2017). HGMD policy is to enter a variant into the database even if its

pathological relevance is questionable. CAROL, VEST, MutationTaster, M-CAP, REVEL were trained on DM variants from this database.

ClinVar is a freely available database that combines different interpretation of the same variants and it allows users to see any concordance or conflict between different submissions with supporting evidence. Moreover, it applies the standardized method recommended by the American College of Medical Genetics and Genomics (ACMG) for clinical interpretation of variants. Since its release in 2013, ClinVar has grown rapidly, and became the database best representing current understanding of the relationship between genotypes and medically important phenotypes. It includes both germline and somatic variants (Landrum, et al. 2016). This database is almost new; hence, it was used as testing data in the past rather than training data.

1.4.2.1.2. Drawbacks in pathogenicity prediction methods

Several traditional computational methods are developed to prioritize the variants, but they suffer from high false positive or false negatives. To overcome this problem, ensemble prediction tools applied machine-learning algorithms and were trained on known pathogenic and neutral nsSNVs mostly from HGMD or UNIPROT databases. While those databases provide important information about variants associated with diseases, they have known limitations. Dorscher and his group re-evaluated 239 unique variants classified as disease causing in HGMD. They founding showed only 16 unique autosomal dominant variants and 1 autosomal recessive variant pair were consistent with a pathogenic or likely pathogenic category (Dorschner, et al. 2013). In another study, Bell et al. re-examined recessive disease causing genes' labels and found 27% of annotations were incorrect or are common polymorphisms (Bell, et al. 2011). These incorrect annotations resulted as different submitters applied different criteria. Many genetic laboratories

had developed their own criteria and followed in-home criteria to predict any variant clinical significance. Since they used varying and inconsistent levels of evidence and methodologies, there were many conflicts in the interpretations.

To overcome this discrepancy, American College of Medical Genetics and Genomics (ACMG) in collaboration with Association for Molecular Pathology (AMP) recommended a guideline for interpreting the variants in a clinical context (Richards, et al. 2015). ClinVar database recommended users to apply this guideline in order to consistent and better variant interpretation. Today this database becomes a valuable source that archives relationship of variants with their clinical phenotype.

Even though several programs are designed to predict variants' impact, further improvement is needed, especially concerning prediction of clinical relevance. Having access to higher confidence variants make developing these tools more feasible. In Chapter 5 of this thesis, I will explore this problem.

1.5. Rationale and Objectives of Study

Recent advances in next generation sequencing have been transformative in medicine. Next generation high-throughput sequencing technologies have been widely used to identify germ-line mutations underlying Mendelian disorders and somatic mutations in various cancers. Understanding the underlying genetic factors in diseases is important in diagnosis and treatment of patients and identification of at-risk family members.

Family history plays an important role in risk of development for all cancers. Pancreatic ductal adenocarcinoma accounts for more than 90% of pancreatic cancers (PC). Although 10% of this type of cancer occurs in families, the genetic basis underlying familial pancreatic cancer in 85-90% of patients is still unknown. Considering DNA repair genes are widely implicated in gastrointestinal malignancies, and most known pancreatic genes are involved in DNA repair mechanisms, I hypothesize that additional DNA repair genes are involved in hereditary PC. Since application of WES in inherited forms of cancers has proved to be successful in revealing causal genetic events in a cost-effective manner, the first objective of this thesis is to identify new pancreatic cancer susceptibility genes, using WES as the primary investigative tool. I will investigate this hypothesis by applying WES to 109 familial pancreatic cases, performing a filter-based candidate gene approach focused on DNA repair genes.

For the second part of the thesis, another cancer type that typically has poor prognoses will be studied. Triple-negative breast cancer (TNBC) is an aggressive breast cancer subtype and resistance to chemotherapy is the underlying cause of death in these patients. Although many researchers have tried to investigate drug resistance in Triple Negative Breast Cancer, the mechanism underlying resistance in this subtype, which accounts for the large proportion of all

breast cancer mortality, is still unknown. We hypothesize that the treatment of TNBC tumors with chemotherapy would lead to the enrichment and/or selection of genomic alterations that are associated with resistance to chemotherapy. Therefore, as the second objective of this thesis, we will explore drug resistance factors in TNBC tumors using WES, in combination with RNAseq and aCGH data.

Although WES is the primary investigating tool for the first two objectives, a serious challenge in using this technology is interpreting the effect of discovered variants. Several traditional computational methods are developed to prioritize the variants, but they suffer from high false positive or false negative rates. Therefore, further improvement is needed, especially concerning prediction of clinical relevance. I hypothesize that by having access to higher confidence variants and better algorithms, we can develop tools that are more efficient. Given this, as my third main objective, I will propose a new method to identify disease-relevant nonsynonymous single nucleotide variants.

Chapter 2: Materials and methods

2.1. Whole exome sequencing

2.1.1. Library preparation

We performed exome sequencing on familial pancreatic cancer and triple negative breast cancer samples. The samples have different qualities; therefore, library preparation was different in these two projects.

2.1.1.1. Familial pancreatic cancer susceptibility project

Pancreatic cancer (PC) cases enrolled in the Ontario or Quebec Pancreas Cancer Studies (OPCS, QPCS) were selected. We performed WES for 109 high-risk PC cases from 93 families. These cases were selected on the basis of the family history. They were not carriers of mutations in known PC susceptibility genes such as BRCA2, BRCA1, PALB2, ATM, CDKN2A, PRSS1, SPINK1 and mismatch repair genes. The study group consisted of 8 early onset cases (less than 50 years old at the time of diagnosis) and 101 cases from 85 families with two or more PC-affected relatives. Of the familial cases, WES data were performed for 70 cases with available DNA from a single affected family member and on 15 families with available DNA from multiple PC-affected family members (Table 2.1). We also performed WES on available matched tumor DNA from cases 32B and 72. In one case, 58B, we used existing tumor whole genome sequencing (WGS) data. WES was performed on lymphocyte or white blood cell DNA.

Library capture varied among samples, as during patient collection sequencing technologies evolved rapidly. In this project, exome capture was performed by applying standard protocols and using the Illumina TruSeq Exome Enrichment Kit in 69 cases, Agilent SureSelect Human All Exon V4 in 14 cases and Roche NimbleGen SeqCap EZ kit v3.0 in 26 cases. The

sequencing subsequently performed on Illumina HiSeq2000 platforms with 100 base paired-end reads (Illumina Inc., San Diego, CA, USA). We obtained mean read depth of 61.8±39.8 for target regions. The average percentage of Consensus Coding Sequence (CCDS) bases covered by at least 5, 10 and 20 reads were 94.0, 89.4 and 77.5 respectively. It should be mentioned that the coverage was higher in samples collected later (sequenced by the newer generation capture kits). The mean read depth of sequencing for the tumor samples was 130.8±3.3 with 97.1, 96.1 and 94.5 for average percentage of CCDS bases (covered by at least 5, 10 and 20 reads respectively).

2.1.1.2. Triple negative breast cancer project

Samples for the second hypothesis (TNBC) were collected from five hospital centers (4 in Montreal, QC and 1 in Chicago, IL). Fifty-nine patients were enrolled in this study (Table 2.2). A minimum of 500 ng of DNA was used to generate DNA libraries using Agilent's SureSelect standard protocol. In two samples (Neo42 and Neo28), very low quantities of DNA extracted; therefore, in these two cases libraries were generated using the Nextera DNA library protocol (Illumina Inc.). DNA from matched lymphocytes, pre-chemotherapy breast tumors, and post-chemotherapy tumors (if available) were sequenced. In one case (Neo31) we sequenced a normal breast tumor instead of matched lymphocyte as a control for germline variants. The tumor cell cellularity index or percentage was calculated for each tumor sample based on the report from the pathologist.

Table 2. 1. Clinical characteristics of the 109 PC cases from 93 families at high-risk for hereditary PC that underwent whole exome sequencing.

Characteristics	
Age at diagnosis, mean±SD (range)	61.3±13.2 (20-93)
Gender, n (%)	
Male	54 (49.5)
Female	55 (50.5)
# PC affected per kindred (n=93), n (%)	
1 (young onset)	8 (8.6)
2	51 (54.8)
3	22 (23.7)
≥ 4	12 (12.9)
Stage, n (%)	
0	1 (0.9)
IA	1 (0.9)
IB	3 (2.8)
IIA	15 (13.8)
IIB	36 (33.0)
III	14 (12.8)
IV	38 (34.9)
Unknown	1 (0.9)
Resected, n (%)	
Y	50 (45.9)
N	58 (53.2)
Unknown	1 (0.9)
Chemotherapy, n (%)	
Y	73 (67.0)
N	15 (13.8)
Unknown	21 (19.3)

Table 2. 2. Types of chemotherapy and clinical characteristics of TNBC patients

	Age at diagnosis	Clinical stage at diagnosis	Grade	RCB Score	Neoadjuvant therapy	ER (%)	PR (%)	Her2
NEO-01	42	II	2	0	AT	1-3%	<1	0
NEO-02	51	II	3	3	AT	0	0	0 - 1+
NEO-03	59	I	2	1	AT	0	0	0
NEO-04	32	II	3	2	AT	<1 (30% after revision)	<1	0
NEO-05	57	II	3	3	T	<1	<1	0
NEO-06	52	II	3	0	AT	0	<1	0
NEO-07	57	II	2	2	T	0	0	0
NEO-08	31	II	3	0	AT	0	<1	2+
NEO-09	54	II	3	0	TA	0	0	0
NEO-10	42	II	3	0	Other	<1	<1	2+
NEO-11	55	II	3	0	AT	0	0	0
NEO-12	33	II	3	0	AT	0	0	2+
NEO-13	65	III	3		AT	<1	0	2+
NEO-14	68	III	3	0	AT	<1	<1	0
NEO-15	43	I	3	2	AT	<1	<1	1+
NEO-16	48	II	3	3	Other	<1	1-10%	0
NEO-17	50	III	3	3	AT	<1	0	1+
NEO-18	43	II	3	0	TA	<1	<1	1+
NEO-19	28	III	2	2	AT	<1	0	0
NEO-20	57	II	3	2	T	1-2	5	2+
NEO-21	39	III	3	1	AT	0	0	1+
NEO-22	51	I	3	0	AT	0	0	1+
NEO-23	51	II	3	0	TA	0	0	1+
NEO-24	63	II	3	3	TA	<1	<1	1+
NEO-25	45	II	3	3	AT	<1	<1	1+
NEO-26	36	II	3	1	AT	0	0	1+
NEO-27	40	II	2	2	AT	0	0	2+
NEO-28	52	II	2	2	AT	5	<1	2+
NEO-29	26	II	2	0	AT	0	<1	1+
NEO-30	65	II	3		AT	0	0	1+
NEO-31	51	II	3	2	AT	na	na	2+
NEO-32	40	II	2	2	AT	0	0	0
NEO-33	42	II	3	0	Other	0	0	0
NEO-34	39	III	3	0	Other	<1	0	0

	Age at diagnosis	Clinical stage at diagnosis	Grade	RCB Score	Neoadjuvant therapy	ER (%)	PR (%)	Her2
NEO-35	49	II	3	2	AT	0	0	1+
NEO-36	63	II	2	3	AT	0	0	1+
NEO-37	49	II	3	0	AT	0	<1	0
NEO-38	49	III	3	3	Other	0	3	1+
NEO-39	82	II	3	2	T	0	0	1+
NEO-40	41	II	3	0	AT	1	50	1+
NEO-41	47	II	3	0	AT	0	0	1+
NEO-42	43	III	2	3	Other	0	0	0
NEO-43	66	II	2	1	T	<1	0	0
NEO-44	47	II	3	2	AT	<1	<1	1+
NEO-45	78	II	3	2	T	0	0	1+
NEO-46	54	II	2		AT	0	0	2+
NEO-47	43	III	2	2	AT	3	0	1+
NEO-48	54	II	2	0	AT	0	0	1+
NEO-49	48	III	3	0	AT	<1	0	2+
NEO-50	47	II	2	2	T	0	0	2+
NEO-51	38	II	3	2	AT	0	0	1+
NEO-52	48	II	3	0	TA	0	5	0
NEO-53	33	III	3	0	AT	<1	10	0 or 1
NEO-54	50	III	3	2	AT	4	0	1+
NEO-55	41	II	3	0	TA	0	0	2+
NEO-56	34	II	3			0	15	2+
NEO-57	32	II	3	1	TA	<1	5-10	0
NEO-58	33	II	3	0	TA	0	0	1+
NEO-59	47	II	3	2	TA	<1	80	0
NEO-60	76	II	3	0	TA	0	0	1+
Total		Stage I:3 Stage II:45 StageII:12	Grade 2:15 Grade 3:45	RCB 0:24 RCB 1:5 RCB 2:18 RCB 4:9	T:7 AT:36 TA:10 Other:6			

A: Anthracycline, T: Taxane

2.1.2. Pipeline of Whole exome sequencing data analysis

Whole exome sequencing was processed using our pipeline (Figure 2.1). Briefly, the following workflow was performed.

FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) was performed in order to control the quality of generated reads from the sequencer. Then adaptor sequences and low-quality bases were removed using the Fastx toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). High-quality trimmed reads were aligned to the human reference genome (UCSC hg19) using the Burrows-Wheeler Alignment tool (BWA 0.5.9)(Li and Durbin 2010). Insertions/deletions (indels) were re-aligned using Genome Analysis Tool Kit (GATK)(McKenna, et al. 2010). PCR duplicates were marked with Picard (DePristo, et al. 2011). Capture efficiency and coverage of consensus coding sequence (CCDS) bases were assessed by GATK (McKenna, et al. 2010). Single nucleotide variants (SNVs) and indels were called by means of the SAMtools mpileup software. At the end, variants were annotated with ANNOVAR and custom in-house scripts (Li, et al. 2009; Wang, et al. 2010).Variants were annotated for frequency in dbSNP (Database of Single Nucleotide Polymorphisms (dbSNP). Bethesda (MD): National Center for Biotechnology Information), 1000 Genomes Project (Consortium, et al. 2010), NHLBI Exome Variant Server (Exome Variant Server), Exome Aggregation Consortium (ExAC)(Exome Aggregation Consortium (ExAC)), COSMIC (Forbes, et al. 2015) and ClinVar (Landrum, et al. 2014), as well as for in silico pathogenicity prediction scores, SIFT (Kumar, et al. 2009), PolyPhen 2 (Adzhubei, et al. 2010), GERP (Cooper, et al. 2005b) and CADD.(Kircher, et al. 2014)

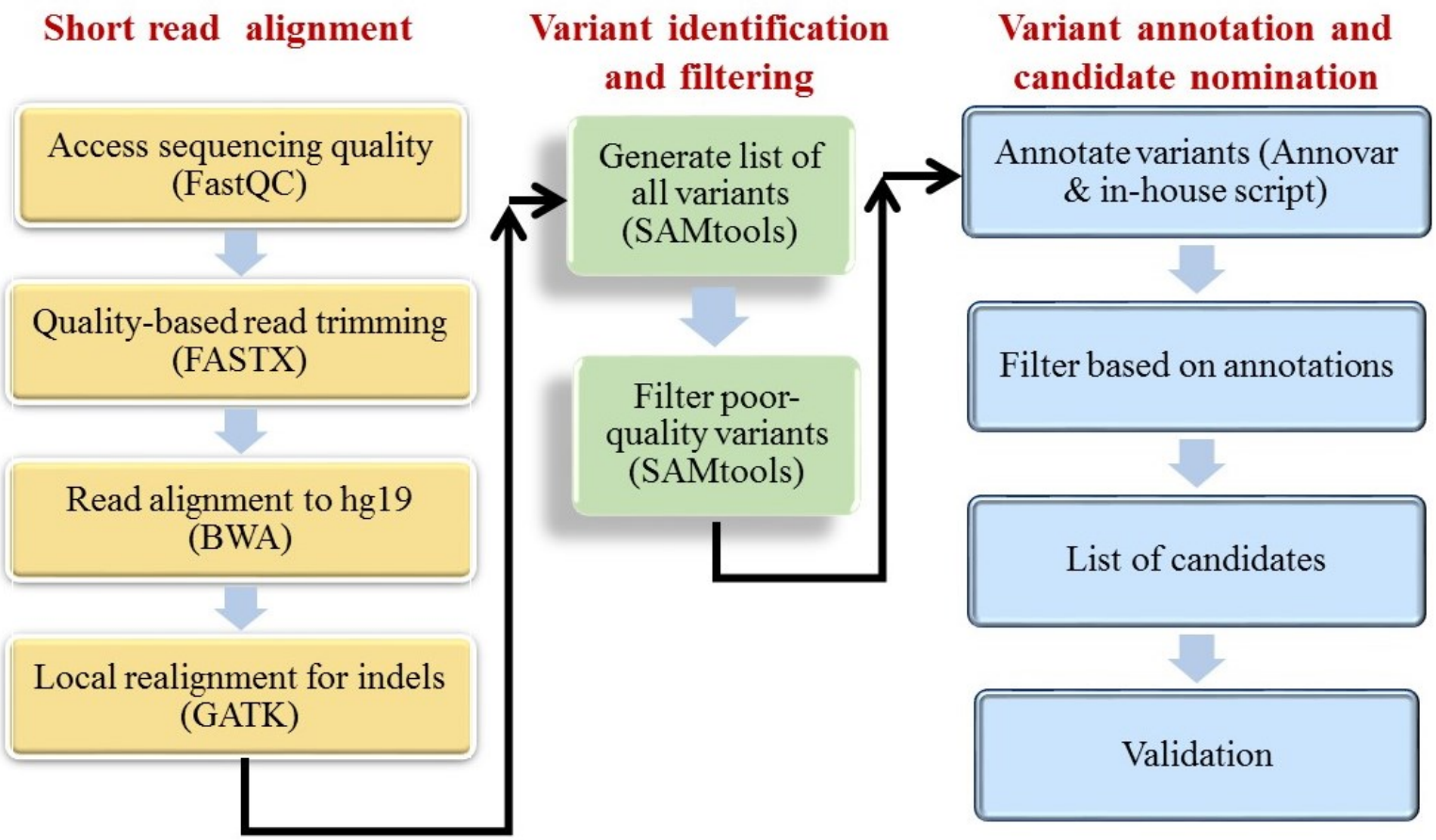


Figure 2. 1. Pipeline for whole exome sequencing analysis

2.1.3. Mutation detection

Since we were looking for different goals in the first two projects (finding germline susceptibility genes in FPC samples and somatic mutations in TNBCs), the criteria that was used for mutation detection varied between these two studies.

2.1.3.1. Variant detection in Pancreatic Cancer

The following criteria were performed to filter variants based on their quality:

1) Base quality score ≥Q20

2) Number of reads ≥3, with at least 2 alternate reads

3) Alternate allele fraction >0.2 for SNVs or >0.15 for indels

Protein-truncating variants (PTVs) most likely to affect protein function were selected. These variants consist of nonsense, frameshift indels and canonical splice-site variants. Since PTVs in the genes that cause PC are predicted to be rare, we filtered any variant with minor allele frequency (MAF) >0.005 in public control databases (dbSNP, 1000 Genomes Project, NHLBI Exome Variant Server) as well as in our in-house database (1,045 exomes from unaffected individuals run through the same pipeline) (Consortium, et al. 2010; Database of Single Nucleotide Polymorphisms (dbSNP). Bethesda (MD): National Center for Biotechnology Information ; Exome Variant Server). Next, any homozygote variants in our case series, the in-house control exomes, or the NHLBI Exome Variant Server were excluded as we hypothesized that those causal genes follow an autosomal dominant inheritance pattern. Then, we retained rare PTVs in genes implicated in DNA repair genes (set of 513 genes) and in recognized PC susceptibility genes not associated with DNA repair (i.e., CDKN2A, PRSS1 and SPINK1). The list of 513 DNA repair gene was compiled from genes identified as DNA repair genes in the Gene Ontology project (via

AmiGO browser)(Carbon, et al. 2009), genes included in the REPAIRtoire database (Milanowska, et al. 2011), and other genes identified through PUBMED literature search. The selected variants were visualized by Integrative Genomics Viewer (IGV) to exclude any sequencing artifacts (Robinson, et al. 2011). The selected list of variants was then validated by Sanger sequencing.

Finally, we searched for any missense and in-frame indel variants in DNA repair genes harbouring at least one Sanger-confirmed PTV. We used the same quality and allele frequency filtering criteria mentioned before for PTV selection. Only missense variants predicted as pathogenic by four in silico prediction algorithms were selected. The criteria were as follow: SIFT score <0.05), PolyPhen 2 score >0.909, GERP score >2 and CADD_Phred score >15. Variants annotated in ClinVar (Landrum, et al. 2014) as "benign" or "likely benign" were also excluded. Visual inspection by IGV was performed and final selected variants were confirmed by Sanger sequencing

2.1.3.2. Mutation detection in TNBC samples

The following criteria were performed to filter variants in these samples:

1) Any variants with minimum 10 reads in that location

2) Mapping Quality (MAPQ) of reads < 40

3) Any variant present in more than 2 non-cancer samples or ranking of the gene as less than 100th in the ranking of common mutated genes

For further analysis, we included only somatic nonsynonymous SNVs, stopgains, frameshift indels, and non-frameshift indels not occurring in repetitive regions. We excluded any UTRs or

synonymous single nucleotide variants (SNVs). To estimate the real allele frequency (AF), final AFs were corrected by dividing the allele frequency by tumor cellularity:

$$\text{Corrected allele frequency} = \frac{\frac{\text{alternate base reads}}{\text{Total reads}}}{\text{tumor cellularity}}$$

We considered any variant as deleterious if the CADD-PHRED score >10, SIFT < 0.5 or Polyphen2 > 0.5. In order to analyze DNA copy number from WES data, the Nexus Copy Number software (Illumina) was employed to import BAM files from whole exome sequencing.

2.2. Segregation analysis

Segregation analysis was performed on families using available WES data from multiple PC-affected family members. In cases, where archived formalin-fixed paraffin-embedded (FFPE) non-tumor tissue samples were available from relatives affected with PC, genomic DNA was extracted to be tested for segregation using Sanger sequencing. For family without samples from PC-affected family members, we used DNA from unaffected family members.

2.3. Loss of heterozygosity

In two cases WES tumor data and in one case WGS tumor data were available (52B, 72 and 58B). Loss of heterozygosity (LOH) or somatic inactivation was assessed in these samples. In cases where archived FFPE tumor blocks were available, Sanger sequencing was performed to assess LOH. LOH was determined by visually comparing allelic ratios of tumor and normal tissue DNA.

2.4. RNAseq analysis

In TNBC project, we performed RNA-seq of 14 Post-chemotherapy/Pre-chemotherapy pairs. All samples were sequenced using Illumina HiSeq 2000, 100 nucleotide paired-end reads, generating a sample of approximately 50 million reads. The sequencing data were first trimmed with CutAdapt and mapped to the human reference genome RefSeq (hg19) using STAR aligner. The STAR-Fusion tool was employed for fusion detection, and only fusions that had at least one junction read and six spanning flags were kept for further screening.

2.5. Array CGH

Dr. Basik group in Jewish general hospital performed array CGH for TNBC project. Briefly, sample preparation and hybridization was done based on the manufacturer's protocol. Copy number alterations were identified by array CGH analysis using the 244 000 (244 K) oligonucleotide probe microarray slides (Agilent Technologies, Santa Clara, CA, USA).

Reading, pre-processing, and segmentation of aCGH Agilent FE files were performed using "limma", "cghMCR" , "CNTools" and "DNAcopy" packages of Bioconductor.

2.6. Pathogenicity prediction model

2.6.1 Training dataset

The ClinVar database dated January 2016 was downloaded and nonsynonymous variants categorized as (1) benign or likely benign or (2) pathogenic or likely pathogenic, were selected as our negative (Benign) and positive (Pathogenic) labels respectively. All variants with conflicting interpretations in the clinical significance reports were excluded. We restricted our training data to the high confidence variants with review status of "criteria provided" from submitter or

"reviewed by expert panel". As a result, 32,910 variants were picked. Next, the variants added to ClinVar prior to January 2013 were eliminated to minimize overlap with training data of the component features of our predictors and the tools being compared. Since the training data of PolyPhen-2 and CADD overlap with our training set, to prevent type 1 circularity, any variants existing in their training data were excluded from our data set. Only missense variants were retained, resulting in the 11,082 variants, with 7,059 labeled as Benign and 4,023 labeled as Pathogenic.

2.6.2. Test datasets

We assembled eight test datasets. The first independent test dataset (ClinVarTest) was constructed from missense variants that were added to the ClinVar database after January 2016 to minimize any overlap with the training data of our features as well as other available deleteriousness prediction tools' training data. Any variant that was evaluated before 2016 was excluded from this data. To further investigate generalizability of our model with respect to data collection method, we constructed our second and third datasets from different sources.

The second distinct database comprised pathogenic variants in mutagenetix database (http://mutagenetix.utsouthwestern.edu). This is a database of phenotypes and mutations produced through random germline mutagenesis induced with N-ethyl-N-nitrosourea (ENU) in mice. Phenotypic mutations responsible for a particular phenotype were obtained from the mutagenetix database. The UCSC genome browser LiftOver tool was applied to convert genome coordinates and annotation from mouse to human GRCH37. Only variants that cause the same amino acid changes in humans and mice were kept. We obtained our neutral SNVs for the second test data from the VariSNP database, which is the benchmark database for neutral-SNVs (Schaafsma and Vihinen 2015). In order to prevent the type II circularity that arises when all the variants in a gene

are labelled either pathogenic or benign (Grimm, et al. 2015), we retained only genes that contained both benign (VariSNP) and pathogenic (mutagenetix) variants to create the MouseVariSNP test data (Figure 2.2).

The third dataset consisted of variants from DoCM, a database of curated mutations in cancers derived from the literature (Ainscough, et al. 2016). We retained only missense variants labelled as pathogenic and likely pathogenic to form the DoCM test data. Since this database contains only pathogenic variants, we used this test set to compare the sensitivity of the different methods.

In order to determine if performance differs between the gain of function and the loss of function gene products, we next constructed four distinct subset datasets from ClinVarTest. Oncogene test data consists of 242 Benign and 112 Pathogenic variants in genes defined as oncogenes according to the ONGene database.(Liu, et al. 2017) The tumor suppressor gene (TSG) dataset consists of 635 variants (475 Benign and 160 Pathogenic) on the basis of genes defined as TSG in the TSGene database. (Zhao, et al. 2016) Gain of function (GOF) and loss of function (LOF) datasets were collected according to the gene-disorder relationship as curated by the Orphanet database (http://www.orpha.net/). A description of the datasets is given in Table 2.3.Any variants that existed in our training data and the training data of our features were discarded from all test datasets.

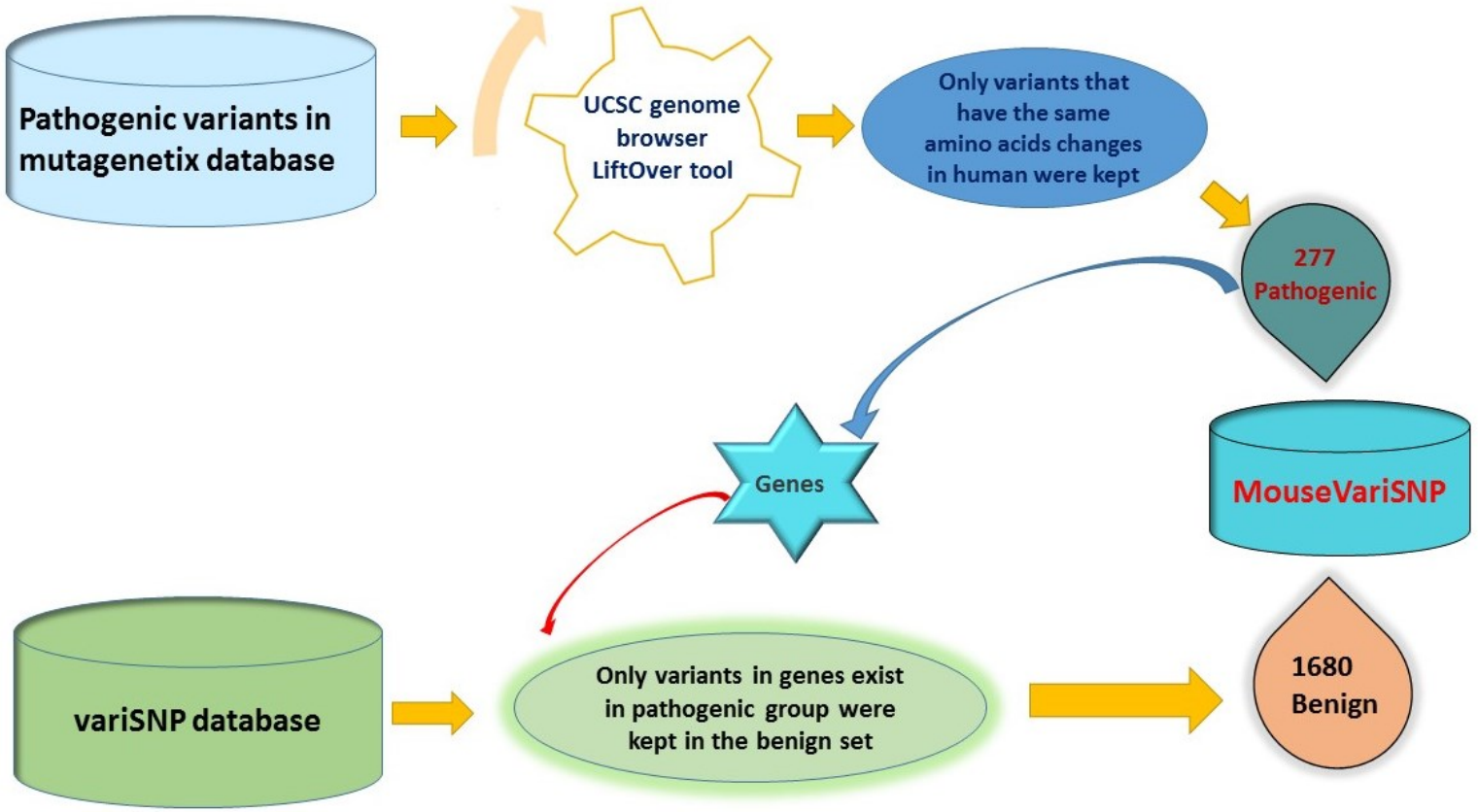


Figure 2. 2. Description of the MouseVariSNP dataset.

This dataset was compromised of pathogenic variants in mutagenetix database and neutral variants from VariSNP database.

Importantly, to test the application of ClinPred in clinically relevant data, the next dataset comprised 31 exome cases of rare disease obtained from the FORGE Canada, Care4Rare Canada Consortia and collaborators (Beaulieu, et al. 2014). These samples were considered solved if the variant under consideration was in a known gene and the referring clinician provided feedback that this gene explained the affected individual's phenotype. Also, in the case of novel disease genes, the variant was considered likely causative for the clinical phenotype in the presence of genetic validation (multiple families with mutations in the same gene and similar phenotype) and/or strong functional evidence. Since all of these data were novel and published after mid-2015, they have not been used to train any predictor.

Finally, to evaluate how ClinPred matches the results of large-scale functional assay data, we constructed the BRCA1 dataset from "A Database of Functional Classifications of BRCA1 Variants based on Saturation Genome Editing "(Findlay, et al. 2018). This test data consists of 437 missense loss of function (LOF) and 1464 functional variants from genome editing in 13 BRCA1 exons that encode critical RING and BRCT domains.

Table 2. 3. Description of datasets that were used in chapter 5

Data		Total variants	Benign	Pathogenic
Training data		11082	7059	4023
Test data	ClinVar Test	5759	4169	1590
	MouseVariSNP	1897	1680	217
	DoCM	1189	0	1189
	LossFunction	1066	776	290
	GainFunction	293	160	133
	Oncogene	354	242	112
	TSG	635	475	160
	BRCA1	1901	437	1464

2.6.3. Features

Having collected the high confidence sets of SNVs, we annotated them with the latest version of ANNOVAR using dbNSFP version 3.3a to generate the required prediction scores from different component tools. Allele frequencies (AF) of each variant in different populations were obtained from the gnomAD database (all exome, African/African American [AFR], Latino [AMR], Ashkenazi Jewish [ASJ], East Asian [EAS], Finnish [FIN], Non-Finnish European [NFE], South Asian [SAS], Other [OTH]). These AFs were assigned zero if the variant was not represented. The potential clinical relevance of each variant is predicted by incorporating AFs and 16 individual

prediction scores: SIFT, PolyPhen2 HDIV, PolyPhen2 HVAR, LRT, MutationAssessor, PROVEAN, CADD, GERP++, DANN, phastCons, fitCons, PhyloP, and SiPhy.(Choi, et al. 2012; Gulko, et al. 2015; Kircher, et al. 2014; Quang, et al. 2015; Reva, et al. 2011) These features were selected as they represent a wide range of orthogonal information; they were not trained on any datasets or their training data is publicly available, thus allowing exclusion from our data and preventing type-I circularity.

2.6.4.. Model definition

We applied random forest (cforest) and gradient boosted decision tree (xgboost) models and used the default missing value predictions these algorithms provide for cases where individual scores for component predictors are not available.

We trained each model using either balanced or equal weights:

- Equal weights assign a weight of one to each example in the training set. For highly unbalanced datasets (e.g. 90% Pathogenic, 10% Benign), these models may be trivial/sub-optimal.
- Balanced weights assign a weight to each example so that the total weight of each class is equal. For example, if there are 900 Pathogenic and 100 Benign variants, we assign each Pathogenic variant a weight of 1 (total weight = 1*900=900), and each Benign variant a weight of 9 (total weight=9*100=900).

As the results from the balanced weight model were not significantly different from the equal weight one (data not shown), for simplicity, we show only balanced weight results for our final models. The output of each model is a score between zero and one, with zero corresponding to Benign and one to Pathogenic variants. In addition, in order to maximize the sensitivity for

detecting pathogenic variants, we defined the higher score of either of these two models (xgboost, cforest) as the ClinPred score.

2.6.5. Comparing the Performance of Individual Predictors

For each input feature and the comparator models, we learned a univariate model based on each training set as follows. We learn the sign (-1 if smaller scores are more likely to be pathogenic; 1 otherwise) and threshold (score *sign <= threshold predicts Benign; otherwise Pathogenic) that minimizes the number of incorrect predictions using all non-missing features in the training data. This threshold was used to compute evaluation metrics.

To quantitatively compare our models with individual features and other models, we performed 5-fold cross validation on training, ClinVarTest and MouseVariSNP data. Each data set was randomly partitioned into five equal sized subsamples. In each round of cross-validation, our models were trained on 80% of training data and tested on 20% of the test data. In order to allow for fair comparison with available methods, the thresholds of other models were learned during the cross validation to maximize accuracy. Thus, the performance of our models was compared to other recent state-of-the-art tools such as VEST3, MetaSVM, MetaLR, M-CAP, fathmm-MKL, Eigen, GenoCanyon and REVEL (Carter, et al. 2013; Dong, et al. 2015; Ioannidis, et al. 2016; Ionita-Laza, et al. 2016; Jagadeesh, et al. 2016; Lu, et al. 2015; Shihab, et al. 2015). Following the guidelines for reporting and using prediction tools, we computed seven evaluation metrics on each test based on the learned threshold described above. These metrics include sensitivity (true positive rate), specificity (1-false positive rate), accuracy, precision, F1 Score and Matthew correlation coefficient (MCC) as well as the area under the test receiver operating characteristic curve (AUC) (Niroula and Vihinen 2016; Vihinen 2013).

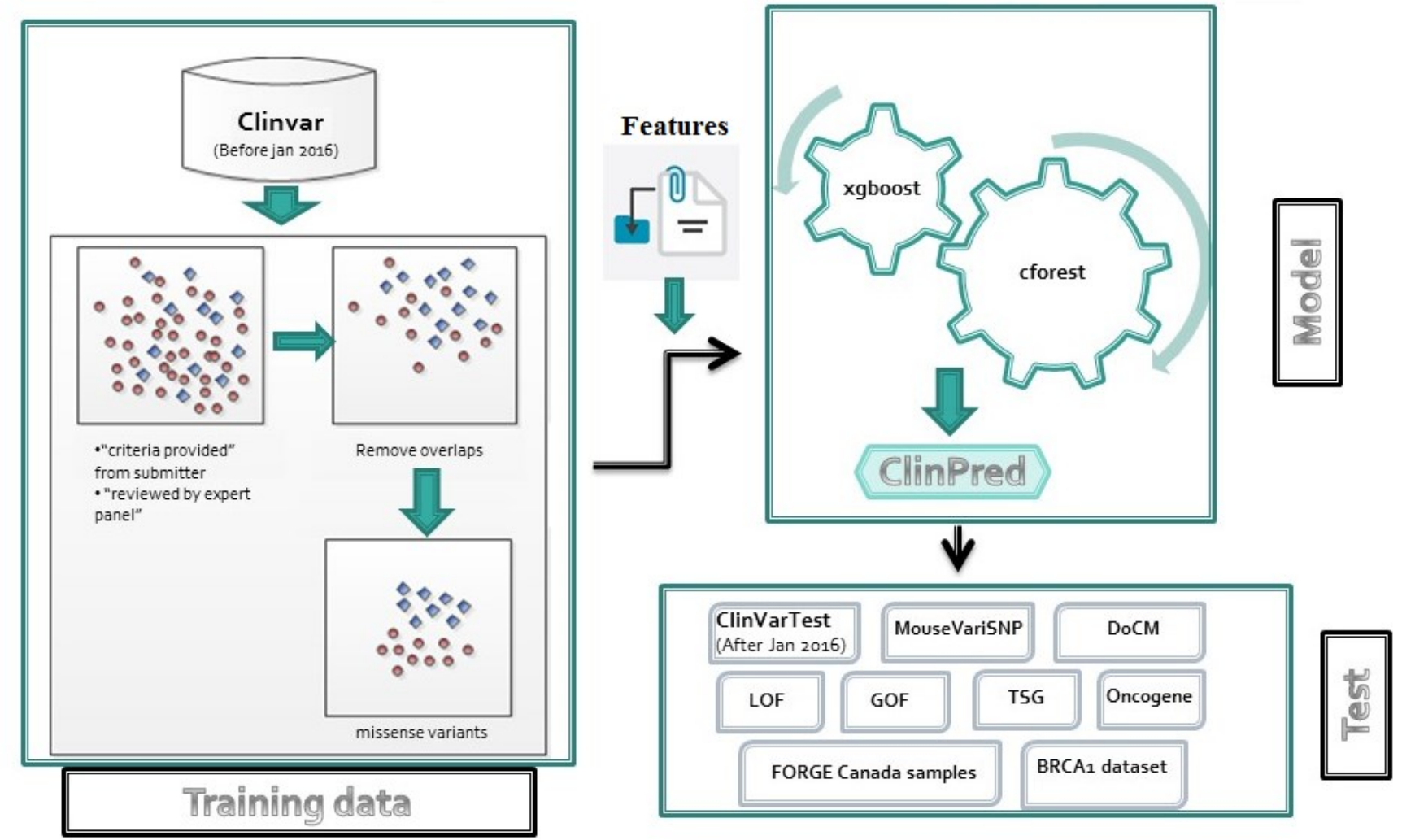


Figure 2. 3. Description of the ClinPred method

Chapter 3: Candidate DNA repair susceptibility genes in familial pancreatic cancer

Part of the figures and tables from this chapter are published as:

"Detecting Candidate DNA Repair Susceptibility Genes in Patients at Increased Risk of Hereditary Pancreatic Cancer Using Exome Sequencing" Alyssa L. Smith, Najmeh Alirezaie, Ashton Connor, Michelle Chan-Seng-Yue, Robert Grant, Iris Selander, Claire Bascuñana, Ayelet Borgida, Anita Hall, Thomas Whelan, Spring Holter, Treasa McPherson, Sean Cleary, Gloria M. Petersen, Atilla Omeroglu, Emmanouil Saloustros, John McPherson, Lincoln D. Stein, William D. Foulkes, Jacek Majewski, Steven Gallinger, and George Zogopoulos. Cancer Letter, 2016 January 28; 370(2): 302–312.

Author contributions are stated in the Contribution of the Authors section.

3.1. Introduction

Pancreatic ductal adenocarcinoma (PC) has a very poor prognosis. The majority of the affected patients are diagnosed late when the tumor is inoperable. Thus, current treatment options are limited and largely ineffective. Around 10 percent of PDACs occur in families. In a family affected by pancreatic cancer (FPC), at least one pair of first-degree relatives is affected. Although some environmental factors or stochastic effects can be the underlying cause in some FPCs, many are thought to be due to an underlying genetic susceptibility; hence, programs that help to detect individuals at increased risk may improve clinical outcomes (Shi, et al. 2009). Although PC develops over a decade following the primary somatic mutation, screening strategies that just consider family history were not effective (Al-Sukhni, et al. 2012; Langer, et al. 2009; Yachida, et al. 2010). Therefore, full knowledge of causative germline mutations will help to identify high-risk individuals and allow for more specific screening programs.

Hereditary PC occurs either alone or as part of a tumor spectrum in families. In 10 to 15 percent of FPCs, the increased risk of pancreatic cancer can be ascribed to known FPC susceptibility genes such as BRCA1, BRCA2, CDKN2A, MLH1, MSH2, MSH6, PMS2, PRSS1, SPINK1, STK11/LB1, PALB2 and ATM (Jones, et al. 2009). The genetic basis underlying this disease predisposition in the remaining 85-90 percent of patients is still unknown. Therefore, there are other FPC susceptibility genes yet to be discovered.

Findings from prior researches on FPC patients suggest autosomal dominant inheritance of a rare allele(s) with variable penetrance for remaining unknown susceptibility genes (Klein, et al. 2002). Further linkage and genome-wide association studies were not successful to find novel medium or high penetrant PC susceptibility loci (Childs, et al. 2015; Klein 2013). Since DNA

repair genes are widely implicated in gastrointestinal malignancies, and account for the majority of hereditary PC attributable to known PC predisposition genes (BRCA1, BRCA2, ATM, PALB2, mismatch repair genes), we hypothesized that additional DNA repair genes are involved in hereditary PC (Bartsch, et al. 2012; Hruban, et al. 2010; Rubinstein and Weissman 2008). Therefore, we employed a DNA repair candidate gene approach to interrogate whole exome sequencing (WES) data for novel susceptibility genes (Bartsch, et al. 2012; Hruban, et al. 2010).

3.2. Results

3.2.1. Whole exome sequencing

WES data was performed on 109 high-risk PC cases from 93 families. The list of 513 DNA repair gene was compiled from genes identified as DNA repair genes in the Gene Ontology project (via AmiGO browser)(Carbon, et al. 2009), genes included in the REPAIRtoire database(Milanowska, et al. 2011) and other genes identified through PUBMED literature search (Table 3.1). Patient selection criteria and WES process was explained in method section and the algorithm briefly outlined in Figure 3.1.

Table 3. 1. List of 513 DNA repair genes compiled from the Gene Ontology project, REPAIRtoire database and PUBMED.

DNA Repair Genes							
AATF	CDKN2A	ERCC8	INO80D	NTHL1	PSMD3	SIRT1	TP73
ABL1	CDKN2D	ESCO1	INO80E	NUDT1	PTTG1	SIRT6	TREX1
ACTR5	CEBPG	ESCO2	INTS3	OGG1	RAD1	SLC30A9	TREX2
AKT1	CEP164	ESR1	IRS1	OTUB1	RAD17	SLX1A	TRIP12
ALKBH1	CEP170	ETS1	JMY	PALB2	RAD18	SLX4	TRIP13
ALKBH2	CETN2	EXO1	JUN	PAPD7	RAD21	SMAD2	TTC5
ALKBH3	CHAF1A	EXO5	KAT5	PARG	RAD23A	SMAD3	TWIST1
AP5S1	CHAF1B	EYA1	KDM2A	PARP1	RAD23B	SMAD4	TYMS
AP5Z1	CHD1L	EYA2	KIAA0101		RAD50	SMAD7	UBA1
APEX1	CHD4	EYA3	KIAA0430	PARP2	RAD51	SMARCA1	UBA52
APEX2	CHEK1	EYA4	KIAA2022	PARP3	RAD51AP1	SMARCA2	UBB
APITD1	CHEK2	FAM175A	KIF22	PARP4	RAD51B	SMARCA4	UBC
APLF	CHRNA4	FAN1	KIN	PARP9	RAD51C	SMARCA5	UBE2A
APTX	CIB1	FANCA	KPNA2	PARPBP	RAD51D	SMARCAD1	UBE2B
ASCC3	CINP	FANCB	LIG1	PCNA	RAD52	SMARCB1	UBE2D3
ASF1A	CLSPN	FANCC	LIG3	PLK1	RAD54B	SMARCC2	UBE2I
ASTE1	COPS5	FANCD2	LIG4	PLK3	RAD54L	SMARCD1	UBE2N
ATF2	CRB2	FANCE	MAD2L2	PMS1	RAD9A	SMARCD2	UBE2NL
ATM	CREB1	FANCF	MBD4	PMS2	RAD9B	SMC1A	UBE2T
ATMIN	CREBBP	FANCG	MC1R	PNKP	RASSF1	SMC2	UBE2U
ATR	CRY1	FANCI	MCM9	POLA1	RB1	SMC3	UBE2V2
ATRIP	CRY2	FANCL	MCPH1	POLB	RBBP4	SMC4	UBE4B
ATRX	CSNK1D	FANCM	MDC1	POLD1	RBBP7	SMC5	UHRF1
ATXN3	CSNK1E	FBXO18	MDM2	POLD2	RBBP8	SMC6	UIMC1
AXIN2	CUL4A	FBXO6	MDM4	POLD3	RBM14	SMG1	UNG
BABAM1	CUL4B	FEN1	MED17	POLD4	RBX1	SMUG1	UPF1
BAP1	CYP19A1	FGF10	MEIOB	POLDIP3	RDM1	SMURF2	USP1
BARD1	CYP1A1	FHIT	MEN1	POLE	REC8	SOD1	USP28
BAX	DAPK1	FIGN	MGME1	POLE2	RECQL	SP1	USP3
BAZ1B	DBF4	FIGNL1	MGMT	POLE3	RECQL4	SPATA22	USP47
BCCIP	DCLRE1A	FOS	MLH1	POLE4	RECQL5	SPIDR	USP7
BLM	DCLRE1B	FOXM1	MLH3	POLG	RELA	SPO11	UVRAG
BRAP	DCLRE1C	FTO	MMS19	POLG2	REV1	SPP1	UVSSA
BRCA1	DDB1	FZR1	MMS22L	POLH	REV3L	SPRTN	VCP
BRCA2	DDB2	GADD45A	MNAT1	POLI	RFC1	SSRP1	WDR16
BRCC3	DDR1	GADD45G	MORF4L1	POLK	RFC2	STAT1	WDR33
BRE	DDX1	GEN1	MORF4L2	POLL	RFC3	STRA13	WDR48
BRIP1	DEK	GPS1	MPG	POLM	RFC4	SUMO1	WEE1
BTG2	DHX9	GSTP1	MRE11A	POLN	RFC5	SUPT16H	WHSC1
BUB1	DMAP1	GTF2H1	MSH2	POLQ	RFWD2	SWI5	WRN
BUB1B	DMC1	GTF2H2	MSH3	POLR2A	RFWD3	SWSAP1	WRNIP1
C11orf30	DNA2	GTF2H2C	MSH4	POLR2B	RHNO1	SYCP1	WWP1
C17orf70	DOT1L	GTF2H3	MSH5	POLR2C	RNASEH2A	TAOK1	WWP2
C19orf40	DTL	GTF2H4	MSH6	POLR2D	RNF168	TAOK2	XAB2
CASP3	DTX3L	GTF2H5	MTA1	POLR2E	RNF169	TAOK3	XPA
CCNA1	DUSP3	H2AFX	MUM1	POLR2F	RNF8	TCEA1	XPC
CCNA2	DYRK2	HDAC1	MUS81	POLR2G	RPA1	TDG	XRCC1
CCNB1	E2F1	HDAC2	MUTYH	POLR2H	RPA2	TDP1	XRCC2
CCND1	E2F2	HELQ	MYC	POLR2I	RPA3	TDP2	XRCC3
CCNE1	E2F4	HERC2	NABP1	POLR2J	RPA4	TELO2	XRCC4

DNA Repair Genes							
CCNH	E2F6	HIC1	NABP2	POLR2K	RPAIN	TERF1	XRCC5
CCNO	EEPD1	HINFP	NBN	POLR2L	RPS27A	TERF2	XRCC6
CDC14B	EGFR	HIST3H2A	NCOA6	PPM1D	RPS27L	TERF2IP	XRCC6BP1
CDC25A	EME1	HMGB1	NEIL1	PPP1CA	RPS3	TEX12	YY1
CDC25B	EME2	HMGB2	NEIL2	PPP2R2A	RRM2B	TEX15	ZBTB32
CDC25C	ENDOV	HUS1	NEIL3	PPP2R5A	RTEL1	TICRR	ZFYVE26
CDC45	EP300	HUS1B	NEK1	PPP2R5B	RUVBL1	TMEM161A	ZNF350
CDC6	EPC2	HUWE1	NEK11	PPP2R5C	RUVBL2	TNP1	ZRANB3
CDH13	ERBB2	IFI16	NFKB1	PPP2R5D	SETD2	TONSL	ZSWIM7
CDK1	ERCC1	IGF1	NHEJ1	PPP2R5E	SETMAR	TOP1	
CDK2	ERCC2	IGHMBP2	NINL	PPP4C	SETX	TOP2A	
CDK4	ERCC3	IKBKG	NME1	PPP4R2	SFPQ	TOP3A	
CDK7	ERCC4	INIP	NONO	PRKDC	SFR1	TOPBP1	
CDKN1A	ERCC5	INO80	NSMCE1	PRMT6	SHFM1	TP53	
CDKN1B	ERCC6		NSMCE2	PRPF19	SHPRH	TP53BP1	

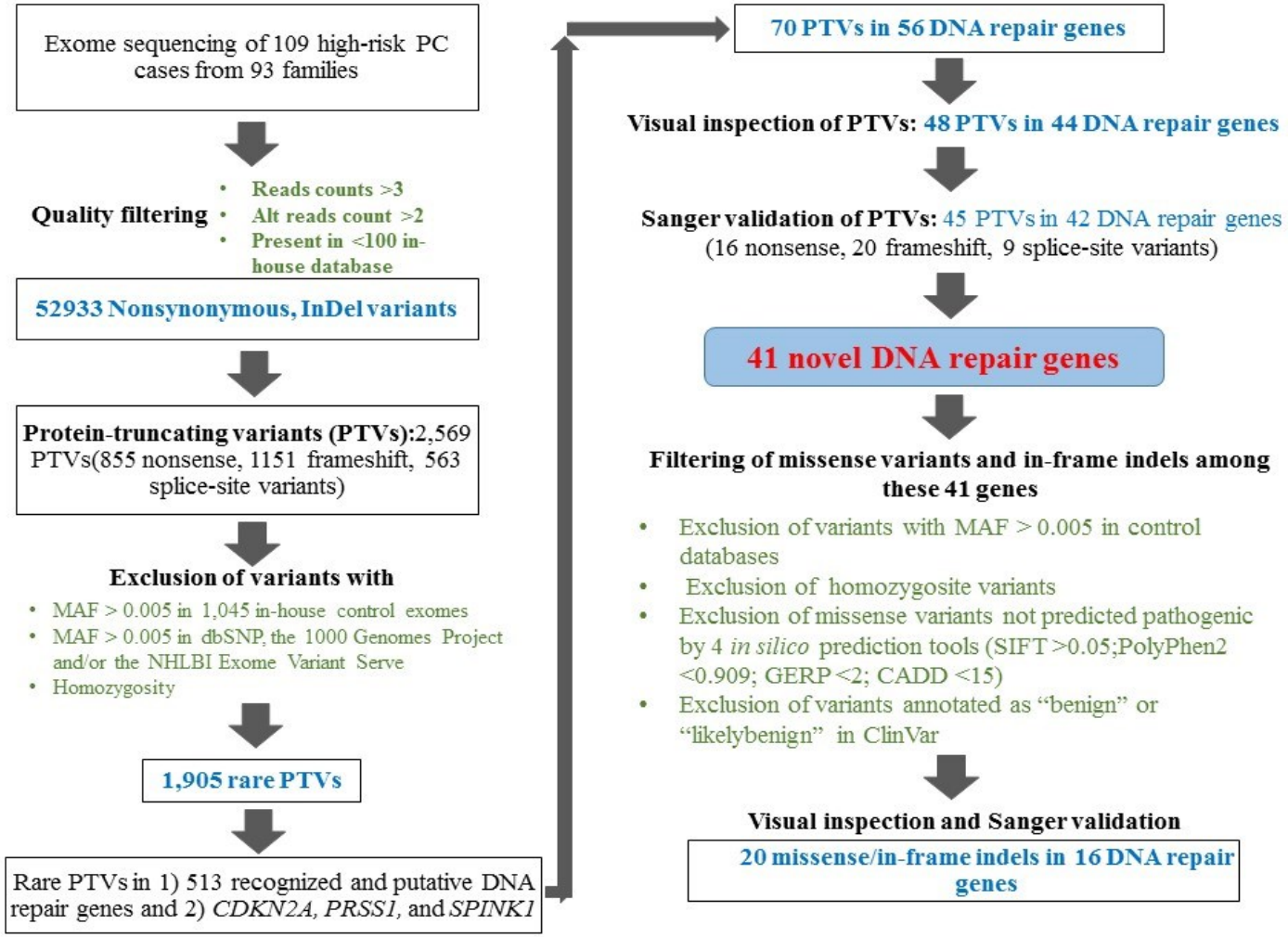


Figure 3. 1. Schematic of the exome sequencing data analysis in FPC project.

Variants remaining after each filtering step are indicated. SNV, single nucleotide variant; indel, insertion/deletion; PTV, protein-truncating variant; MAF, minor allele frequency.

3.2.2. Identification of DNA repair gene variants

Variant filtering as described in method section (Figure 3.1) was applied to the germline WES from 109 samples. After performing quality filtering, a total of 52,933 nonsynonymous and indel variants remained. Only variants that cause protein truncating (PTVs) were kept (n=2569). We then exclude homozygous variants and retained only variants with MAF < 0.005 in our in-house control exome database (1045 unaffected samples), the 1000 Genomes Project or the NHLBI Exome Variant Server (n=1905 rare PTVs). Any variant that was located in our 513 DNA repair gene list or CDKN2A, PRSS1, and SPINK1 genes was picked for further evaluation. We could identify 70 variants in 56 DNA repair genes. Then these variants were visually inspected by IGV to exclude any false positive. After excluding 22 false positives variants, remaining variants (48 variants located in 44 genes) were further evaluated by Sanger sequencing. Among them, 93.8% (45 PTVs in 42 DNA repair genes) were validated by Sanger sequencing. Of these variants, 16 were nonsense, 20 were frameshift indels, and nine were splice-site variants. Description of these 45 PTVs variants was given in table 3.2. Forty-one PC cases in 37 families had one or more PTVs in a DNA repair gene. One of these variants were located in BRCA2 that is a known PC susceptibility gene [BRCA2:c.4691dupC (p.Thr1566Aspfs*9)]. Therefore, we excluded this variant.

Of the remaining 41 novel genes identified in 36 families, four genes (FANCL, MC1R, NEK1 and RHNO1) had PTVs in multiple families. Seven individuals were carriers of two PTVs; one individual was a carrier of three PTVs, while two families had different affected family members carrying different PTVs (Table 3.2).

Table 3. 2. Description of the 45 PTVs validated by Sanger sequencing.

Modified from Smith et al., Cancer Letter, 2016.

Gene	Chr. Pos.	Variant (HGVS nomenclature)	MAF IH[a]	MAF EVS	MAF 1000G	MAF ExAC
AATF	chr17:35307578	c.158_159dup (p.Gly54Trpfs*157)	0	0	0	0
BARD1	chr2:215595181	c.1935_1954dup (p.Glu652Valfs*69)	0	0	0	0
BCCIP	chr10:127520177	c.599+1G>A	0.00087	0.00038	0	2.36E-04
BLM	chr15:91292792	c.298_299del (p.Gln100Glufs*42)	0	0	0	0
BRCA2	chr13:32913182	c.4691dupC (p.Thr1566Aspfs*9)	0	7.99E-05	0	0
C17orf70	chr17:79518071	c.449G>A (p.Trp150*)	0	0	0	0
CDC6	chr17:38449789	c.743del (p.Gln248Argfs*16)	0	0	0	0
CEP164	chr11:117282575	c.4228C>T (p.Gln1410*)	0.0017	0.0016	0.0004	7.89E-04
CHD1L	chr1:146742591	c.1086-2A>G	0	0	0	0
DCLRE1A	chr10:115612530	c.C412T (p.Arg138*)	0.0048	0.0024	0.002	0.0027
DNA2	chr10:70182188	c.2493-2A>G	0	0	0	0
ENDOV	chr17:78395694	c.295C>T (p.Arg99*)	0.00087	7.80E-05	0.0044	0.0012
ERCC6	chr10:50680423	c.2923C>T (p.Arg975*)	0	0	0	2.44E-05
FAN1	chr15:31214513	c.2128C>T (p.Arg710*)	0	7.70E-05	0	2.44E-05
FANCG	chr9:35074472	c.1652_1655del (p.Tyr551Phefs*7)	0	0	0	0
FANCL	chr2:58386928	c.1096_1099dup (p.Thr367Asnfs*13)	0.0013	0.0025	0	0
HUS1	chr7:48018013	c.357+1G>A	0	0	0	2.44E-05
IGHMBP2	chr11:68701332	c.1488C>A (p.Cys496*)	0	0.00015	0.0002	1.47E-04
MC1R	chr16:89985733	c.67C>T (p.Gln23*)	0	0.00092	0.0014	4.01E-04
MC1R	chr16:89985750	c.86dup (p.Asn29Lysfs*14)	0.0044	0.003	0	0
MC1R	chr16:89986122	c.456C>A (p.Tyr152*)	0	0.00023	0.0002	6.67E-04
MGMT	chr10:131565137	c.593G>A (p.Trp198*)	0.00044	0	0	1.22E-04
MLH3	chr14:75509094	c.3367C>T (p.Gln1123*)	0	0.00015	0.0004	8.95E-05
NEIL1	chr15:75641315	c.330_331insAGGC (p.Ala111Argfs*46)	0.0017	0	0	0

Gene	Chr. Pos.	Variant (HGVS nomenclature)	MAF IH[a]	MAF EVS	MAF 1000G	MAF ExAC
NEK1	chr4:170428209	c.1687_1688del (p.Ala563Tyrfs*36)	0	0	0	0
NEK11	chr3:130828766	c.455+1G>A	0	0	0	1.63E-05
NINL	chr20:25434092	c.4142_4143del (p.Ser1381Cysfs*17)	0.00044	0	0	0
PARG	chr10:51363054	c.1018_1019insG (p.Lys340Argfs*11)	0	0	0	2.66E-05
PARP3	chr3:51978471	c.401del (p.Lys134Argfs*33)	0	0	0	0
PMS1	chr2:190719824	c.1826G>A (p.Trp609*)	0	0	0	0
POLE3	chr9:116172359	c.127del (p.Val43Serfs*15)	0	0	0	0
POLL	chr10:103345072	c.573+1G>A	0	0	0	0
POLN	chr4:2230817	c.133del (p.Thr45Leufs*4)	0	0	0	0
POLQ	chr3:121217455	c.2021dup (p.Lys675Glufs*16)	0.00044	7.99E-05	0	0
RFC2	chr7:73646495	c.1006C>T (p.Gln336*)	0	0	0	0
RHNO1	chr12:2997158	c.250C>T (p.Arg84*)	0.0017	0.0016	0.0008	0.0012
RHNO1	chr12:2997245	c.337C>T (p.Arg113*)	0	0	0	0
SMC2	chr9:106875703	c.1365_1366del (p.Arg456Thrfs*2)	0	0	0	0
SPP1	chr4:88901197	c.94-1G>A	0.00087	0.00077	0.000599	8.46E-04
TEX15	chr8:30700833	c.5699_5700del (p.Arg1900Asnfs*22)	0.0026	0.00072	0	0
TONSL	chr8:145668147	c.490del (p.Leu164Serfs*72)	0	0	0	0
UBE2U	chr1:64707361	c.622C>T (p.Gln208*)	0.00044	0.0032	0.002	9.52E-04
WDR48	chr3:39125749	c.1278_1279del (p.Gly427Aspfs*8)	0	0	0	2.44E-05
WRN	chr8:30999118	c.3138+2T>G	0	0	0	0
ZSWIM7	chr17:15897070	c.98+1G>A	0	0	0	8.18E-06

Chr. Pos. chromosomal position; MAF, minor allele frequency; IH, in-house; EVS, NHLBI Exome Variant Server; 1000G, 1000 Genomes Project; ExAC, Exome Aggregation Consortium;

a In-house unaffected control exomes (n=1,045)

In order to prioritize candidate genes, we then searched for missense and in-frame indel variants in the WES data for these 41 DNA repair genes confirmed by Sanger sequencing (Figure 3.1). We applied the same quality and control filtering that we used for PTVs. Any variant labeled as "benign" in ClinVar was also excluded. Just missense variants predicted to be pathogenic by four in silico prediction tools were retained, resulting 18 missense variants and 2 in-frame indels in 16 DNA repair genes (Table 3.3). Sanger sequencing confirmed all of these variants. Twenty-two PC cases in 19 families had one or more missense variant or in-frame indel. Interestingly, three PC cases were carriers of multiple nonsynonymous variants and five cases were carriers of both a PTV and one or more nonsynonymous variant (Table 3.4). Five genes (DCLRE1A, FAN1, POLQ, TEX15, and TONSL) had a missense variant or in-frame indel in multiple families.

Table 3. 3. Description of the 20 missense and in-frame indels validated by Sanger sequencing

Modified from Smith et al., Cancer Letter, 2016.

Gene	Chr. Pos.	Variant (HGVS nomenclature)	MAF IH[a]	MAF EVS	MAF 1000G	MAF ExAC	SIFT	PolyPhen-2	CADD	GERP
AATF	chr17:35311130	c.755A>G (p.Asn252Ser)	0	0	0	0	0.02	0.986	23.8	5.88
CEP164	chr11:117244534	c.1220C>T (p.Ser407Phe)	0.00044	0.0011	0	8.05E-04	0	0.912	18.48	5.18
CHD1L	chr1:146756048	c.1730G>A (p.Gly373Asp)	0	7.70E-05	0	0	0.01	1	22.7	5.65
DCLRE1A	chr10:115602192	c.2575A>T (p.Ile859Phe)	0.0035	0.004	0.0018	0.0027	0.02	0.926	24.1	3.66
ENDOV	chr17:78399353	c.647G>T (p.Ser171Ile)	0	0.00024	0	1.97E-04	0	0.934	19.78	4.9
ERCC6	chr10:50690906	c.1996C>T (p.Arg666Cys)	0.0026	0.0015	0.001	0.0017	0	0.987	27.1	5.57
FAN1	chr15:31197015	c.149T>G (p.Met50Arg)	0.0013	0.0027	0.0018	0.002	0.01	0.974	22.8	5.15
MC1R	chr16:89986522	c.862_864del (p.Ile288del)	0	0	0	8.29E-06	.	.	.	.
MLH3	chr14:75514503	c.1856A>T (p.Lys619Ile)	0	0	0	8.13E-06	0.01	0.925	16.23	2
NEK1	chr4:170398474	c.2235T>G (p.Asn648Lys)	0.0031	0.0044	0.0016	0.0038	0.01	0.999	21.8	5.57
POLN	chr4:2097622	c.2021G>A (p.Arg674Lys)	0	0	0	0	0	1	16.16	3.68
POLQ	chr3:121151236	c.7688A>G (p.Glu2563Gly)	0.00044	0	0	1.63E-05	0	1	22.8	4.81
POLQ	chr3:121155119	c.7393G>A (p.Glu2465Lys)	0.00044	0.00054	0	3.17E-04	0	1	32	5.81
POLQ	chr3:121168167	c.7259A>G (p.Tyr2420Cys)	0.00044	0.00046	0.0002	3.01E-04	0	0.999	21.2	5.41

Gene	Chr. Pos.	Variant (HGVS nomenclature)	MAF IH[a]	MAF EVS	MAF 1000G	MAF ExAC	SIFT	PolyPhen-2	CADD	GERP
RHNO1	chr12:2994578	c.45_46delinsAG (p.Leu16Val)	0.0048	0	0.0008	0.0037	0	0.999	16.25	4.44
TEX15	chr8:30701070	c.5464T>A (p.Leu1822Ile)	0	7.70E-05	0	1.55E-04	0	0.999	17.18	5.54
TEX15	chr8:30704934	c.1585_1599del (p.Ile529_Glu533del)	0	8.02E-05	0	0	.	.	.	.
TONSL	chr8:145660507	c.2899C>T (p.Arg967Cys)	0	0	0	0	0	0.997	21.2	3.92
TONSL	chr8:145662005	c.1950C>G (p.Asp650Glu)	0.00044	0.0015	0.0004	9.76E-04	0.03	0.991	17.41	2.51
WRN	chr8:31012237	c.3785C>G (p.Thr1262Arg)	0.0017	0.0035	0.0008	0.002749	0	0.974	19.22	5.48

Chr. Pos. chromosomal position; MAF, minor allele frequency; IH, in-house; EVS, NHLBI Exome Variant Server; 1000G, 1000 Genomes Project; ExAC, Exome Aggregation Consortium. a In-house unaffected control exomes (n=1,045)

3.2.3. Segregation analyses

In the next step, we performed segregation analysis for all validated variants. We looked at available WES sequencing data from affected family members and available DNA from affected or unaffected relatives (see method). We should note that we could not find samples for segregation analysis for all the variants. Out of 36 variants that we have family samples for segregation analysis, 18 variants in 14 genes showed segregation with PC in two or more affected family members (Table 3.5). Interestingly, five genes had variants segregated in two families. These genes were AATF, CHD1L, FAN1, NEK1 and RHNO1. Descriptions of these variants are shown in table 3.5.

3.2.4. Loss of heterozygosity analyses

LOH was assessed in all cases where tumor WES data were available or in cases where archived FFPE tumor blocks were available. We could assess LOH for 27 variants in 29 tumors. Out of these variants loss of the wild-type allele observed for three variants [MGMT:c.593G>A (p.Trp198*), RHNO1:c.250C>T (p.Arg84*), WDR48:c.1278_1279del (p.Gly427Aspfs*8)] (Table 3.6). Twenty-two variants in 24 tumors did not show LOH. Loss of the alternate allele was seen in two variants [MLH3:c.1856A>T (p.Lys619Ile) and PMS1:c.1826G>A (p.Trp609*)]. We could not find any second somatic mutation in two WES tumor data that was available for 52B and 72 patients with germline mutation in MC1R and NINL. Similarly, the WES analysis of 58B patient tumor (with germline mutation in FAN1) did not show somatic mutation in this gene (Table 3.6).

Table 3. 4. PC cases with more than one variant in a putative DNA repair gene.

Modified from Smith et al., Cancer Letter, 2016.

Sample ID	Gene	Variant
2	CDC6	c.743del (p.Gln248Argfs*16)
2	TONSL	c.490del (p.Leu164Serfs*72)
3A	HUS1	c.357+1G>A
3A	TONSL	c.2899C>T (p.Arg967Cys)
14	MC1R	c.67C>T (p.Gln23*)
14	UBE2U	c.622C>T (p.Gln208*)
16	CEP164	c.4228C>T (p.Gln1410*)
16	ENDOV	c.647G>T (p.Ser171Ile)
16	POLN	c.2021G>A (p.Arg674Lys)
24A	TEX15	c.5699_5700del (p.Arg1900Asnfs*22)
24A	DCLRE1A	c.2575A>T (p.Ile859Phe)
31B	POLL	c.573+1G>A
31B	RFC2	c.1006C>T (p.Gln336*)
43	NEIL1	c.330_331insAGGC (p.Ala111Argfs*46)
43	RHNO1	c.250C>T (p.Arg84*)
47	FANCL	c.1096_1099dup (p.Thr367Asnfs*13)
47	PARG	c.1018_1019insG (p.Lys340Argfs*11)
47	POLN	c.133del (p.Thr45Leufs*4)
53A	CEP164	c.1220C>T (p.Ser407Phe)
53A	POLQ	c.7393G>A (p.Glu2465Lys)
53B	ERCC6	c.2923C>T (p.Arg975*)
53B	CEP164	c.1220C>T (p.Ser407Phe)
68A	POLQ	c.7688A>G (p.Glu2563Gly)
68A	TEX15	c.5464T>A (p.Leu1822Ile)
72	NINL	c.4142_4143del (p.Ser1381Cysfs*17)
72	WDR48	c.1278_1279del (p.Gly427Aspfs*8)
78A	NEK1	c.1687_1688del (p.Ala563Tyrfs*36)
78A	SPP1	c.94-1G>A
78B	BLM	c.298_299del (p.Gln100Glufs*42)
78B	SPP1	c.94-1G>A
81	ZSWIM7	c.98+1G>A
81	DCLRE1A	c.2575A>T (p.Ile859Phe)
89	MC1R	c.862_864del (p.Ile288del)
89	NEK1	c.2235T>G (p.Asn648Lys)

Table 3. 5. Eighteen variants in 14 genes segregated in families.

We could performed segregation analysis for 36 variants. Modified from Smith et al., Cancer Letter, 2016.

Gene	Chr. Pos.	Variant (HGVS nomenclature)	Samples
AATF	chr17:35307578	c.158_159dup (p.Gly54Trpfs*157)	76A, 76B
AATF	chr17:35311130	c.755A>G (p.Asn252Ser)	32
BLM	chr15:91292792	c.298_299del (p.Gln100Glufs*42)	78B
CHD1L	chr1:146742591	c.1086-2A>G	90
CHD1L	chr1:146756048	c.1730G>A (p.Gly373Asp)	25
FANCG	chr9:35074472	c.1652_1655del (p.Tyr551Phefs*7)	50
MC1R	chr16:89986122	c.456C>A (p.Tyr152*)	52B
NEIL1	chr15:75641315	c.330_331insAGGC (p.Ala111Argfs*46)	43
NEK1	chr4:170428209	c.1687_1688del (p.Ala563Tyrfs*36)	17, 78A
NEK1	chr4:170398474	c.2235T>G (p.Asn648Lys)	89
NEK11	chr3:130828766	c.455+1G>A	68C, 68B
RHNO1	chr12:2997245	c.337C>T (p.Arg113*)	18
RHNO1	chr12:2994578	c.45_46delinsAG (p.Leu16Val)	70A, 70B
SPP1	chr4:88901197	c.94-1G>A	78A, 78B
CEP164	chr11:117244534	c.1220C>T (p.Ser407Phe)	53A, 53B
FAN1	chr15:31197015	c.149T>G (p.Met50Arg)	58A, 58B, 34
TONSL	chr8:145662005	c.1950C>G (p.Asp650Glu)	86
WRN	chr8:31012237	c.3785C>G (p.Thr1262Arg)	44

Table 3. 6. LOH was assessed for 27 variants in 29 tumors.

Modified from Smith et al., Cancer Letter, 2016.

Gene	Chr. Pos.	Variant (HGVS nomenclature)	Samples	LOH
AATF	chr17:35307578	c.158_159dup (p.Gly54Trpfs*157)	76A, 76B	N, -
BARD1	chr2:215595181	c.1935_1954dup (p.Glu652Valfs*69)	62	N
BRCA2	chr13:32913182	c.4691dupC (p.Thr1566Aspfs*9)	12	N
C17orf70	chr17:79518071	c.449G>A (p.Trp150*)	74	N
CDC6	chr17:38449789	c.743del (p.Gln248Argfs*16)	2	N
DNA2	chr10:70182188	c.2493-2A>G	64	N
FANCL	chr2:58386928	c.1096_1099dup (p.Thr367Asnfs*13)	47, 55B	N, -, -
MC1R	chr16:89986122	c.456C>A (p.Tyr152*)	52B	N[a]
MGMT	chr10:131565137	c.593G>A (p.Trp198*)	63A	Y
NEIL1	chr15:75641315	c.330_331insAGGC (p.Ala111Argfs*46)	43	N
NEK11	chr3:130828766	c.455+1G>A	68C, 68B	N[b]
NINL	chr20:25434092	c.4142_4143del (p.Ser1381Cysfs*17)	72	N[a]
PMS1	chr2:190719824	c.1826G>A (p.Trp609*)	13	Y (alt.)
RHNO1	chr12:2997158	c.250C>T (p.Arg84*)	43	Y
RHNO1	chr12:2997245	c.337C>T (p.Arg113*)	18	N
SMC2	chr9:106875703	c.1365_1366del (p.Arg456Thrfs*2)	73	N
TEX15	chr8:30700833	c.5699_5700del (p.Arg1900Asnfs*22)	24A	N
TONSL	chr8:145668147	c.490del (p.Leu164Serfs*72)	2	N
WDR48	chr3:39125749	c.1278_1279del (p.Gly427Aspfs*8)	72	Y[c]
AATF	chr17:35311130	c.755A>G (p.Asn252Ser)	32	N
CHD1L	chr1:146756048	c.1730G>A (p.Gly373Asp)	25	N
DCLRE1A	chr10:115602192	c.2575A>T (p.Ile859Phe)	24A, 81	N, -
ERCC6	chr10:50690906	c.1996C>T (p.Arg666Cys)	10	N
FAN1	chr15:31197015	c.149T>G (p.Met50Arg)	58A, 58B, 34	-, N[d], N
MLH3	chr14:75514503	c.1856A>T (p.Lys619Ile)	8	Y (alt.)
RHNO1	chr12:2994578	c.45_46delinsAG (p.Leu16Val)	70A, 70B	N, N
WRN	chr8:31012237	c.3785C>G (p.Thr1262Arg)	44	N

a Absence of LOH or somatic second hit in tumor exome sequencing data
b LOH assessed in tumor from PC-affected family member who was found to be a carrier
c LOH identified in tumor exome
d Absence of LOH or somatic second hit in tumor whole genome sequencing data

3.2.5. Top candidate genes

Next, we prioritized the 41 DNA repair genes with identified PTVs in them based on

1- If there is more than 1 kindred with a PTV in that gene

2- If the predicted pathogenic variant in the gene was segregated in at least one kindred

3- If LOH was found in the corresponding wild-type allele.

Based on these criteria we believed there are 17 genes with stronger genetic evidence supporting their roles as candidate novel PC predisposition genes (Table 3.7). Among these genes, FAN1, NEK1 and RHNO1 were our top three candidate genes as they harbor variants presented in three kindred and a variant segregated at least in two kindred. (Figures 3.2, 3.3 and 3.4)

Table 3. 7. Seventeen top candidate PC susceptibility genes

Modified from Smith et al., Cancer Letter, 2016.

Gene	Chr. Pos.	Variant (HGVS nomenclature)	Samples	>1 PTV	Segregation	LOH
AATF	chr17:35307578	c.158_159dup (p.Gly54Trpfs*157)	76A, 76B		Y	
AATF	chr17:35311130	c.755A>G (p.Asn252Ser)	32		Y	
BLM	chr15:91292792	c.298_299del (p.Gln100Glufs*42)	78B		Y	
CEP164	chr11:117282575	c.4228C>T (p.Gln1410*)	16			
CEP164	chr11:117244534	c.1220C>T (p.Ser407Phe)	53A, 53B		Y	
CHD1L	chr1:146742591	c.1086-2A>G	90		Y	
CHD1L	chr1:146756048	c.1730G>A (p.Gly577Asp)	25		Y	
FAN1	chr15:31214513	c.2128C>T (p.Arg710*)	42			
FAN1	chr15:31197015	c.149T>G (p.Met50Arg)	58A, 58B, 34		Yx2	
FANCG	chr9:35074472	c.1652_1655del (p.Tyr551Phefs*7)	50		Y	
FANCL	chr2:58386928	c.1096_1099dup (p.Thr367Asnfs*13)	47, 55B	Y		
MC1R	chr16:89985733	c.67C>T (p.Gln23*)	14	Y		
MC1R	chr16:89985750	c.86dup (p.Asn29Lysfs*14)	69	Y		
MC1R	chr16:89986122	c.456C>A (p.Tyr152*)	52B	Y	Y	
MC1R	chr16:89986522	c.862_864del (p.Ile288del)	89	Y		
MGMT	chr10:131565137	c.593G>A (p.Trp198*)	63A			Y
NEIL1	chr15:75641315	c.330_331insAGGC (p.Ala111Argfs*46)	43		Y	
NEK1	chr4:170428209	c.1687_1688del (p.Ala563Tyrfs*36)	17, 78	Y	Y	

Gene	Chr. Pos.	Variant (HGVS nomenclature)	Samples	>1 PTV	Segregation	LOH
NEK1	chr4:170398474	c.2235T>G (p.Asn745Lys)	89	Y	Y	
NEK11	chr3:130828766	c.455 + 1G>A	68C, 68B		Y	
RHNO1	chr12:2997158	c.250C>T (p.Arg84*)	43	Y		Y
RHNO1	chr12:2997245	c.337C>T (p.Arg113*)	18	Y	Y	
RHNO1	chr12:2994578	c.45_46delinsAG (p.Leu16Val)	70A, 70B	Y	Y	
SPP1	chr4:88901197	c.94-1G>A	78A, 78B		Y	
TONSL	chr8:145668147	c.490del (p.Leu164Serfs*72)	2			
TONSL	chr8:145660507	c.2899C>T (p.Arg967Cys)	3A			
TONSL	chr8:145662005	c.1950C>G (p.Asp650Glu)	86		Y	
WDR48	chr3:39125749	c.1278_1279del (p.Gly427Aspfs*8)	72			Y
WRN	chr8:30999118	c.3138 + 2T>G	51			
WRN	chr8:31012237	c.3785C>G (p.Thr1262Arg)	44		Y	

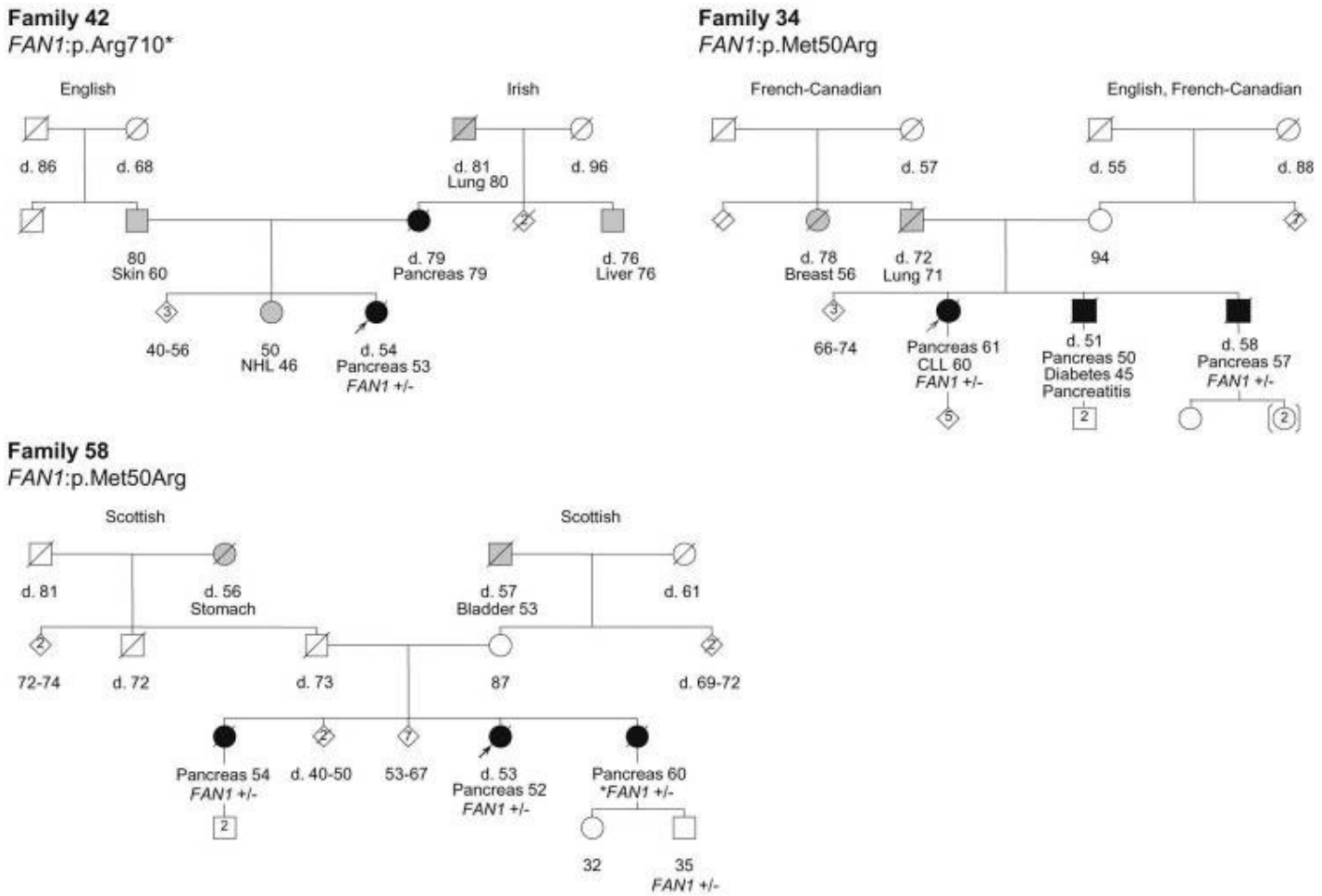


Figure 3. 2. Pedigrees of the families with FAN1 variants.

We analyzed all the cases in which germline DNA was available to be tested. +/- indicates heterozygous carrier status. +/+ indicates wild-type. Probands are indicated with an arrow. Individuals shaded in black are PC individuals, while individuals had other type of tumors shaded in grey. Ages of living family members, ages at diagnoses and the ages of death (d.) for deceased individuals are indicated in years. Other illnesses with ages in years at diagnosis (if known) are shown. NHL, non-Hodgkin's lymphoma; CLL, Chronic lymphocytic leukemia. Adapted from Smith et al., Cancer Letter, 2016.

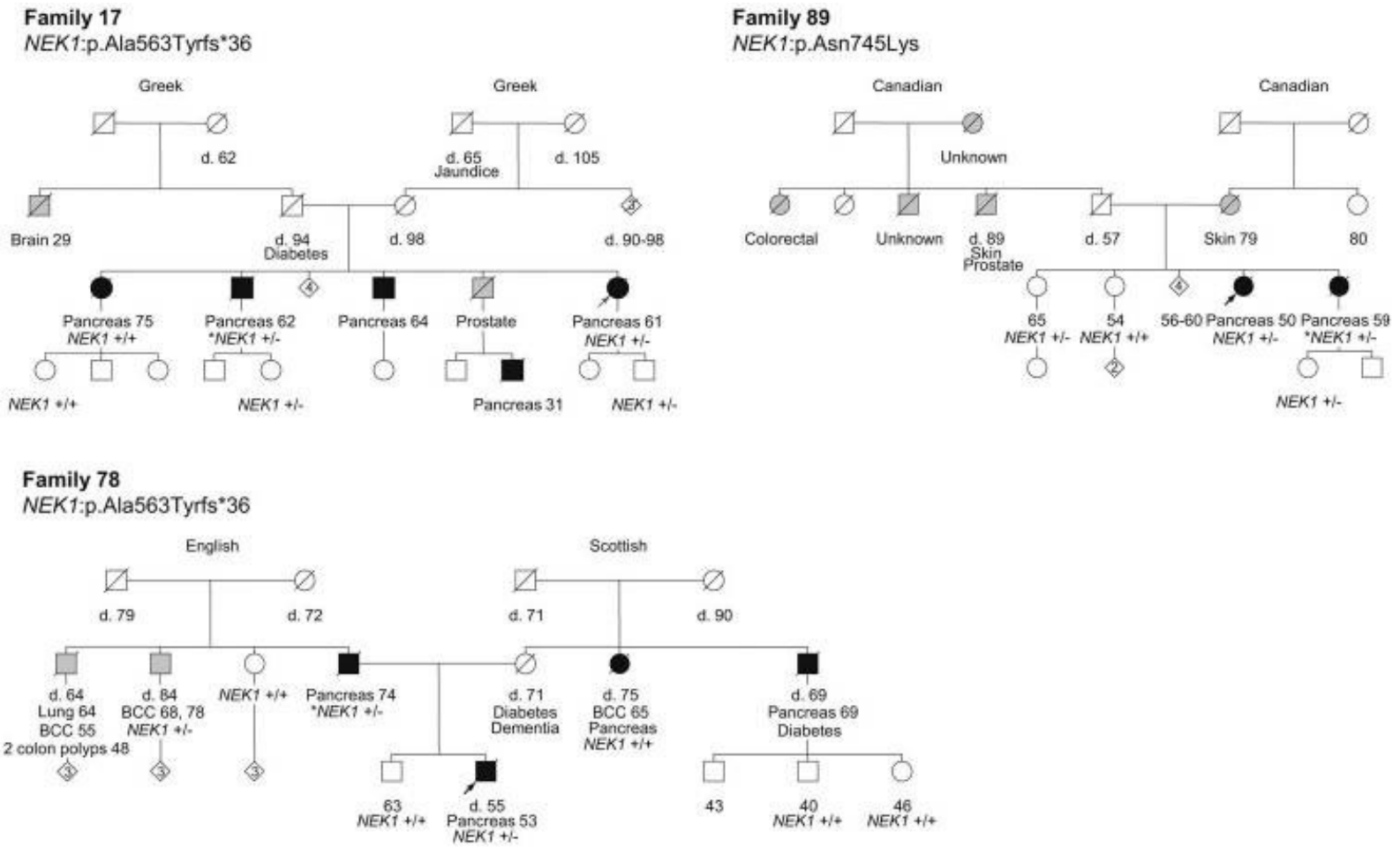


Figure 3. 3. Pedigrees of the families with NEK1 variants.

+/- indicates heterozygous carrier status. +/+ indicates wild-type. Probands are indicated with an arrow. Black indicates PC individuals, while grey indicates other type of tumors. BCC, basal cell carcinoma. Adapted from Smith et al., Cancer Letter, 2016

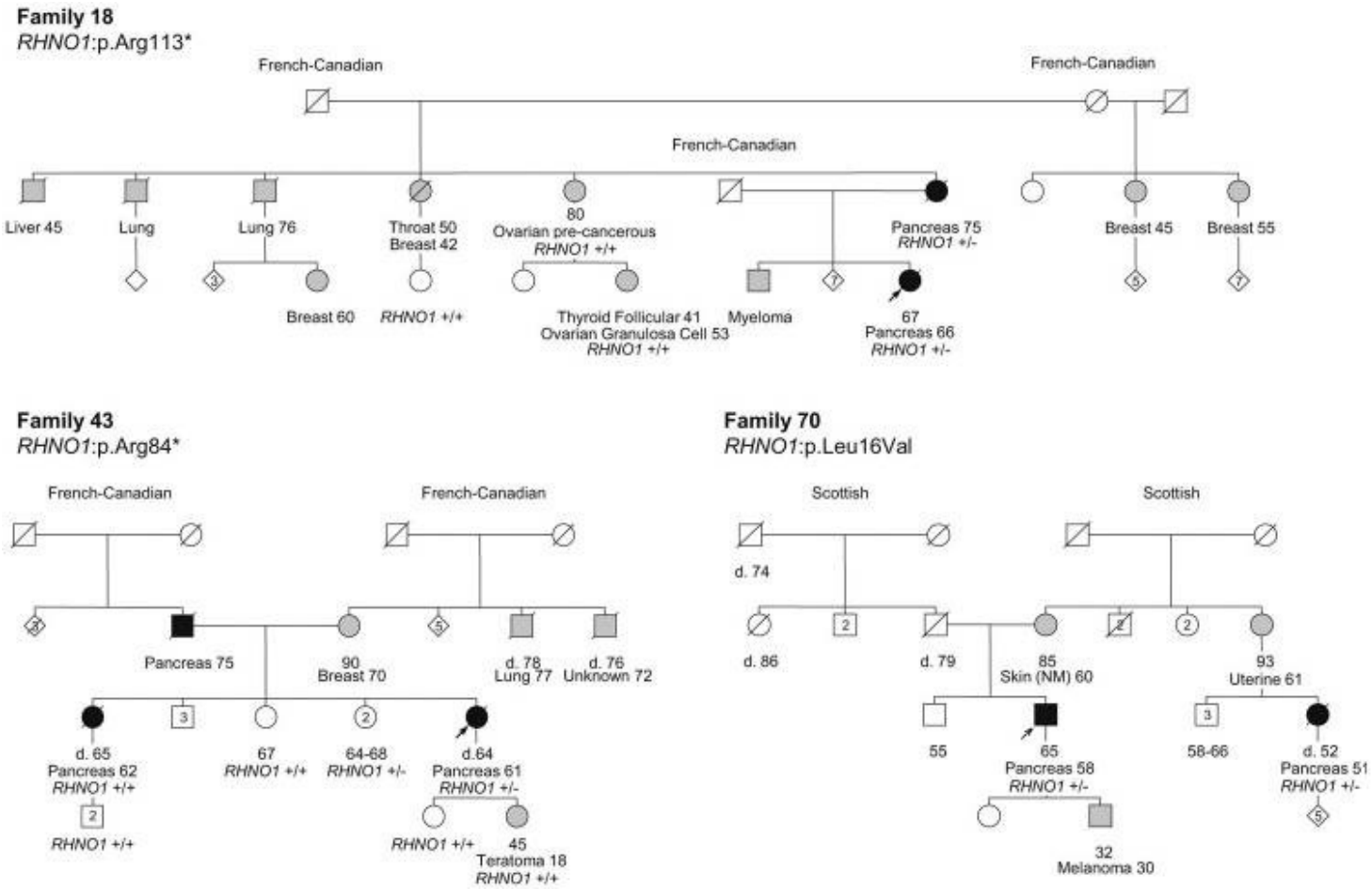


Figure 3. 4. Pedigrees of the families with RHNO1 variants.

+/- indicates heterozygous carrier status. +/+ indicates wild-type. Probands are indicated with an arrow. NM, non-melanoma. Adapted from Smith et al., Cancer Letter, 2016

3.3. Conclusion

In our study, we performed a large-scale NGS to identify novel genetic predisposition factors in hereditary PC. We performed WES on 109 selected cases with increased risk of genetic PC predisposition from 93 families. As we hypothesized that DNA repair genes can be involved in this particular cancer type, we performed a filter-based candidate gene approach to find new genes responsible for FPC. We found PTVs in 42 DNA repair genes among 37 families. One of these PTVs was in a known FPC susceptibility gene (BRCA2). This variant was missed in the screening phase but was detected by WES, showing the ability of our approach to identify causal variants. We not only looked for DNA repair genes in our approach but also considered mutations in known PC predisposition genes (for example CDKN2A, SPINK1 and PRSS1) that are not involved in DNA repair.

We further looked for the other variants in these 41 genes and performed segregation analysis and LOH analysis in order to prioritize our list of 41 candidate PC susceptibility genes. From this information, 17 genes ranked at the top of the list according to the following criteria:

1- genes with more than 1 family with a PTV
2- genes with segregation of a predicted-pathogenic variant in at least one kin
3- and/or genes with LOH associated with at least one predicted-pathogenic variant.

We propose FAN1, NEK1 and RHNO1 as the strongest candidates based on the available data to be further investigated and validated by additional families with PC.

In our study, we also found that rare germline protein-truncating DNA repair gene variants are common in PC. This could indicate that double haploinsufficiency may have a role in PC

development. There were also other noteworthy variants that can be potentially important in pancreatic cancer development. Most of these variants were located in genes implicated in other hereditary cancer syndromes. In particular FANCG , FANCL, POLQ, BLM , and BARD1 genes

In summary, our findings suggest that several novel DNA repair genes may have a role in hereditary PC. Previous linkage studies could not identify major causal loci for this disease and to date, 12 genes were reported to have role in predisposition to PC. Therefore, the remaining cause of familial PC may be due to several genes, with each gene accounting for only a small fraction of PC susceptibility.

Chapter 4: The Genomic landscape of triple negative breast cancer in neoadjuvant chemotherapy

Aguilar-Mahecha A, Alirezaie N, Bareke E, Przybytkowski E, Lan C, Lafleur J, Cavallone L, Salem M, Pelmus M, Aleynikova O, C. Greenwood, Lovato A, Nabavi S, Tonellato P, Kiu R, Ferrario C, Boileau JF, Robidoux A, Mihalciou C, Roy JA, Markus E, Seguin C, Discepola F, Sala S, Chabot C, Sirois I, Majewski J, Basik M

A version of chapter 4 is under preparation for publication. Author contributions are stated in the Contribution of the Authors section.

4.1. Introduction

Breast cancer is the second leading cause of cancer-related deaths and the most frequently diagnosed cancer among Canadian women. Resistance to chemotherapy is the underlying cause of most cancer fatalities. Moreover, administration of ineffective chemotherapeutic agents increases the probability of side effects and decreases the quality of life of many cancer patients, which further emphasizes the need to develop more efficient drugs and target them appropriately.

The cause of drug resistance in Triple Negative Breast Cancer (TNBC), which has a much higher proportion of all breast cancer mortality, is still unknown. Although some mechanisms, such as decreased cellular uptake, alter signaling pathways, dysregulated apoptosis, DNA repair, change in autophagy, paracrine effects, and changes in microRNA expression have been reported in publications, none of them has been validated in clinical studies. Therefore, we hypothesize that by using NGS approaches, we will be able to identify the novel genes associated with resistance.

4.2. Results

4.2.1. Clinical results

From 59 patients enrolled in this study, we were able to obtain pre-chemotherapy tumor samples from 54 of them. Although these patients received different types of chemotherapy (Table 4.1), most of them (47/54) were on anthracycline and taxane-based chemotherapy treatment. Seven patients received only taxane, while 42 patients received biphasic chemotherapy- either anthracycline followed by a taxane or the reverse. We further assessed the residual tumors by calculating the Residual Cancer Burden on a scale of 0-3 as described by Symmans et al (RCB-0 [no residual cancer or pathologic complete response], RCB-I [minimal], RCB-II [moderate], RCB-III [extensive residual]) (Symmans, et al. 2007)). Twenty-two of 47 patients had pathologic complete response (pCR) while 32 patients had residual tumors. Interestingly none of the pCR samples received taxane monotherapy. In patients who became resistant to chemotherapy, we tried to obtain DNA samples from matched pre-treatment and post-treatment tumors. However, in only 18 out these 32 patients we had enough pre-treatment tumor DNA and/or RNA extraction and the rest we could not obtain DNA because of the small size of the biopsies and the variations in tumor content.

Since the prognosis of RCB1 was shown to be almost identical to patients with RCB scores of zero (pCR) (Symmans, et al. 2007), we considered these patients in the same group for our analysis. By looking at the quality of the data, we had to exclude four pCR samples from further analysis because of low cellularity. We also excluded Neo4 from the resistant group after it was re-evaluated as non-TNBC.

Table 4. 1. Chemotherapy treatments and response in all samples

Category	Treatment	RCB0/1	RCB2/3	Total
1. Anthracycline–based and Taxane-based therapies	AC x 4 - T x (11-12)	6	8	14
	AC x 1 - TC x 4	0	1	1
	AC x 4 - T x (3-8)	5	4	9
	TAC x (3-6)	2	3	5
	EC x 4 - T x 12	1	1	2
	FEC x 3 - D x 3	2	1	3
	FEC x 3 - T x (9-12)	0	4	4
	TC x 4 - EC x 4	1	0	1
	TC x 3 - AC x 1	1	0	1
	T x 12 - AC x 4	2	2	4
	TC x 4 - FEC x (2-4)	2	1	3
Total for each response group		22 (46.8)	25 (53.1%)	47
2. Taxane-based only	T x (9-12)	0	4	4
	TC x (2-4)	0	3	3
Total for each response group		0 (0%)	7 (100%)	7

A, Adryamicin; C, Cyclophosphamide; T, Taxol; F, 5-fluouracil; E, Epirubicin; D, dose dense

We evaluated tumor size before, at midpoint of the treatment and after therapy before surgery in the patients. The response of the tumor was not the same in all tumors (Figure 4.1). In four patients (Neo07, Neo30, Neo27 and Neo50) the tumor did not respond to treatment according to RECIST criteria (>25% decrease in tumor size); therefore, these samples were labeled "non-responders". In one of these cases (Neo07), the tumors even grew while on paclitaxel monotherapy. Seven tumors, labeled "responders", showed significant response to chemotherapy.

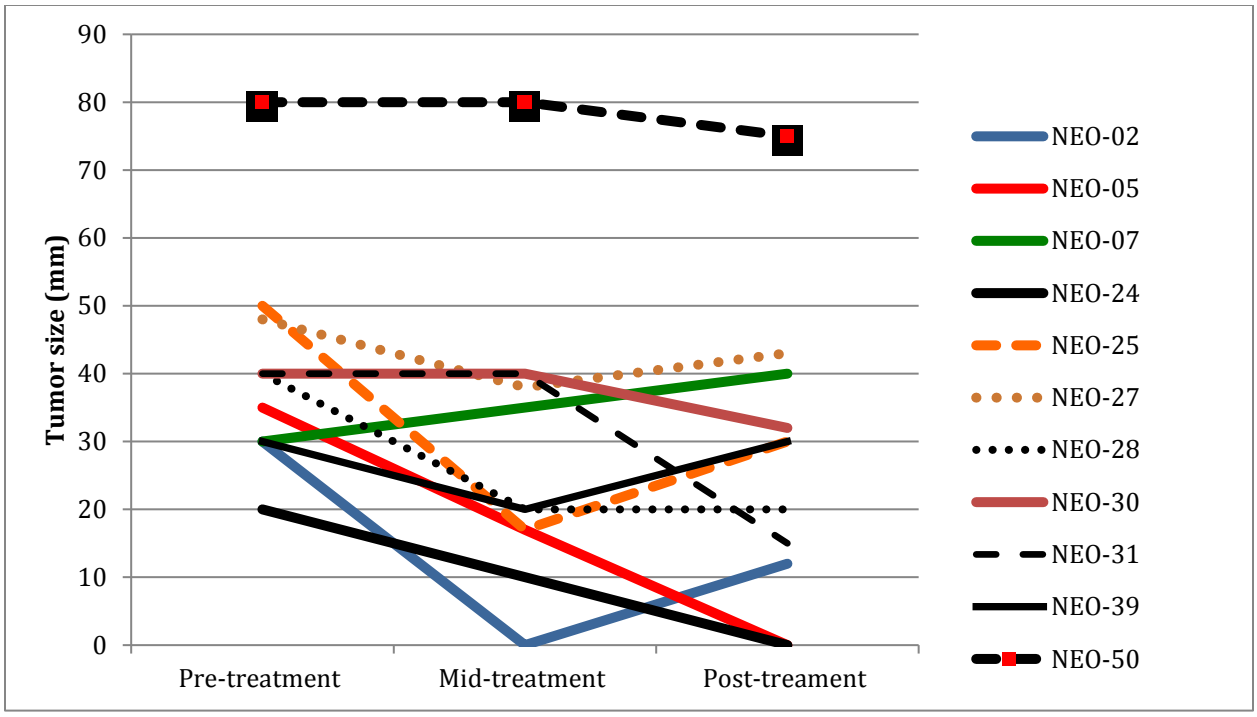


Figure 4. 1. Clinical evaluation of tumor size before, at midpoint of the treatment and after therapy before surgery

4.2.2. Whole exome sequencing results

Whole exome sequencing (WES) data was generated on samples from 25 patients. For our analysis, we included only somatic nonsynonymous SNVs, stopgains, frameshift indels, and non-frameshift indels not occurring in repetitive regions as discussed in method section. There were 11 pairs of pre- and post-treatment samples. In four pre-treatment samples from RCB2/3 tumors, we could not get adequate genomic data from the post-treatment biopsies. In two cases, sufficient post-treatment tissue was only available. Seven pre-treatment samples did not have residual tumor (RCB0 or RCB1) .We also had four samples from the metastatic lesions in three patients (Table 4.2).

We explored whether there was any shared mutation between samples. Germline BRCA2 mutations were seen in six patients, while germline BRCA1 mutations were detected in three patients. The most commonly mutated gene was TP53 in 21 of 25 tumors. NOTCH1 and MAGI2 genes mutated in four tumors, and five genes (RB1, PLXNC1, HUWE1, PDZRN4, and IGSF10) mutated in three tumors. No other gene mutated in more than two patient samples. On average, there were 80 somatic mutations detected in any tumor. The chemo-resistant tumors (RCB2/3 group) had more variants (average 89/tumor) than the chemo-sensitive (RCB0/1 group) tumors (average 59/tumor). Less than 5 percent of detected variants were presented in the COSMIC database; therefore, most of these variants were novel. These findings emphasize the genomic heterogeneity of TNBC tumors.

Table 4. 2. Number of remaining samples in each group

	Lymphocyte /normal tissue	Pre - chemotherapy	Post - chemotherapy	Metastasis
pCR	5	5	-	-
I	3	3	-	-
II, III	17	15	14	4
Total	25	23	14	4

In the next step, we performed three types of analyses to obtain a complete knowledge of tumor resistance mechanisms: an analysis of post- versus pre-treatment matched pairs, an analysis of the post-chemotherapy genomic landscape, and an analysis comparing all pre-treatment samples from “chemo-resistant” to “chemo-sensitive” tumors.

4.2.2.1 WES analysis of pre- and post-chemotherapy samples

We compared post-treatment samples with its respective pre-treatment one to find if possible changes had occurred during treatment. In total, 1148 somatic variants were observed in all 11 pairs of matched tumors. The number of variants varied in different tumors. Figure 4.2 shows the number of somatic variants in pre- and post-chemotherapy and the number of variants present just in pre or post or both. The great majority of variants (average 78%) were present in both pre- and post-treatment samples, although some variants were just in the pre-treatment samples (12%) and others (10%) in the post-treatment tumor samples.

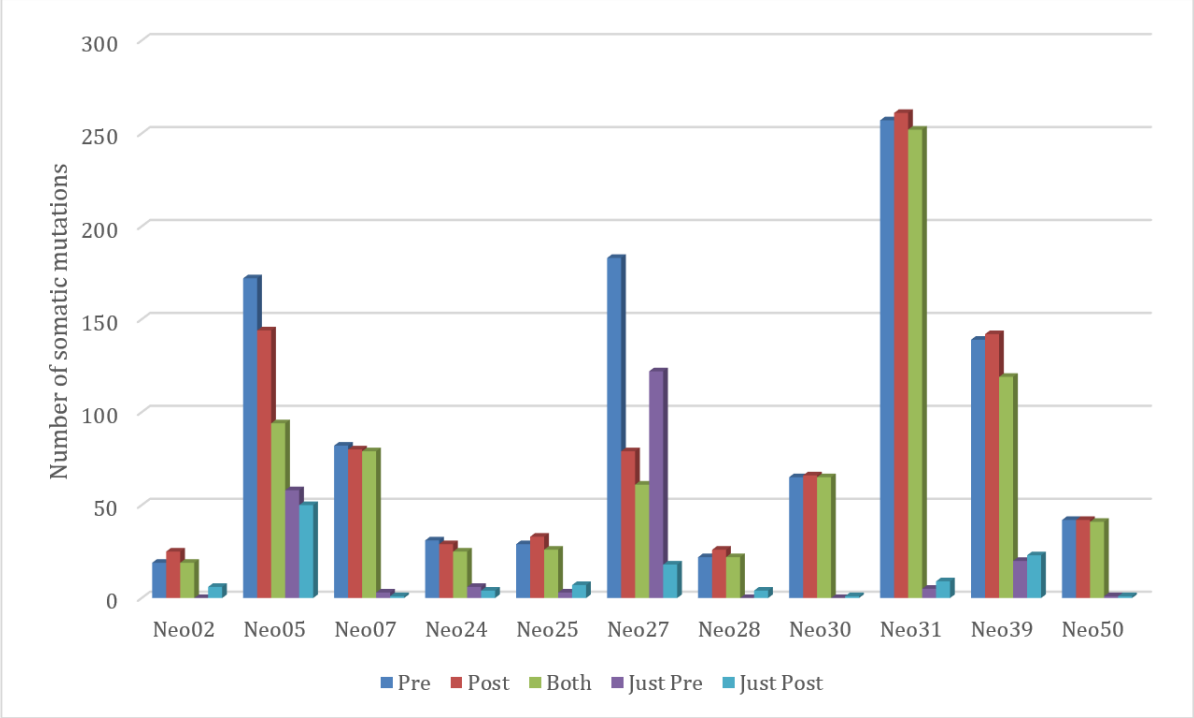


Figure 4. 2. Comparing number of somatic variants in pre and post-chemotherapy samples

In four tumors, at least 95% of the variants were conserved, suggesting little genomic change from pre- to post-treatment. This trend was in line with the degree of clinical tumor response in these four patients; none of them showed any initial response to treatment (non-responders). In one case (Neo27) a large proportion of variants in pre-chemo was lost, but this patient did not have a major response to chemotherapy (Figure 4.3).

In order to validate these results, we selected 3-5 variants per tumor pair to be tested by digital droplet PCR technology in the same tumor DNA sample used for sequencing. Except in one tumor (Neo27), all the tested WES variants were detected by ddPCR with almost similar AF. In Neo27, ddPCR did not detect 67% of SNVs (10 of 15 selected variants) in the pre-chemo tumor. This finding suggests that the majority of these variants were due to sequence artifacts. Therefore, this sample was excluded from further analysis.

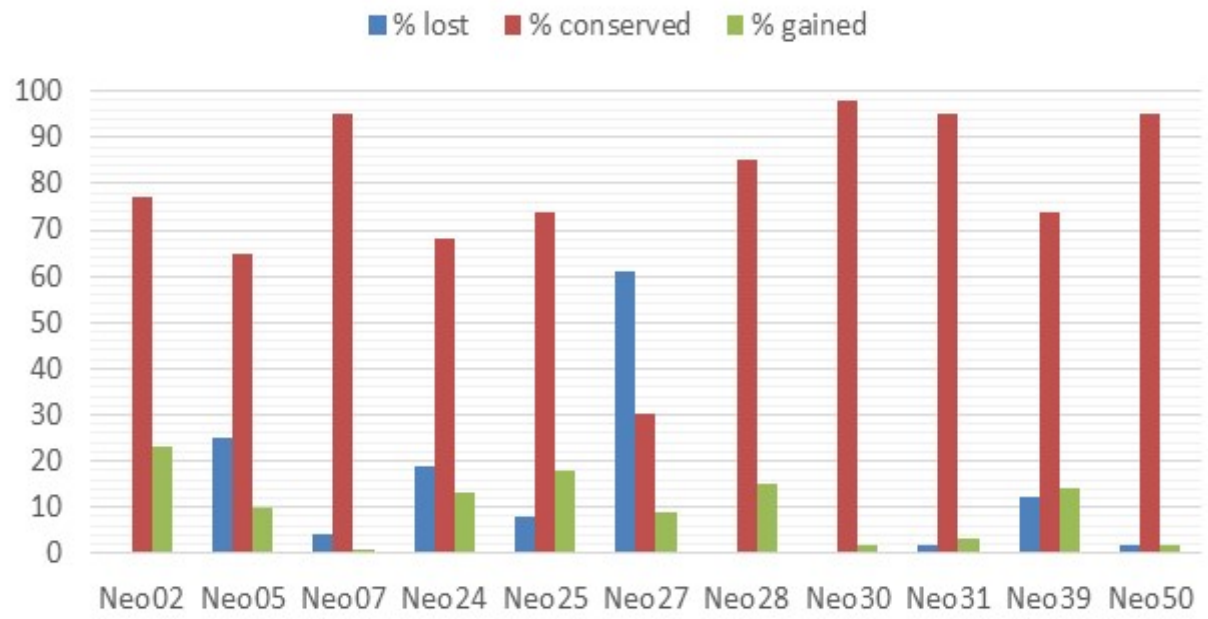


Figure 4. 3. Proportion of variants conserved, only detected in post-chemo sample (gained) or only detected in pre-chemo sample (lost) in each matched pre/post samples

We then asked if the "gained" or "lost" mutations were acquired in response to chemotherapy, or these variants existed in very low frequencies before chemotherapy. To explore the answer to this question, we selected seven different variants in four tumor pairs, which showed either a "gain" or a "loss" from pre- to post-treatment to be further validated by digital droplet PCR technology (Table 4.3). Three of these variants were detected by ddPCR in both pre- and post- blood samples. This finding suggests that de novo "gained" mutations or "lost" variants may have been present at low frequencies in subclonal tumors and may have not been sampled because of intra-tumoral heterogeneity.

Table 4. 3. Comparison of tumor and plasma variant allele frequencies

Sample	Gene	Variant	Position	Ref	Alt	Protein change	Tumor VAF		Plasma VAF (%ctDNA)		
							Pre-chemo	Post-chemo	Pre-chemo	Mid-chemo	Post-chemo
Neo27	ROBO3	non synonymous SNV	124744718	G	T	p.W662C	0	0.69	1.48	1.79	0.27
Neo27	NOL7	non synonymous SNV	13615603	C	G	p.R5G	0	0.55	0	0	0.78
Neo02	UTS2R	non synonymous SNV	80332825	C	G	p.R209G	0	0.4	5.14	0.51	3.23
Neo05	TRPM3	non synonymous SNV	73168179	G	T	p.L916M	0	0.99	0	0	0
Neo05	ROBO2	non synonymous SNV	77147279	C	T	p.T59M	0	0.97	0	0	0
Neo05	HDAC 9	frameshift deletion	18788734	AG	A	p.E670fs	0.7	0	0.68	0.05	3.17
Neo39	RASA1	non synonymous SNV	86633905	G	C	p.E338D	0.72	0	0	0	0

4.2.2.2. Post-chemotherapy genomic landscape

To provide a picture of the mutation landscape of post-chemotherapy samples, we retrieved all the mutated genes present in the 14 post-chemotherapy tumor samples with a minimum allele frequency of 0.3 (n=588 genes) . This threshold was chosen to favor more clonal variants. Besides TP53 and RB1, only 14 of these genes were mutated in more than one sample (Table 4.4). Among 16 recurrent somatic mutations in post-chemo, a variant in the ROBO2 gene showed the highest gain in AF (0.97), while another variant in the same gene was associated with an AF of 0.55 in the post-chemo Neo38 sample (no pre-chemo sample was available for this tumor).

We performed pathway analysis on these 588 genes to find any pathway enriched in these samples. Our results did not show any pathway enriched with a p-value and a q-value less than 0.05. Of these 588 gene variants, we identified 67 genes with stopgain or frameshift indel variants, including frameshifts affecting TP53, RB1 and PTEN (Table 4.5). Pathway analysis by DAVID for these 67 genes showed enrichment in "regulation of cell cycle" (FDR = 0.034). The following genes are in this group: TP53, RB1, TACC3, PTEN, MOV10L1, HERC2, INSR, ZZEF1 and CDK13. Interestingly, three genes of the dynein family of microtubule-associated motor proteins (DNAH2, DNAH3, and DNAH5) showed stopgain mutations in different post-chemotherapy residual tumor samples.

Table 4. 4. Recurrently mutated genes in post-chemo tumors

Gene	Position	Variation	Protein Change	Tumor	Post chemo AFs
GAPVD1	Chr9:128061231	nonsynonymous SNV	p.H11D	Neo05	0.92
GAPVD1	Chr9:128092405	nonsynonymous SNV	p.S673L	Neo38	0.51
GPR158	Chr10:25839971	nonsynonymous SNV	p.G491S	Neo39	0.70
GPR158	Chr10:25887599	nonsynonymous SNV	p.T1015I	Neo50	0.50
HIVEP1	Chr6:12123407	stopgain	p.Q1127X	Neo27	0.33
HIVEP1	Chr6:12122228	stopgain	p.Q734X	Neo39	0.30
MAGI2	Chr7:78256520	nonsynonymous SNV	p.G152R	Neo05	0.40
MAGI2	Chr7:79082626	nonsynonymous SNV	p.S4N	Neo30	0.39
MCM3AP	Chr21:47685317	nonsynonymous SNV	p.P1051Q	Neo05	0.66
MCM3AP	Chr21:47692523	nonsynonymous SNV	p.F806Y	Neo39	0.35
MEF2D	Chr1:156446983	nonsynonymous SNV	p.V226I	Neo25	0.80
MEF2D	Chr1:156449594	splicing-extended	NA	Neo50	0.44
NOTCH1	Chr9:139410015	nonsynonymous SNV	p.S608C	Neo07	0.96
NOTCH1	Chr9:139399537	nonsynonymous SNV	p.C1536R	Neo24	0.50
OSBPL9	Chr1:52238388	nonsynonymous SNV	p.L210P	Neo05	0.47
OSBPL9	Chr1:52253122	stopgain	p.R522X	Neo31	0.40
PBXIP1	Chr1:154919168	nonsynonymous SNV	p.E328Q	Neo05	0.44
PBXIP1	Chr1:154920756	nonsynonymous SNV	p.E166K	Neo24	0.39
PHF21B	Chr22:45312350	nonsynonymous SNV	p.A113G	Neo05	0.72
PHF21B	Chr22:45279150	nonsynonymous SNV	p.R417H	Neo17	0.64
PLXNC1	Chr21:94543548	nonsynonymous SNV	p.I267M	Neo05	0.50
PLXNC1	Chr21:94543039	nonsynonymous SNV	p.R98W	Neo30	0.34
ROBO2	Chr3:77612365	nonsynonymous SNV	p.G523R	Neo05	0.97
ROBO2	Chr3:77147279	nonsynonymous SNV	p.T59M	Neo38	0.55
SPTBN1	Chr2:54858620	nonsynonymous SNV	p.A1133S	Neo07	0.42
SPTBN1	Chr2:54853291	nonsynonymous SNV	p.Q509K	Neo39	0.36
WDR90	Chr16:711894	nonsynonymous SNV	p.R1290C	Neo31	0.44
WDR90	Chr16:711414	nonsynonymous SNV	p.G1196S	Neo17	0.33

Table 4. 5. Stopgains and frameshift indels in post-chemo samples

	Variation	Protein Change	Position	Ref	Alt	Sample	AF
ACO1	stopgain	p.Y759X	32440492	C	A	Neo27	0.51
ACSM1	stopgain	p.Y461X	20638555	G	T	Neo30	0.37
ADD1	frameshift deletion	p.E281fs	2900008	AAGAGGAAAAAGTTTTGATTCAGA	AAGA	Neo39	0.54
AKAP6	stopgain	p.E1877X	33292648	G	T	Neo05	0.95
ANKLE1	frameshift insertion	p.X652delinsX	17397522	T	TGA	Neo25	0.48
ARHGAP22	stopgain	p.W4X	49701523	C	T	Neo05	0.62
ARMC9	stopgain	p.G543X	2.32E+08	G	T	Neo05	0.85
ATG7	frameshift deletion	p.I565fs	11406143	ATTGC	A	Neo39	0.31
BBS1	frameshift deletion	p.P61fs	66281895	GGCCCTGGTGGGCAGCAGCCC	GGCCC	Neo27	0.34
CCR5	frameshift insertion	p.V154fs	46414853	GT	GTT	Neo30	0.32
CD2AP	stopgain	p.Y548X	47576870	C	A	Neo07	0.49
CDK13	stopgain	p.R626X	40037097	C	T	Neo31	0.59
CHD1	frameshift deletion	p.K1502fs	98195690	GTTTTTT	GTTTTT	Neo39	0.54
CPA6	stopgain	p.R184X	68419108	G	A	Neo31	0.32

	Variation	Protein Change	Position	Ref	Alt	Sample	AF
CR2	frameshift deletion	p.N239fs	2.08E+08	AC	A	Neo07	0.96
DDX51	frameshift deletion	p.R67fs	1.33E+08	GCCGTC	GC	Neo39	0.40
DENND4B	frameshift deletion	p.L1065fs	1.54E+08	GGGAGTGGCGGGAAGGAGTG	GGGAGTG	Neo05	0.70
DHDH	frameshift deletion	p.F329fs	49448167	ACCTTCCC	ACC	Neo07	0.97
DHX36	stopgain	p.Y66X	1.54E+08	G	T	Neo05	0.47
DHX37	stopgain	p.S177X	1.25E+08	G	T	Neo30	0.89
DNAH2	stopgain	p.R3387X	7721017	C	T	Neo39	0.44
DNAH3	stopgain	p.E1216X	21073877	C	A	Neo17	0.32
DNAH5	stopgain	p.Y3465X	13758979	A	C	Neo07	0.38
DOCK9	frameshift deletion	p.L1209fs	99512732	CA	C	Neo30	0.40
ETV3	frameshift deletion	p.E402fs	1.57E+08	AGTGCCCTCTTCTTGAGTGTGC	AGTGC	Neo05	0.75
FAM111A	frameshift insertion	p.Y263fs	58919923	GCAGAT	GCAGATTCTTTCAGAT	Neo39	1.00
FLT3	frameshift deletion	p.V36fs	28644682	TAAAA	TAAA	Neo31	0.32
FMN2	frameshift deletion	p.P991fs	2.4E+08	TCCCCC	TCCCC	Neo05	0.43
GNA13	stopgain	p.R165X	63010731	G	A	Neo07	0.97
GON4L	stopgain	p.R1639X	1.56E+08	G	A	Neo28	0.49

	Variation	Protein Change	Position	Ref	Alt	Sample	AF
GPRASP1	frameshift deletion	p.R330fs	1.02E+08	CAG	C	Neo39	0.48
GPRASP2	frameshift deletion	p.G622fs	1.02E+08	TGGG	TGG	Neo39	0.50
HCLS1	stopgain	p.E326delinsX	1.21E+08	C	CAGGCCCAGGCTA	Neo07	0.78
HEATR3	stopgain	p.R11X	50100073	C	T	Neo31	0.41
HERC2	stopgain	p.Q4424X	28366494	G	A	Neo07	0.44
HIVEP1	stopgain	p.Q1127X	12123407	C	T	Neo27	0.33
HIVEP1	stopgain	p.Q734X	12122228	C	T	Neo39	0.30
HSPB1	stopgain	p.E108X	75932351	G	T	Neo24	0.34
INSR	stopgain	p.Y319X	7184344	G	T	Neo05	0.56
KCNA2	stopgain	p.R65X	1.11E+08	G	A	Neo07	0.46
KCNF1	stopgain	p.E213X	11053189	G	T	Neo39	0.33
KDM5A	stopgain	p.G474X	443477	C	A	Neo05	0.45
KIAA2022	stopgain	p.E300X	73963494	C	A	Neo39	0.47
MGA	stopgain	p.R618X	41989060	C	T	Neo38	0.56
MOV10L1	frameshift deletion	p.T39fs	50530443	GAAAA	GAAA	Neo39	0.37
NAA40	stoploss	p.X217W	63721951	A	G	Neo38	0.36
ORMDL1	stopgain	p.E73X	1.91E+08	C	A	Neo39	0.49
OSBPL9	stopgain	p.R522X	52253122	C	T	Neo31	0.40
PARD3	stopgain	p.W735X	34630588	C	T	Neo07	0.46
PHKG1	frameshift deletion	p.G101fs	56151394	TCCCC	TCCC	Neo05	0.50

	Variation	Protein Change	Position	Ref	Alt	Sample	AF
PKP3	stopgain	p.E551X	400619	G	T	Neo17	0.67
PTEN	frameshift deletion	p.C71fs	89690802	GTT	GT	Neo31	0.48
PTPRD	stopgain	p.E538X	8507336	C	A	Neo05	0.37
RAB8A	frameshift deletion	p.Q183fs	16243037	CA	C	Neo17	0.34
RB1	stopgain	p.G442X	48951162	G	T	Neo07	0.92
RB1	stopgain	p.R358X	48942685	C	T	Neo38	0.83
SH2D4A	stopgain	p.R209X	19221636	C	T	Neo31	0.52
SMAD9	frameshift deletion	p.P333fs	37427678	AGCCGCTGGGGATCTTGCAGACGGTAGCTG	AG	Neo50	0.37
SMTN	stopgain	p.E82X	31483975	G	T	Neo07	0.88
TACC3	frameshift insertion	p.G219fs	1729786	A	AGCCGAGGAGGAAT	Neo25	0.37
TCEAL3	frameshift deletion	p.P101fs	1.03E+08	GCCC	GCC	Neo31	0.33
TP53	stopgain	p.R81X	7578212	G	A	Neo02	0.81
TP53	stopgain	p.R81X	7578212	G	A	Neo25	0.40
TP53	frameshift deletion	p.I200fs	7574029	CGG	CG	Neo30	0.99
TP53	stopgain	p.W14X	7579528	C	T	Neo50	0.63

	Variation	Protein Change	Position	Ref	Alt	Sample	AF
TP53	frameshift deletion	NA	7578505	GGGCAGGTCTTGGCCAGTTGGCAAAACATCTTGT	G	Neo17	0.97
TRIM46	frameshift deletion	p.T399fs	1.55E+08	GCACAC	GCAC	Neo05	0.87
TRIM5	stopgain	p.C55X	5701243	G	T	Neo17	0.40
TTBK1	frameshift deletion	p.R1003fs	43251479	GCCCCCC	GCCCCC	Neo31	0.31
USP49	frameshift insertion	p.C632fs	41766440	CCAC	CCACCACAC	Neo28	0.68
ZCCHC2	stopgain	p.R429X	60217665	C	T	Neo31	0.34
ZZEF1	frameshift deletion	p.G558fs	4005596	AGACCCACCACTCACC	A	Neo25	0.34

4.2.2.3: Mutated genes associated with response to chemotherapy

In the next step, to determine if mutations in the pre-chemotherapy samples could predict pathological complete response to chemotherapy, we compared somatic mutations in pre-chemotherapy samples of chemo-sensitive tumors with chemo-resistant one. We picked all potentially deleterious variants detected at >0.3 MAF in 15 pre-treatment chemo-resistant tumor samples (RCB2/3), that none of these genes showed somatic variants in the chemo-sensitive tumors (RCB0/1) (675 variants). Only 15 genes had variants in more than one tumor and none was mutated in more than two tumors (Table 4.6). Somatic variants in BRCA1 (in Neo32), and in Tubulins (TUBE1, TUA1B, TUBA3E) were among those 675 variants. These data indicate the genomic heterogeneity of TNBCs.

We also compiled a list of "chemo-sensitivity-associated gene variants" from the genes mutated at AF>0.3 uniquely in the eight pre-treatment tumors (n=263 genes) in RCB0/1 group and never mutated in resistant tumors. COL1A2 and PRDM15 were the only two recurrent genes in this list.

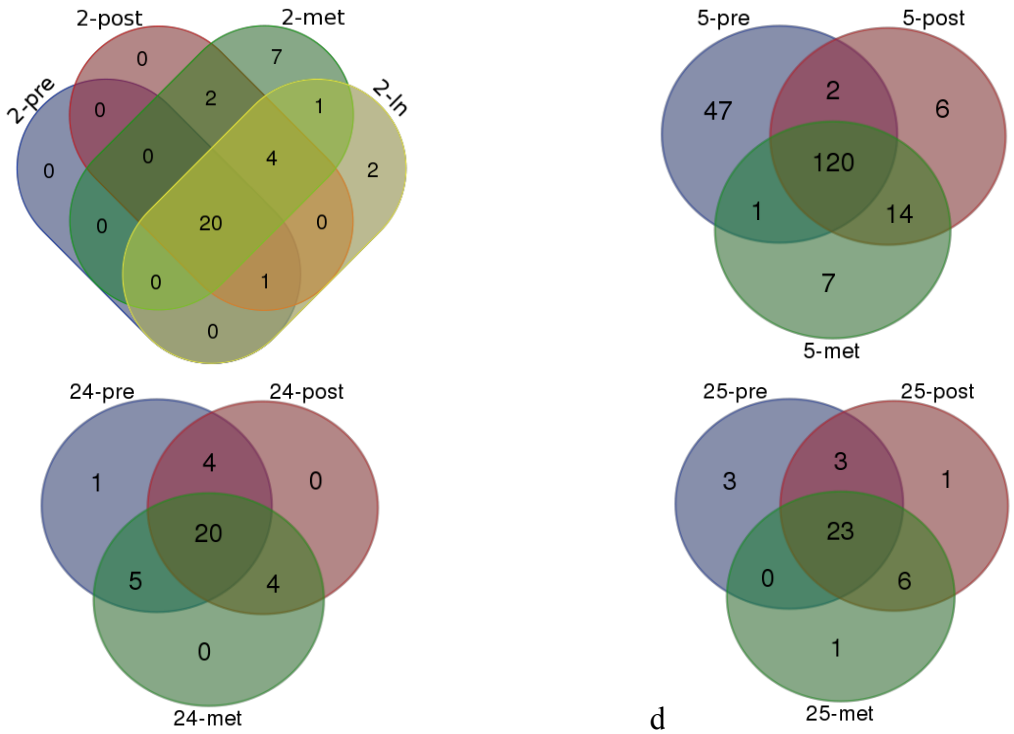
Table 4. 6. Recurrently mutated genes in pre-chemotherapy samples of non-responsive tumors

Gene	Position	Variation	Ref	Alt	Tumor	PreC AF
BRD3	136918572	frameshift deletion	CGGGGG	CGGGG	Neo31	0.48
BRD3	136905238	nonsynonymous SNV	C	G	Neo05	0.55
CHD5	6206925	nonsynonymous SNV	G	C	Neo27	0.71
CHD5	6195354	nonsynonymous SNV	C	T	Neo31	0.37
DLEC1	38139341	nonsynonymous SNV	C	G	Neo50	0.47
DLEC1	38081073	nonsynonymous SNV	G	C	Neo44	0.69
GPR158	25839971	nonsynonymous SNV	G	A	Neo39	0.85
GPR158	25887599	nonsynonymous SNV	C	T	Neo50	0.49
HCN1	45262787	nonsynonymous SNV	C	T	Neo30	0.46
HCN1	45696113	nonsynonymous SNV	G	A	Neo44	0.42
HOXD11	176972330	nonsynonymous SNV	A	G	Neo31	0.45
HOXD11	176973661	nonsynonymous SNV	C	T	Neo44	0.41
MEF2D	156446983	nonsynonymous SNV	C	T	Neo25	0.68
MEF2D	156449594	splicing-extended	C	T	Neo50	0.39
NPAP1	24921477	nonsynonymous SNV	G	A	Neo05	0.45
NPAP1	24923722	nonsynonymous SNV	C	A	Neo39	0.32
ROBO1	78656056	nonsynonymous SNV	C	A	Neo39	0.37
ROBO1	78656003	nonsynonymous SNV	C	T	Neo32	0.36
SAAL1	18127567	nonframeshift insertion	G	GCGC	Neo24	0.65
SAAL1	18127572	nonframeshift insertion	G	GGCC	Neo32	0.64
SCAF4	33074214	nonsynonymous SNV	C	T	Neo39	0.33
SCAF4	33064738	nonsynonymous SNV	C	G	Neo32	0.31
SLC16A2	73744409	nonsynonymous SNV	A	C	Neo39	0.52
SLC16A2	73751248	nonsynonymous SNV	C	T	Neo42	.42
SMARCA2	2116002	nonsynonymous SNV	C	T	Neo28	0.55
SMARCA2	2073595	nonsynonymous SNV	A	G	Neo44	0.48
SPTBN1	54858620	nonsynonymous SNV	G	T	Neo07	0.46
SPTBN1	54853291	nonsynonymous SNV	C	A	Neo39	0.46
TTC3	38538213	nonsynonymous SNV	G	C	Neo05	0.49
TTC3	38533166	splicing-extended	G	C	Neo25	0.32

PreC AF, pre-chemotherapy allele frequency

4.2.2.4. Analysis of metastasis samples

In four patients, we obtained frozen tumor samples from metastatic lesions (regional lymph nodes, skin, and liver). Somatic mutations detected in the metastatic lesions were compared with its respective pre-chemotherapy and post-chemotherapy samples (Figure 4.4). We found that the post-chemotherapy samples contained an average of 89% (79-97%) of the variants observed in the metastatic samples, compared to 62% (58-86%) of the pre- chemotherapy samples. Moreover, 91% of variants present in the metastatic tumors were detected in the post-chemotherapy tumor samples, compared with 85% in the pre-chemotherapy samples. Even though the sample size was limited, these findings suggest that the post-chemotherapy sample is a better indicator of the metastatic genotype than the pre-chemotherapy sample.



d

Figure 4. 4. Number of somatic variants shared between pre, post and metastatic samples in each patient.

4.2.3. Copy number alterations

To find changes in DNA copy number variation during chemotherapy, we performed array CGH on DNA extracted from the pre-treatment and post-treatment biopsies. After applying array CGH in 9 pairs of matched pre and post-treatment tumor samples, we found a region on chromosome 8q (9116962735 – 117887105) that was highly amplified (>2 fold), in both pre and post sample of 5 "drug-resistant" tumors. Three of these five tumors were responded very poorly to chemotherapy (Neo07, Neo27 and Neo50). This region contains three genes: RAD21, UTP23 and EIF3. MYC was also amplified >2-fold in two of nine tumors and there was no other region that amplified in more than two tumors.

In order to find if these copy number changes were functional or not, we combined these data with gene expression changes detected by RNAseq on the nine pairs. We found the mean RNA levels of the 3 genes mentioned above (RAD21, UTP23 and EIF3H) were at least 3 fold higher in the 5 tumors with amplified segments in both pre-chemotherapy and post-chemotherapy samples compared to the 4 tumors in which it was not. This finding indicates the functionality of this amplicon in these tumors.

We then performed copy number variations analysis on our whole exome data (7 patients with RCB0/1 and 15 patients with RCB2/3). Five regions with at least 1 MB size were differentially amplified between RCB0/1 and RCB2/3 tumors (Table4.7). The above-mentioned chromosome (8q23) was the only differentially amplified region in RCB2/3 tumors and was amplified in eight of 15 RCB2/3 and none of RCB0/1 tumors (Figure 4.5). We also found other regions (on chromosomes 1, 3 and 6) that were amplified only in RCB0/1 tumors (Table4.7).

Table 4. 7. Differentially amplified fragments in RCB0/1 vs RCB2/3 tumors by WES of 22 tumors

Chromosome	Genomic Location	Event	Freq. in RCB2/3 (%)	Freq. in RCB0/1 (%)	p-value	No. Genes	Gene Symbols
chr1	242,342,625-243,504,302	CN Gain	0	43	0.023	4	PLD5, CEP170, SDCCAG8
chr1	244,473,418-247,420,078	CN Gain	0	43	0.023	26	C1orf100, ADSS, CATSPERE, DESI2, COX20, HNRNPU, EFCAB2, KIF26B, SMYD3, TFB2M, CNST, SCCPDH, AHCTF1, ZNF695, ZNF670-ZNF695, ZNF670, ZNF669, FLJ39095, C1orf229, ZNF124, MIR3916, VN1R5
chr3	170,723,073-172,852,274	CN Gain	0	43	0.023	13	SLC2A2, MIR569, TNIK, PLD1, TMEM212, TMEM212-AS1, FNDC3B, GHSR, TNFSF10, NCEH1, ECT2, SPATA16
chr6	46,374,581-47,851,579	CN Gain	0	43	0.023	18	RCAN2, CYP39A1, SLC25A27, TDRD6, PLA2G7, ANKRD66, MEP1A, ADGRF5, , ADGRF1, TNFRSF21, CD2AP, ADGRF2, ADGRF4, OPN5, PTCHD4
chr8	117,514,050-119,774,352	CN Gain	53	0	0.022	13	EIF3H, UTP23, RAD21, MIR3610, RAD21-AS1, AARD, SLC30A8, MED30, EXT1, SAMD12, SAMD12-AS1

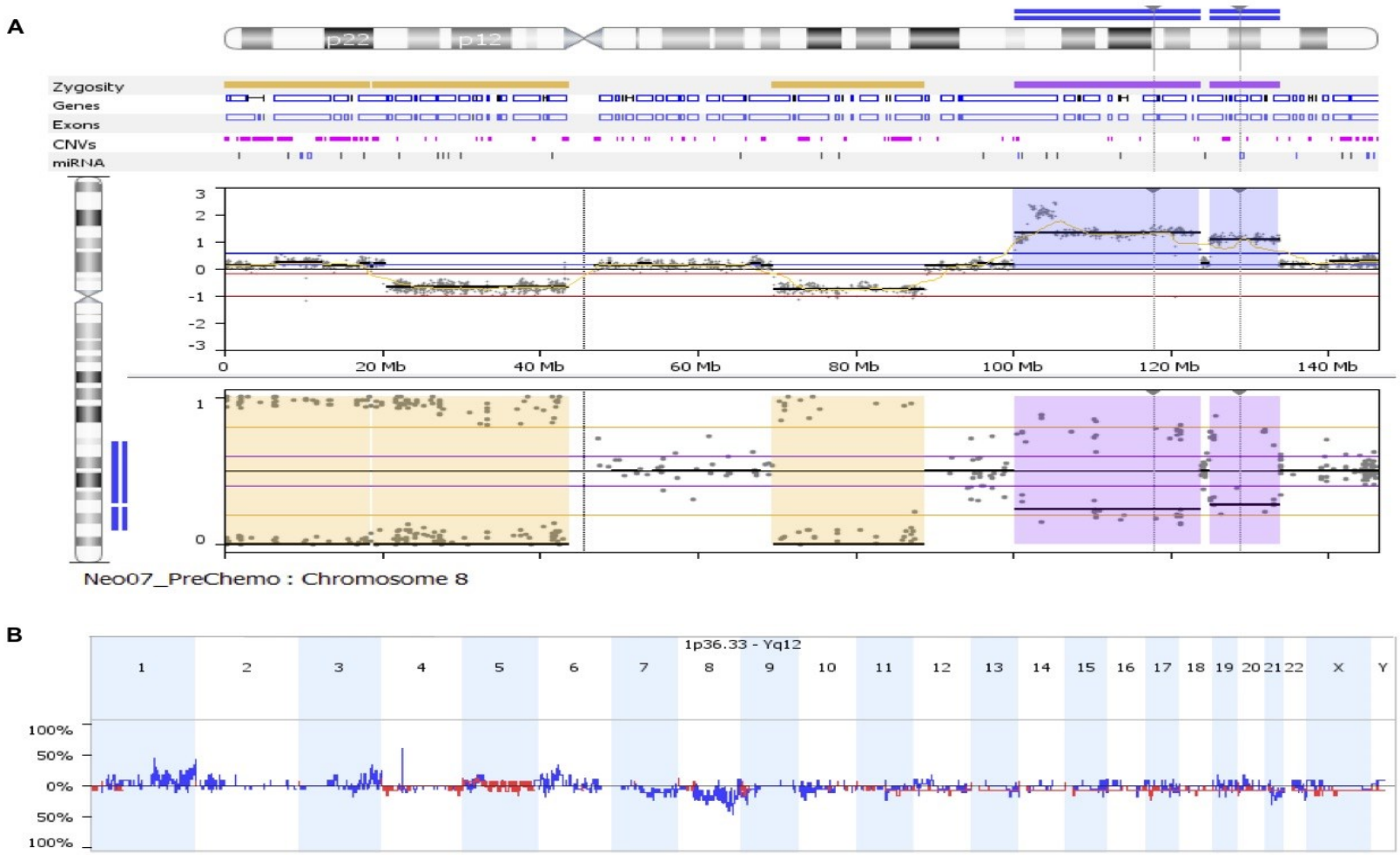


Figure 4. 5. RAD21 Amplification

A. DNA copy number data of chromosome 8 in Neo07 based on whole exome sequencing data using NEXUS software. Vertical line at 118MB indicates location of RAD21 gene and at 128MB, MYC gene. B. Differential DNA copy number changes in RBC0/1 versus RCB2/3 tumors. The Y-scale represents the % of tumors carrying the amplification. There is amplifications in RCB2/3 at chr8q and in RCB0/1 at chr1q.

We further looked at other available RNAseq data in 19 patients and found that mean RAD21 expression was 2.6 fold higher in the chr8 amplified tumors than in the non amplified tumors ($p=0.037$) in the whole group of 28 tumors (similar results for the EIF3H and UTP23 genes). Association of RAD21 with chemo-resistance was demonstrated before in the MDA-MB-231 triple negative breast cancer cell line (Xu, et al. 2011). There was no significant correlation between gene expression and other amplified chromosomal fragments (chr 1, 3 and 6).

When comparing the DNA copy number changes from pre to post-chemo tumors, there were two functional amplicons (9p, 1p) with significant increase gene expression. An amplicon on chromosome 9p in Neo27 was functional in nine genes, including NFIB, CNTLN and FREM1 (Table 4.8). NFIB has been reported to be involved in resistance to cisplatin in ovarian cancer (Kashiwagi, et al. 2011). The other novel amplicon on Chr1p showed increase in copy number and gene expression for two genes, CRYZ (6.5 fold change in RNA) and TYW3 (3.5 fold change in RNA).

Table 4. 8. RNA expression of genes on chr9 amplicon in Neo-27

Gene Name	Gene Id	chrom	end position	aCGH Change (LogRatio)	Neo27 pre (reads)	Neo27 post (reads)	Neo27 Ratio of RNAseq
TYRP1	NM_000550	9	12710266	-0.67	8	14	1.8
LURAP1L	NM_203403	9	12823059	1.95	994	790	0.8
MPDZ	NM_003829	9	13250365	4.22	7124	58872	8.3
FLJ41200	NM_033863	9	13431328	4.52	216	10030	46.4
NFIB	NM_001190738	9	14398982	4.7	21106	149254	7.1
ZDHHC21	NM_178566	9	14693480	4.72	10868	217346	20
CER1	NM_005454	9	14722715	4.46	4	22	5.5
FREM1	NM_144966	9	14910993	4.46	602	15322	25.5
TTC39B	NM_001168340	9	15307358	3.54	4730	3882	0.82
SNAPC3	NM_001039697	9	15461627	3.54	5722	63470	11.1
PSIP1	NM_001128217	9	15511003	-0.6	2862	3776	1.3
C9orf93	NM_173550	9	15971897	0.32	352	540	1.5
BNC2	NM_017637	9	16870786	4.16	2220	482	0.22
CNTLN	NM_017738	9	17503917	5.06	1294	24034	18.6
SH3GL2	NM_003026	9	17797122	-0.17	6	10	1.7

4.2.4. Gene expression analysis

We performed RNAseq on available 28 pre-chemo and 14 post-chemotherapy samples. We analyzed samples in two different ways to provide comprehensive picture of the changes in tumor that became resistant to chemotherapy

1: RCB0/1 versus RCB2/3 pre-chemotherapy samples to identify genes whose expression could predict complete tumor response

2: post-chemo versus pre-chemo in matched tumors in resistant tumors to identify novel gene expression profiles associated with chemotherapy resistance.

There were 160 genes with significant gene expression change between RCB0/1 vs. RCB2/3 tumors ($p<0.05$). Gene ontology analyses on these 160 genes showed a very strong enrichment for genes related to the immune response (Table 4.9).

Involvement of immune response genes in chemo resistance was previously reported by Denkert et al. Their group had published a suggested immune response gene list to predict chemotherapy response to the neoadjuvant therapy (Denkert, et al. 2015). On the other hand, TNBC has different subtypes including immunomodulatory. Therefore, we looked at our samples based on Vanderbilt TNBC subtypes (Lehmann, et al. 2011) to find what portion of the samples is in immunomodulatory sub-group. We found six of 12 RCB0/1 tumors were immunomodulatory sub-type compared to three of 16 RCB2/3 samples ($p = .11$). When we looked at the gene list that was published by Denkert et al, we found high expression of them (above mean of all tumors) in the immunomodulatory genotype (Figure 4.6).

To find genes whose expression was associated with chemotherapy resistance, we removed the nine strongly immune modulated tumors from further analysis and compared the remaining RCB0/1 tumors to the remaining RCB2/3 tumors. There were 40 significantly differentially expressed genes between these groups ($p<0.05$) with MYB and ABCA8 among the top five (Figure 4.6).

Table 4. 9. Gene ontology analysis of the 160 genes whose expression was significantly (p<0.05) different in RCB0/1 vs RCB2/3 tumors

Term	Count	PValue	Genes	FDR
GO:0050853~B cell receptor signaling pathway	10	2.60E-10	IGHG1, CD38, IGHG3, KLHL6, IGHV3-23, IGHA1, ZAP70, NFAM1, IGLC2, IGLC3	4.10E-07
GO:0050776~regulation of immune response	13	1.13E-08	ICAM1, ITGAL, CD96, SH2D1A, TRAC, CD3E, CD247, IGHV3-23, SLAMF6, SLAMF7, IGLC2, IGLV3-1, IGLC3	1.78E-05
GO:0042110~T cell activation	8	6.19E-08	PIK3CG, ITK, NLRC3, CD3E, ZAP70, TNFSF14, IRF4, CD7	9.75E-05
GO:0045087~innate immune response	17	1.75E-07	PIK3CG, IGHG1, ITK, IGHG3, S100A9, SLAMF6, LY9, CLEC10A, CD180, SH2D1A, IGHV3-23, ZAP70, IGHA1, PSTPIP1, IGLC2, CD6, IGLC3	2.76E-04
GO:0050871~positive regulation of B cell activation	6	1.38E-06	IGHG1, IGHG3, IGHV3-23, IGHA1, IGLC2, IGLC3	0.002168607
GO:0006910~phagocytosis, recognition	6	2.03E-06	IGHG1, IGHG3, IGHV3-23, IGHA1, IGLC2, IGLC3	0.003200698
GO:0006955~immune response	15	4.04E-06	TNFSF14, IGLV3-1, IGSF6, CD96, CXCL14, CST7, LAX1, S1PR4, CCR2, IGHV3-23, ZAP70, IGHA1, IL2RG, SPN, CD7	0.006363967
GO:0006911~phagocytosis, engulfment	6	6.43E-06	IGHG1, IGHG3, IGHV3-23, IGHA1, IGLC2, IGLC3	0.010129771
GO:0002250~adaptive immune response	9	1.77E-05	PIK3CG, ITK, SH2D1A, LAX1, ZAP70, SLAMF7, CD6, CLEC10A, CD7	0.027811031
GO:0031295~T cell costimulation	7	2.86E-05	TRAC, CD3E, CD247, TNFSF14, GRAP2, CD5, SPN	0.045037197
GO:0006958~complement activation, classical pathway	7	1.09E-04	IGHG1, IGHG3, IGHV3-23, IGHA1, IGLC2, IGLV3-1, IGLC3	0.172370846
GO:0042742~defense response to bacterium	8	1.23E-04	IGHG1, IGHG3, IGHV3-23, S100A9, IGLC2, S100A14, SPN, IGLC3	0.192882539
GO:0007165~signal transduction	22	1.79E-04	ITGAL, ZNF831, ITK, IL2RB, MPP2, RHPN2, CRABP2, S100A9, TNFSF14, RCAN1, NFAM1, CD38, RAC2, CXCL14, RASAL3, PSTPIP1, CSF2RB, CSF3R, PDE9A, IL2RG, FGF2, SPN	0.282036649

Term	Count	PValue	Genes	FDR
GO:0038096~Fc-gamma receptor signaling pathway involved in phagocytosis	7	4.25E-04	IGHG1, IGHG3, CD247, IGHV3-23, IGLC2, IGLV3-1, IGLC3	0.667122872
GO:0006956~complement activation	6	5.33E-04	IGHG1, IGHG3, IGHV3-23, IGLC2, IGLV3-1, IGLC3	0.837215067
GO:0072540~T-helper 17 cell lineage commitment	3	5.68E-04	SLAMF6, LY9, IRF4	0.891047895
GO:0050852~T cell receptor signaling pathway	7	9.51E-04	ITK, PRKCQ, TRAC, CD3E, CD247, ZAP70, GRAP2	1.487412247
GO:0042113~B cell activation	4	0.001367591	IKZF3, LAX1, ZAP70, BANK1	2.133023801
GO:0046641~positive regulation of alpha-beta T cell proliferation	3	0.002004087	CD3E, CCR2, ZAP70	3.111164147
GO:0038095~Fc-epsilon receptor signaling pathway	7	0.00243433	ITK, PRKCQ, IGHV3-23, GRAP2, IGLC2, IGLV3-1, IGLC3	3.767149836
GO:0006954~inflammatory response	10	0.002492988	PIK3CG, ITGAL, PRKCQ, CCR2, S100A9, PSTPIP1, ZAP70, IL17RE, NFAM1, CD180	3.856262395
GO:0006898~receptor-mediated endocytosis	7	0.003028293	IGHV3-23, IGHA1, CD6, IGLC2, CD5, IGLV3-1, IGLC3	4.665923483
GO:0045060~negative thymic T cell selection	3	0.003031374	CD3E, ZAP70, SPN	4.670564679
GO:0045086~positive regulation of interleukin-2 biosynthetic process	3	0.003619545	PRKCQ, CD3E, IRF4	5.552783712
GO:0032740~positive regulation of interleukin-17 production	3	0.004256369	PRKCQ, SLAMF6, LY9	6.499353486
GO:0042102~positive regulation of T cell proliferation	4	0.010792635	PRKCQ, CD3E, CD6, SPN	15.71457498
GO:0006968~cellular defense response	4	0.011797806	ITK, SH2D1A, CCR2, SPN	17.05384518
GO:0030593~neutrophil chemotaxis	4	0.013964993	PIK3CG, S100A9, CSF3R, TGFB2	19.8738515

Term	Count	PValue	Genes	FDR
GO:0006935~chemotaxis	5	0.014183461	RAC2, CXCL14, CCR2, FGF2, SPN	20.15309019
GO:0001816~cytokine production	3	0.015423483	PIK3CG, ITK, S100A9	21.72085716
GO:0097190~apoptotic signaling pathway	4	0.016971589	CD38, CD3E, CD5, SPN	23.63768279
GO:0007204~positive regulation of cytosolic calcium ion concentration	5	0.019349653	PIK3CG, EDNRB, CD38, S1PR4, CCR2	26.49671124
GO:0060412~ventricular septum morphogenesis	3	0.020464503	HEY2, PROX1, TGFB2	27.80223982
GO:0010621~negative regulation of transcription by transcription factor localization	2	0.022695501	ID1, HEY2	30.34986847
GO:0002291~T cell activation via T cell receptor contact with antigen bound to MHC molecule on antigen presenting cell	2	0.022695501	ICAM1, ITGAL	30.34986847
GO:0030888~regulation of B cell proliferation	2	0.022695501	IKZF3, MZB1	30.34986847
GO:0043547~positive regulation of GTPase activity	10	0.028967731	ICAM1, IL2RB, ACAP1, RASAL3, CSF2RB, IL2RG, DENND5B, FGF2, RAPGEFL1, DENND1C	37.06858481
GO:0035910~ascending aorta morphogenesis	2	0.030146509	HEY2, TGFB2	38.26146402
GO:0038110~interleukin-2-mediated signaling pathway	2	0.030146509	IL2RB, IL2RG	38.26146402
GO:0032496~response to lipopolysaccharide	5	0.036860392	CD96, CSF2RB, ACP5, CD6, S100A14	44.66174092
GO:0007169~transmembrane receptor protein tyrosine kinase signaling pathway	4	0.037029159	ITK, CD3E, ZAP70, CD7	44.81431372
GO:0001895~retina homeostasis	3	0.037244934	AZGP1, IGHG3, IGHA1	45.00881071

Term	Count	PValue	Genes	FDR
GO:0090330~regulation of platelet aggregation	2	0.044879849	PRKCQ, ZAP70	51.49173914
GO:0045577~regulation of B cell differentiation	2	0.044879849	IKZF3, NFAM1	51.49173914
GO:0070374~positive regulation of ERK1 and ERK2 cascade	5	0.044990022	ICAM1, NDRG4, SERPINF2, NPNT, FGF2	51.57981883
GO:0032729~positive regulation of interferon-gamma production	3	0.047991166	CD3E, CCR2, SLAMF6	53.92232758

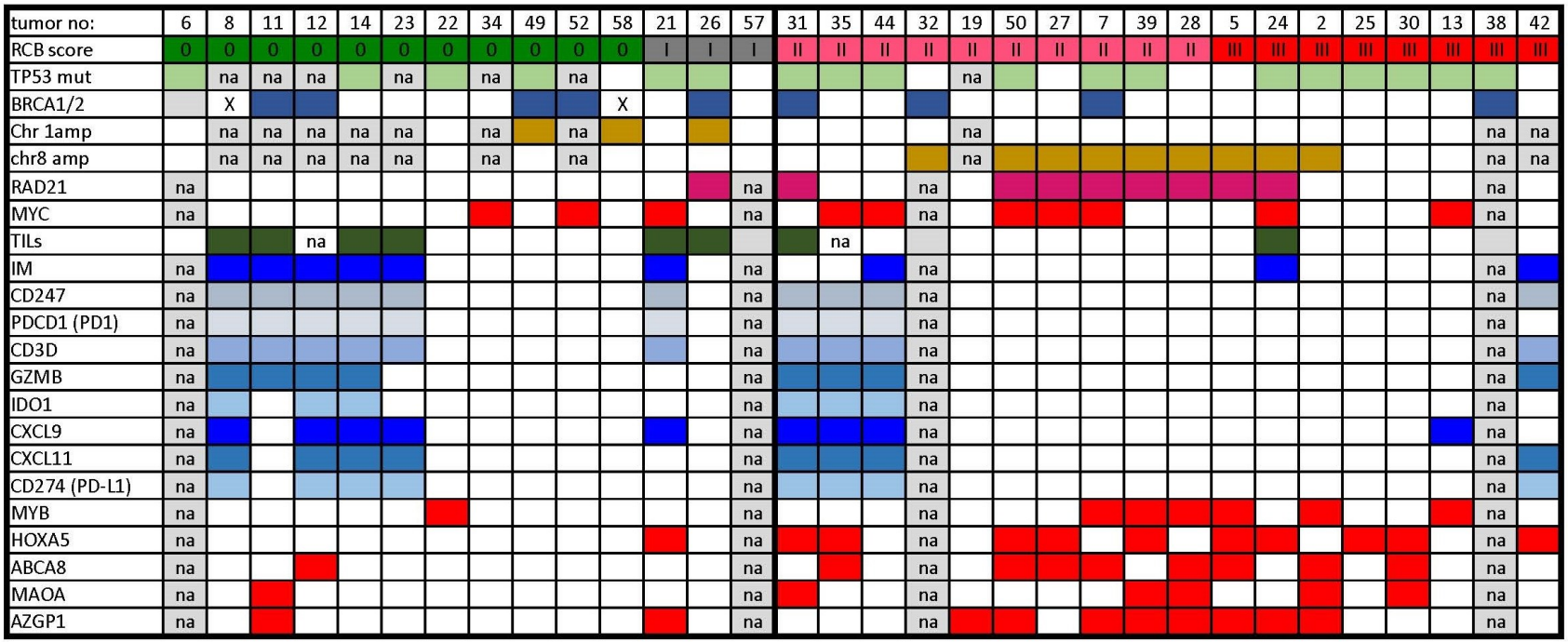


Figure 4. 6. Genomic alterations in pre-chemotherapy tumor samples from 32 Q-CROC-3 TNBC patients.

TP53 ,BRCA1/2 status was shown with filled boxes
RAD21/MYC : filled box: RNAseq reads for that gene in that tumor is above mean for all samples

In the next step, RNAseq of the 14 post-chemo tumors was compared to corresponding matched pre-chemo. Six genes were significantly expressed in post-chemo in comparison to pre-chemo samples. DUSP1 and RGS were increased and GREM1, MMP11, IGHG4 and IGLV1-51 were decreased in the post-chemo samples.

As long as the previous study that was done by Dr. Basik group on resistant TNBC cell lines revealed ABC gene fusion as the main cause of resistance in those samples, we next looked at possible gene fusions to see if the same pattern is happening in patient samples. After removing all fusions involving immunoglobulin genes, we identified 134 gene fusions in these samples. Thirty-five of the fusions represented inter-chromosomal translocations. There were 52 gene fusions in post-chemotherapy samples and in seven of them, RNA expression increase (>4-fold) was detected in at least one of the two genes involved (Table 4.10). Importantly we found fusions that involves the ABCB1 gene, which is associated with chemo resistance in cancer.

Table 4. 10. Highly expressed post-chemotherapy Gene Fusions

Fusion Name (Gene1--Gene2)	Sample	Left Chr	Left Chr BP	Right Chr	Right Chr BP	Gene1 PostC/PreC Expression Ratio	Gene2 PostC/PreC Expression Ratio
SH3GL3--TM6SF1	Neo_05_PostChemo	15	84159654	15	83781549	8.6	2.5
TFG--GPR128	Neo_25_PostChemo	3	100438902	3	100348442	1.6	8.1
DSP--CALCR	Neo_25_PostChemo	6	7563013	7	93116319	1.1	18.1
IL23R--INADL	Neo_25_PostChemo	1	67672738	1	62586853	15.5	0.9
LINGO1--TNNC2	Neo_27_PostChemo	15	78088281	20	44461967	4.1	5.8
SMCO3--C12orf60	Neo_42_PostChemo	12	14967060	12	14975846	32.1	38.5
KIAA1430--ABCB1	Neo_50_PostChemo	4	186124939	7	87230394	0.64	28.2

4.3. Conclusion

In this study, we performed extensive molecular analysis of pre- and post-chemotherapy tumor samples in response to neoadjuvant chemotherapy treatment. Here, we presented the results of array CGH, RNAseq and whole exome sequencing (WES) analysis of tumor samples obtained from 29 patients (array CGH on 9 matched pre/post tumors pairs, RNAseq on 28 pre-chemo samples and 14 matched pre/post pairs and WES on 25 patients). The response of the tumor to chemotherapy was not the same in all tumors and varied from pCR to even growing during chemotherapy.

Our findings suggest that there are two major driving factors in chemotherapy response in the neoadjuvant setting in TNBCs. First one is the role of immune response (PD-L1 expression) and second one is the amplification of chromosome 8q23 particularly RAD21, EIF3H and UTP23genes. We also found a novel focal amplification on chromosome 9 involving the NFIB gene in Neo27 post-chemo sample. This amplification was validated to be functional by RNAseq data and gene expression of nine genes located there was increased in the post chemotherapy sample compared with the pre-chemotherapy one. NFIB role in chemo-resistance to cisplatin was previously reported in ovarian cancer (Kashiwagi, et al. 2011).

WES results showed great genomic heterogeneity of TNBC tumors and any of the samples was almost unique with TP53 somatic mutation as the only shared somatic mutation. In comparison between pre and post samples, 65-95% of the variants were conserved. When limited "gained" or "lost" variants were evaluated by ddPCR, they were mostly detected in plasma of both pre- and post-chemo samples. This finding suggests little genomic change from pre to post-

treatment and the possible role of tumor heterogeneity for some of the changes in AF in detected variants.

In the recent NGS study on 74 post-chemotherapy residual TNBCs after neoadjuvant chemotherapy, TP53 was the most mutated gene and JAK2 and MCL-1 (amplified in 11% and 54% of samples respectively) was proposed as two actionable candidate genes (Balko, et al. 2014). Although our results were similar in TP53 mutations incident to them (89% vs 84%), we just found JAK2 and MCL1 genes amplification in two different tumors. One of the reasons of this discrepancy is usage of formalin-fixed paraffin embedded tumor in their study versus frozen biopsies in ours. In addition, they only used 164 gene panel and RAD21 was not evaluated by them. We also could not reject the role of heterogeneity of TNBCs for these differences. Therefore, evaluation of these genes in larger scale data will help to uncover the drug resistance problem.

Finally, in four patients, we analyzed WES data from matched lymph nodes and metastatic tumor sample. Our findings showed that 91% of variants present in the metastatic tumors were detected in the post-chemotherapy tumor samples, compared with 85% in the pre-chemotherapy samples. Even though the sample size was limited, these findings suggest that the post-chemotherapy sample is a better indicator of the metastatic genotype than the pre-chemotherapy sample. Further genomic and transcriptomic analysis of large scale of metastatic lesions may lead to novel therapeutic approaches to overcome drug resistant TNBC patients.

Chapter 5: ClinPred – A Prediction method to identify clinically relevant nonsynonymous single nucleotide variants.

A version of this chapter is published as:

"ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants." Najmeh Alizeraie, Kristin D. Kernohan, Taila Hartley, Jacek Majewski, Toby Dylan Hocking, Am J Hum Genet. 2018 Oct 4;103(4):474-48

Author contributions are stated in the Contribution of the Authors section.

5.1. Introduction

Immense progress in high throughput sequencing technologies provides new opportunities for identifying genetic determinants of disease. “Next Generation” sequencing is now firmly established in diagnostic and research laboratories. Although recent advances in these technologies make them affordable, interpreting the effect of discovered variants remains a serious challenge. Since the human exome on average contains around 20,000 single nucleotide variants, as compared with the reference,(Shihab, et al. 2014) it is crucial to accurately predict deleteriousness of genomic changes, especially nonsynonymous single nucleotide variants (nsSNVs). Distinguishing pathogenic amino acid changes from background polymorphisms is essential for efficient use of these technologies in personalized medicine. Experimental validation of the pathogenicity of large numbers of variants is not feasible as it is expensive and time consuming. Consequently, many algorithms have been developed to predict the potential impact of a variant on protein structure and/or function. These methods use different properties of the variant, such as relationship to local protein structure, evolutionary conservation and/or physiochemical and biochemical properties of amino acids.

While the current programs provide positive predictive power, their results are often in disagreement with each other, (Ioannidis, et al. 2016; Li, et al. 2014) and there are currently no guidelines as to which predictions are the most reliable. It is believed that individual methods have complementary strengths, depending on their specific features and computational algorithms. (Gonzalez-Perez and Lopez-Bigas 2011; Ioannidis, et al. 2016; Liu, et al. 2011) Hence, recently, new “ensemble” predictors have combined individual predictors in order to achieve higher classification accuracy.

Existing ensemble prediction tools apply machine learning algorithms and have been trained on known pathogenic and neutral nsSNVs mostly from HGMD or UNIPROT databases. While those databases provide important information about variants associated with diseases, they have known limitations. To improve functional annotation of human variation, the more recently developed ClinVar (Landrum, et al. 2018) database recommends that submitters use American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) guidelines (Richards, et al. 2015) for clinical interpretation of variants. Since its release in 2013, ClinVar has grown rapidly and has become the powerful resource representing current understanding of the relationship between genotypes and medically important phenotypes.(Harrison, et al. 2016)

In this chapter, we hypothesized that by developing a machine learning approach trained on the most up to date and highest quality data, we will facilitate more accurate and reliable prediction of variants' relevance to genetic disease.

5.2. Results

The full description of the model is available in the method chapter. Briefly, our classifier, ClinPred, combined random forest and gradient boosting models. As predictive features, we combined commonly used and recently developed individual prediction tool scores, as well as allele frequencies (AF) of the variant in different populations from gnomAD database. Our model is the first to train on variants from the ClinVar database. Moreover, we used AFs in different populations as features, rather than filtering variants based on arbitrary AF cutoffs, as is the case with most currently used approaches. We also assembled large independent test sets to evaluate ClinPred on cancer and rare disease data, and to compare its performance with other existing methods.

5.2.1. Performance comparison of our models and individual component features

Our two models (cforest and xgboost) were superior to all their constituent features and discriminated well between pathogenic and benign variants in ClinVarTest with AUC equal to 0.97 ± 0.004 (mean ± standard deviation in 5-fold CV) for xgboost and cforest (Figure 5.1). They also showed superior performance in MouseVariSNP with respective AUCs of 0.96 ± 0.01 and 0.96 ± 0.02. Although most features and our models demonstrated little change in AUC score between MouseVariSNP and ClinVarTest, DANN and Siphy24-way attained 11% lower AUC in MouseVariSNP compared to ClinVarTest. Overall, the single features with the highest AUC were AF (gnomAD_exome_ALL), followed by PROVEAN, Polyphen-HVAR and CADD (Figure 5.1). Consistent with other research findings, conservation scores (GERP++, phastCons, PhyloP and SiPhy) almost all have lower AUC than functional scores (SIFT, MutationAssessor, PROVEAN, PolyPhen-2 HDIV and HVAR).(Dong, et al. 2015) We further investigated the effect of

excluding/including AF as a feature in the models and found that the inclusion of AF significantly increases AUC as well as increasing sensitivity and specificity (Figure 5.2).

Finally, we found that combining our models by selecting the higher of the two probability scores improved the AUC to 0.98 ± 0.004 and 0.96 ± 0.01 in ClinVarTest and MouseVariSNP respectively, while also achieving the best specificity at 95% sensitivity. Hence, we defined this combined model as ClinPred and used it in subsequent tests.

Figure 5. 1. The performance of our models was compared against their constituting features and other available tools in ClinVarTest and MouseVariSNP.

Analysis is based on the raw scores and was calculated for 5-fold cross validation. Adapted from Alirezaie et al., The American Journal of Human Genetics, 2017.

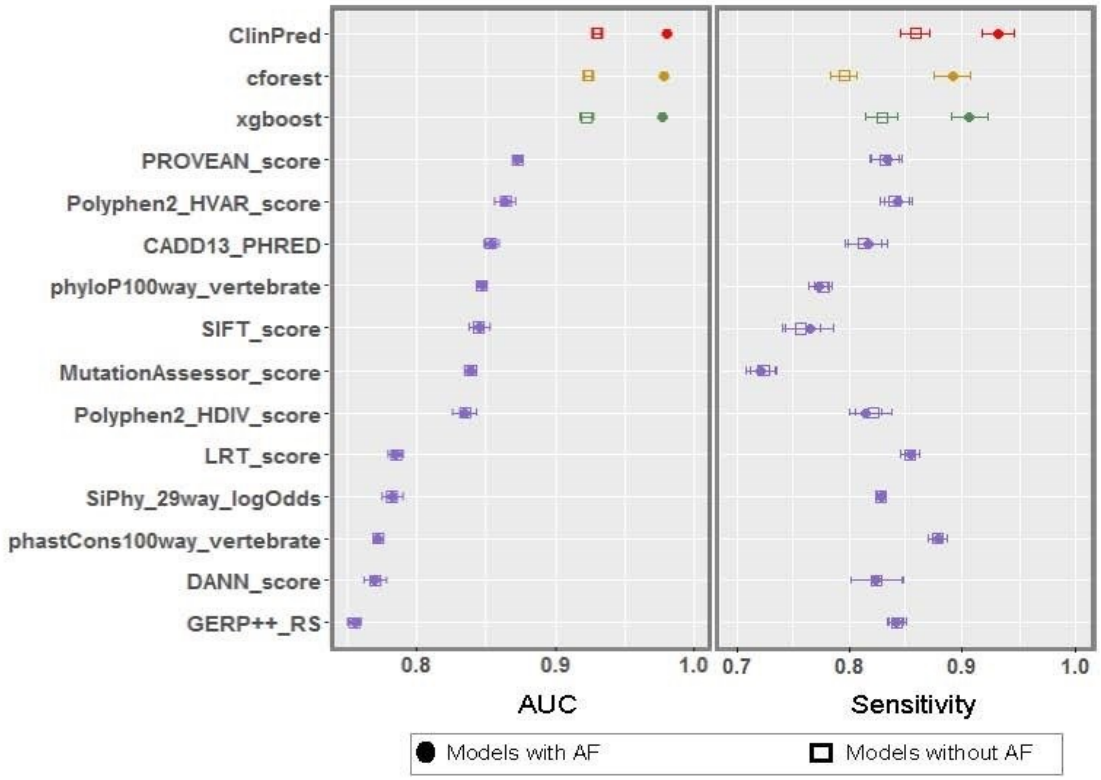


Figure 5. 2. AF boost sensitivity and AUC score when applied as a feature in our models.

Adapted from Alirezaie et al., The American Journal of Human Genetics, 2017.

5.2.2. ClinPred in comparison to other ensemble tools

Using the ClinVarTest dataset, ClinPred outperformed other classifiers with the best AUC (0.98 ± 0.004) (Figure 5.3), sensitivity (93.1 ± 3 %) and specificity (94.2±0.04%). It had the lowest error rate (6.04%) - the sum of false positives and false negatives over total number of labeled variants - in comparison to other tools (Figure 5.4 and Table 5.1) where the error rate ranged from 13.2% (REVEL) to 50.3% (M-CAP).

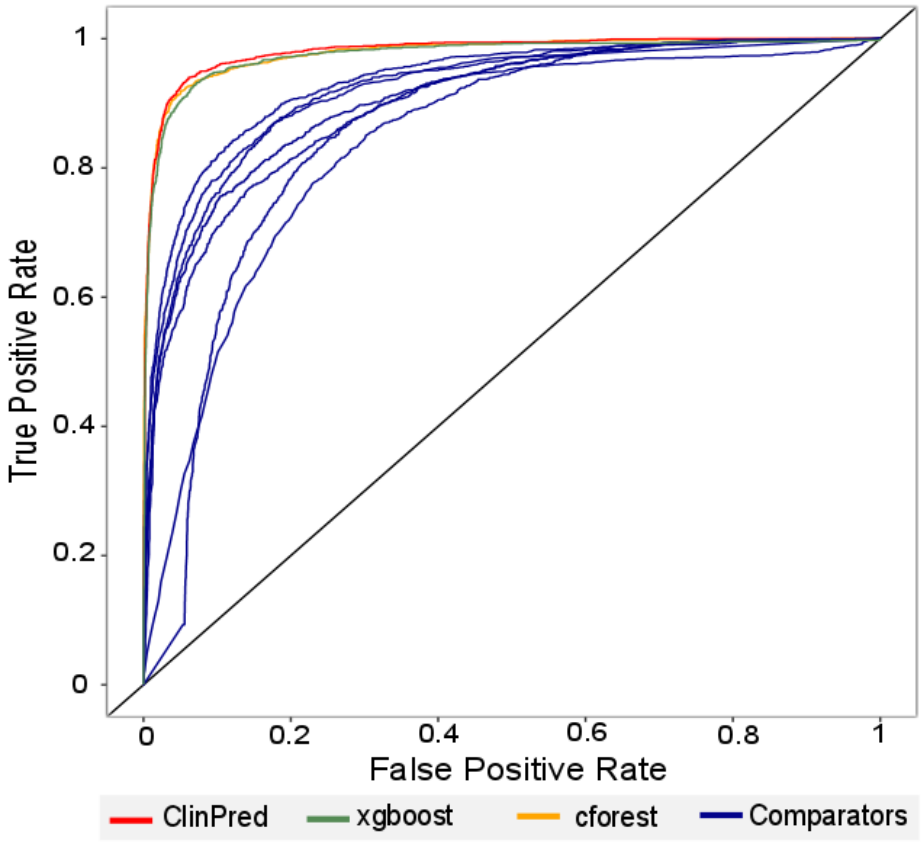


Figure 5. 3. AUC was compared between our models and seven recently developed tools using in ClinVarTest data.

Adapted from Alirezaie et al., The American Journal of Human Genetics, 2017.

Table 5. 1. Overview of performance of ClinPred in comparison to raw scores of other tools in ClinVarTest

Adapted from Alirezaie et al., The American Journal of Human Genetics, 2017.

	sensitivity	specificity	FPR	accuracy	precision	Error Percent	F1 score	MCC
ClinPred	0.94	0.94	0.06	0.94	0.86	6.04	0.90	0.85
xgboost	0.91	0.95	0.05	0.94	0.87	6.42	0.89	0.84
cforest	0.89	0.97	0.03	0.95	0.91	5.49	0.90	0.86
VEST3_score	0.83	0.84	0.16	0.84	0.66	16.48	0.73	0.62
MetaSVM_score	0.78	0.85	0.15	0.83	0.67	16.84	0.72	0.60
MetaLR_score	0.80	0.80	0.20	0.80	0.60	20.18	0.69	0.55
M-CAP_score	0.84	0.36	0.64	0.50	0.34	50.36	0.48	0.20
fathmm-MKL_score	0.84	0.69	0.31	0.73	0.51	26.53	0.64	0.48
Eigen-raw	0.76	0.74	0.26	0.74	0.53	25.58	0.62	0.45
REVEL	0.82	0.89	0.11	0.87	0.74	13.20	0.77	0.68

FPR: False positive rate
MCC: Matthews correlation coefficient

When used on MouseVariSNP, ClinPred again outcompeted other available methods (Table 5.2). VEST3 was the closest competitor with an AUC of 0.88± 0.03. All methods were less accurate in MouseVariSNP than in ClinVarTest; the method with the largest AUC decrease was FATHMM (from 0.78±0.1 to 0.58±0.07) (Figure 5.1). Although ClinPred achieved the highest specificity in MouseVariSNP, it was followed closely by REVEL. This might be due to type I circularity, considering that VariSNP has overlap with the training set of other tools. On the other hand, as pathogenic variants in MouseVariSNP have the least overlap with the training data used by other tools, sensitivity score is the least biased comparator. ClinPred had the highest sensitivity among tools, detecting 92.79±3.04% of pathogenic variants and VEST3 was the next, achieving 85.26±3.34% sensitivity.

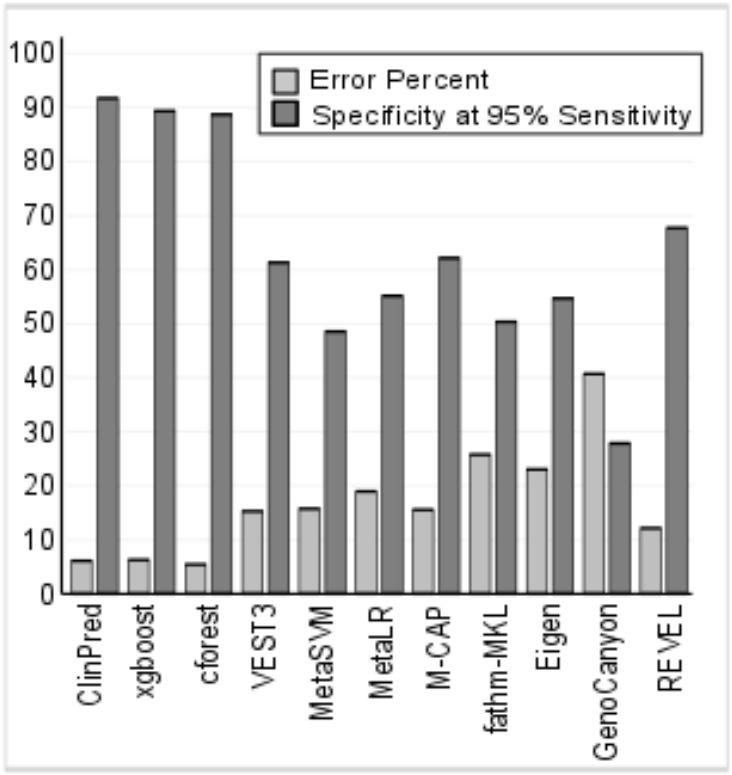


Figure 5. 4. The performance of our models were compared to seven recently developed tools using ClinVarTest data.

Our models had the best specificity at the cut off required to achieve 95% sensitivity. Adapted from Alirezaie et al., The American Journal of Human Genetics, 2017.

In order to visualize the distribution of scores of recently developed ensemble tools, we plotted raw scores for pathogenic and benign variants in different data sets. As demonstrated in Figure 5.5, ClinPred scores were highly concentrated near 1 for pathogenic and 0 in the benign variants across datasets. This analysis provides another way to illustrate the ability of ClinPred to differentiate well between benign and pathogenic variants in comparison to other methods.

Table 5. 2. Overview of performance of ClinPred in comparison to raw scores of other models in MouseVariSNP test

Adapted from Alirezaie et al., The American Journal of Human Genetics, 2017.

Model	Sensitivity	Specificity	FPR	Accuracy	Precision	Error Percent	F1 score	MCC
ClinPred	0.93	0.88	0.12	0.89	0.50	11.44	0.65	0.63
xgboost	0.91	0.89	0.11	0.89	0.51	11.02	0.65	0.63
cforest	0.88	0.92	0.08	0.92	0.60	8.07	0.72	0.69
VEST3_score	0.86	0.78	0.22	0.79	0.34	20.98	0.48	0.45
MetaSVM_score	0.58	0.81	0.19	0.79	0.29	21.24	0.38	0.30
MetaLR_score	0.58	0.75	0.25	0.73	0.23	26.73	0.33	0.23
M-CAP_score	0.66	0.61	0.39	0.62	0.18	37.95	0.29	0.18
fathmm-MKL_score	0.75	0.68	0.32	0.69	0.23	31.15	0.36	0.28
Eigen-raw	0.76	0.73	0.27	0.73	0.27	26.67	0.40	0.34
REVEL	0.71	0.87	0.13	0.86	0.42	14.50	0.53	0.47

FPR: False positive rate
MCC: Matthews correlation coefficient

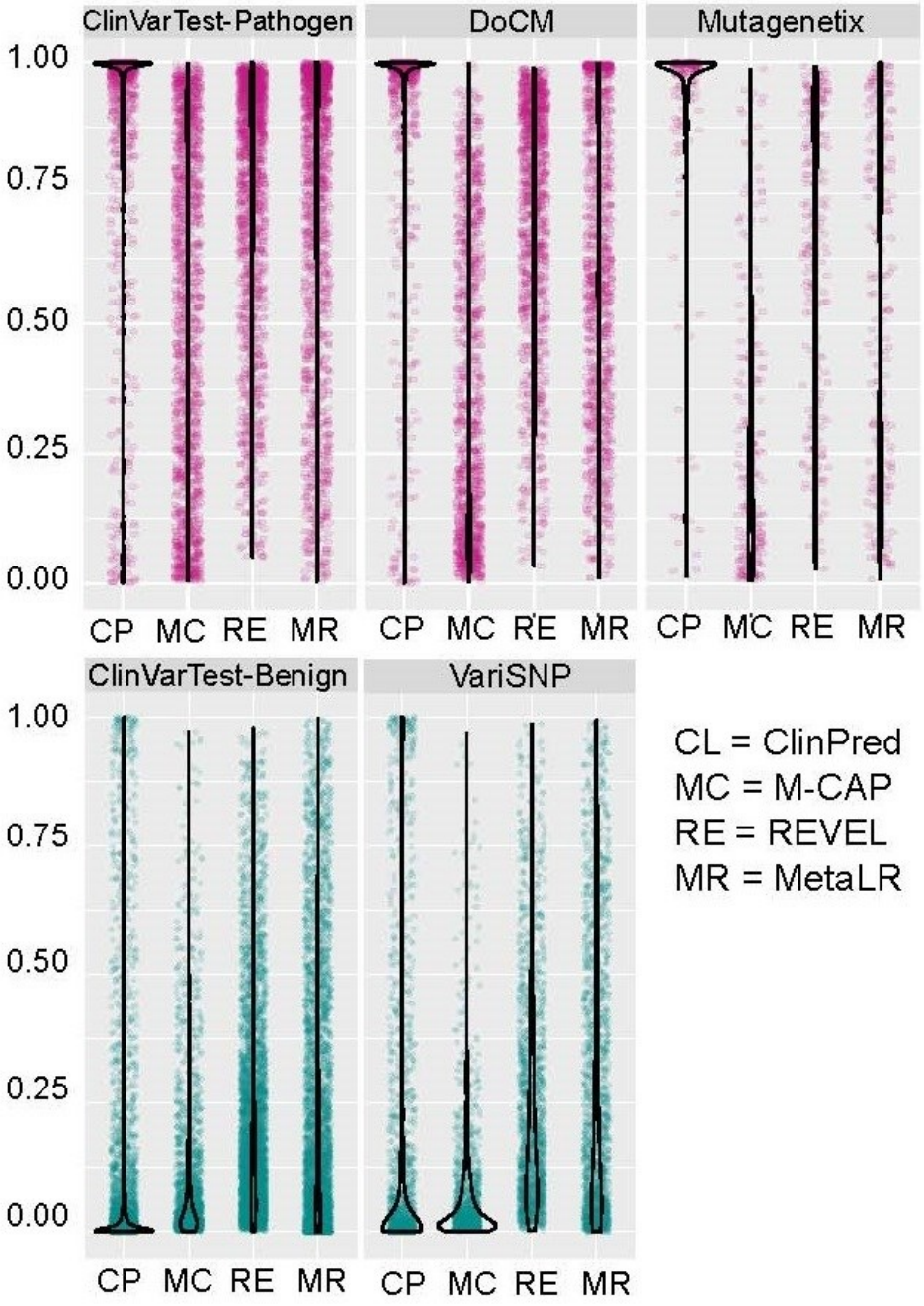


Figure 5. 5. Comparison of raw scores of ClinPred, M-CAP, REVEL, and MetaLR.

Violin plots represent the full distribution of scores for Pathogenic (Pink color) and Benign (Green color) variants in different test data. Adapted from Alirezaie et al., The American Journal of Human Genetics, 2017.

5.2.3. Using set allele frequency cutoffs versus allele frequency as a predictor variable

In most laboratories tasked with analyzing exome data, hard allele frequency cutoffs are used to filter lists of detected variants, and prediction scores are assessed for the remaining variants as part of variant interpretation. As a result, allele frequency has generally not been used explicitly to predict clinical relevance of mutations in previous approaches. This is a sensible approach, since a reasonable estimate of the maximum AF can be based on the mode of inheritance and population frequency of the phenotype. Moreover, this approach is supported by the current ACMG guidelines, where AF is given a higher evidence value than computational predictions (Richards, et al. 2015). However, in many cases, the mode of inheritance may not be evident – for example distinguishing recessive from de novo dominant cases – and population prevalence may not be obvious for non-specific phenotypes. Moreover, our analysis above suggested that population allele frequency is one of the most informative features in our model, and it is likely that it can acquire additional value when used in the machine learning setting alongside other predictor variables. Hence, we investigated whether AF remains an important predictor when the models are used in typical research approaches. We tested our models on datasets filtered according to various AF cutoffs: lower than 0.01, lower than 0.005 and lower than 0.001. In all conditions, ClinPred was superior to other tools, achieving highest AUC, sensitivity and specificity (Figure 5.6 and 5.7). Thus, even when using datasets that are likely to be seen in the research or clinical setting – i.e. filtered using typically applied allele frequency cutoffs – there is still valuable information to be learned from population AF.

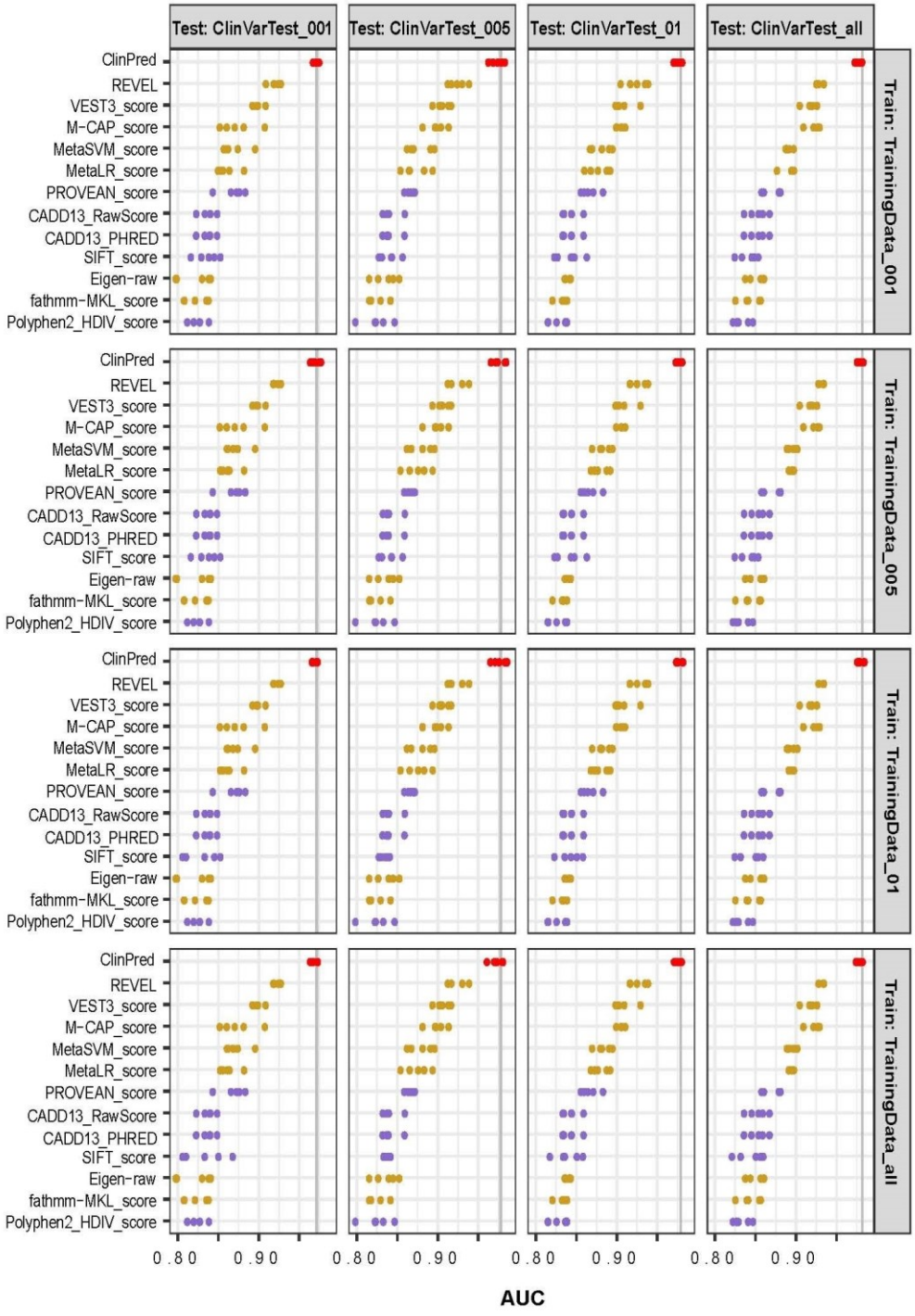


Figure 5. 6. AUC was compared to recently developed and commonly used tools using various AF cutoffs.

Adapted from Alirezaie et al., The American Journal of Human Genetics, 2017.

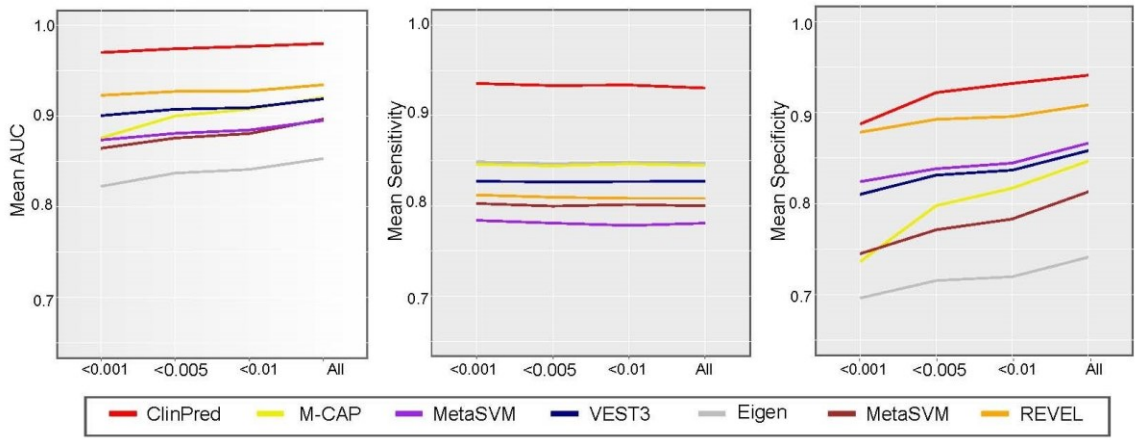


Figure 5. 7. Performance of ClinPred was compared to recently developed ensemble tools in different AFs.

Models were trained on the training data and tested on ClinVarTest using various AF cutoffs: all data set regardless of AF, AF less than 0.01, less than 0.005 and less than 0.001. Adapted from Alirezaie et al., The American Journal of Human Genetics, 2017.

5.2.4. Comparing categorical scores across different tools

Many current tools provide categorical predictions – pathogenic/damaging versus benign/tolerant – according to the authors' recommended pathogenicity thresholds. Hence, we compared the categorical predictions across various ensemble tools. As REVEL does not provide categorical scores, any variant with a raw score lower than 0.5 in REVEL was classified as benign while scores greater than or equal to 0.5 were classified as pathogenic. We used the same threshold for ClinPred. We restricted the comparison to variants where scores were available for any tool (excluding missing values). In ClinVarTest, M-CAP had the highest sensitivity, successfully classifying 95.8% pathogenic variants as damaging. However, this came at the cost of very low specificity, with 58% of benign variants misclassified as damaging. ClinPred achieved the second highest sensitivity (93.6 %), while maintaining a low false positive rate of 6% (Figure 5.8 and Table 5.3).

When tested on MouseVariSNP, ClinPred had the best performance according to both sensitivity and specificity (Figure 5.8, Table 5.4).

Table 5. 3.Overview of performance of ClinPred in comparison to categorical scores of other tools in ClinVarTest.

Adapted from Alirezaie et al., The American Journal of Human Genetics, 2017.

	Sensitivity %	Specificity %	FPR	Accuracy	Precision	Error Percent	F1 Score	MCC
ClinPred	93.58	94.10	0.06	0.94	0.86	6.04	0.90	0.85
xgboost	90.75	94.65	0.05	0.94	0.87	6.42	0.89	0.84
cforest	89.06	96.59	0.03	0.95	0.91	5.49	0.90	0.86
REVEL	82.55	89.27	0.11	0.87	0.75	12.60	0.78	0.70
M-CAP	95.79	41.62	0.58	0.64	0.54	35.79	0.69	0.42
MetaLR	77.93	83.87	0.16	0.82	0.65	17.79	0.71	0.59
Fathmm_mkl	96.48	43.70	0.56	0.58	0.40	41.65	0.56	0.38

FPR: False positive rate

MCC: Matthews correlation coefficient

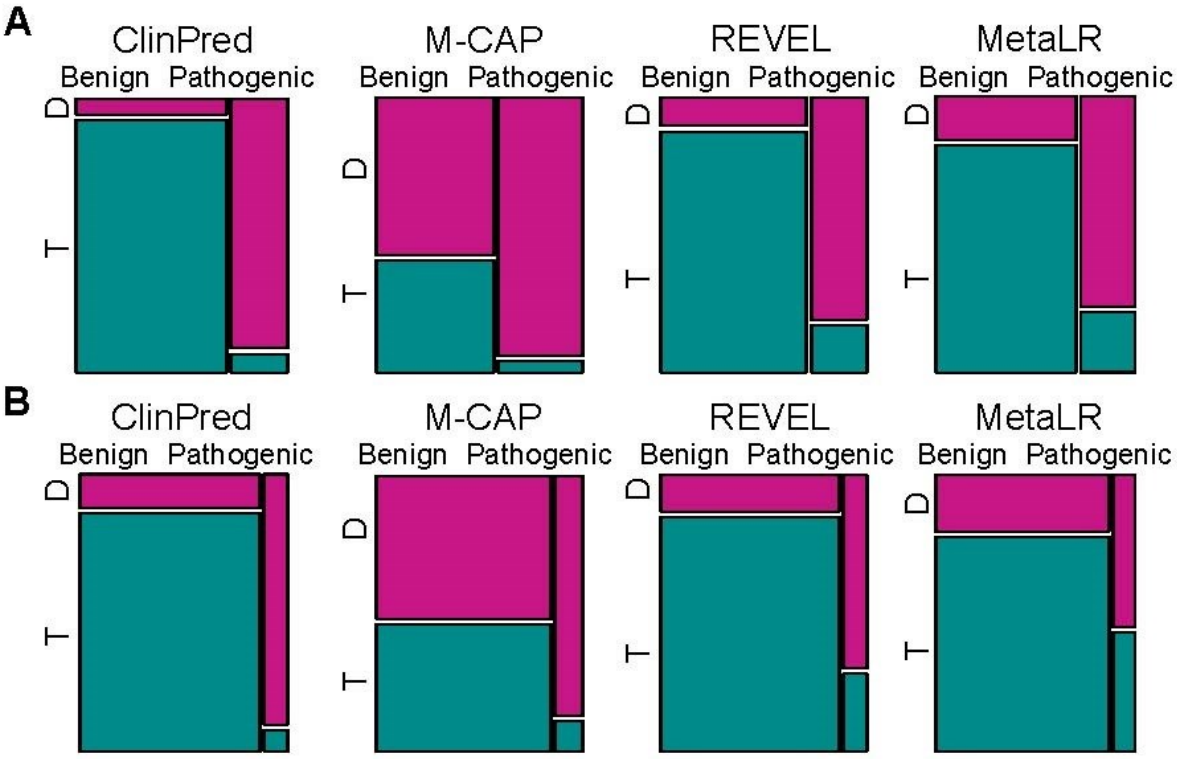


Figure 5. 8. Comparison of ClinPred with categorical predictions available from M-CAP, REVEL, and MetaLR.

REVEL and ClinPred scores lower than 0.5 are defined as tolerant and greater than 0.5 as damaging. We show proportions of benign and pathogenic variants that were classified as Tolerated (T, Green) and Damaging (D, Pink). ClinPred had the best performance in finding as many pathogenic variants possible while minimizing the number of benign variants that are predicted as damaging both in ClinVarTest (A) and MouseVariSNP (B). Adapted from Alirezaie et al., The American Journal of Human Genetics, 2017.

Table 5. 4. Overview of performance of ClinPred in comparison to categorical scores of other tools in MouseVariSNP test.

Adapted from Alirezaie et al., The American Journal of Human Genetics, 2017.

	Sensitivity %	Specificity %	FPR	Accuracy	Precision	Error Percent	F1 Score	MCC
ClinPred	92.63	88.04	0.12	0.89	0.50	11.44	0.65	0.63
xgboost	91.24	88.69	0.11	0.89	0.51	11.02	0.65	0.63
cforest	88.48	92.38	0.08	0.92	0.60	8.07	0.72	0.69
REVEL	71.43	86.65	0.13	0.85	0.41	15.09	0.52	0.46
M-CAP	88.73	47.20	0.53	0.53	0.21	47.16	0.34	0.25
MetaLR	56.28	79.25	0.21	0.77	0.26	23.36	0.35	0.26
Fathmm _mkl	91.16	38.92	0.61	0.45	0.16	55.15	0.27	0.20

FPR: False positive rate
MCC: Matthews correlation coefficient

Subsequently we investigated the performance of our models on the rare variants that are likely to be considered in current clinical testing. As M-CAP scores only rare variants (≤ 1% allele frequency), we restricted our analysis to the same cutoff. ClinPred maintained its performance as a classifier with lowest error rate in both ClinVarTest and MouseVariSNP restricted to AF≤ 1% (Figure 5.9).

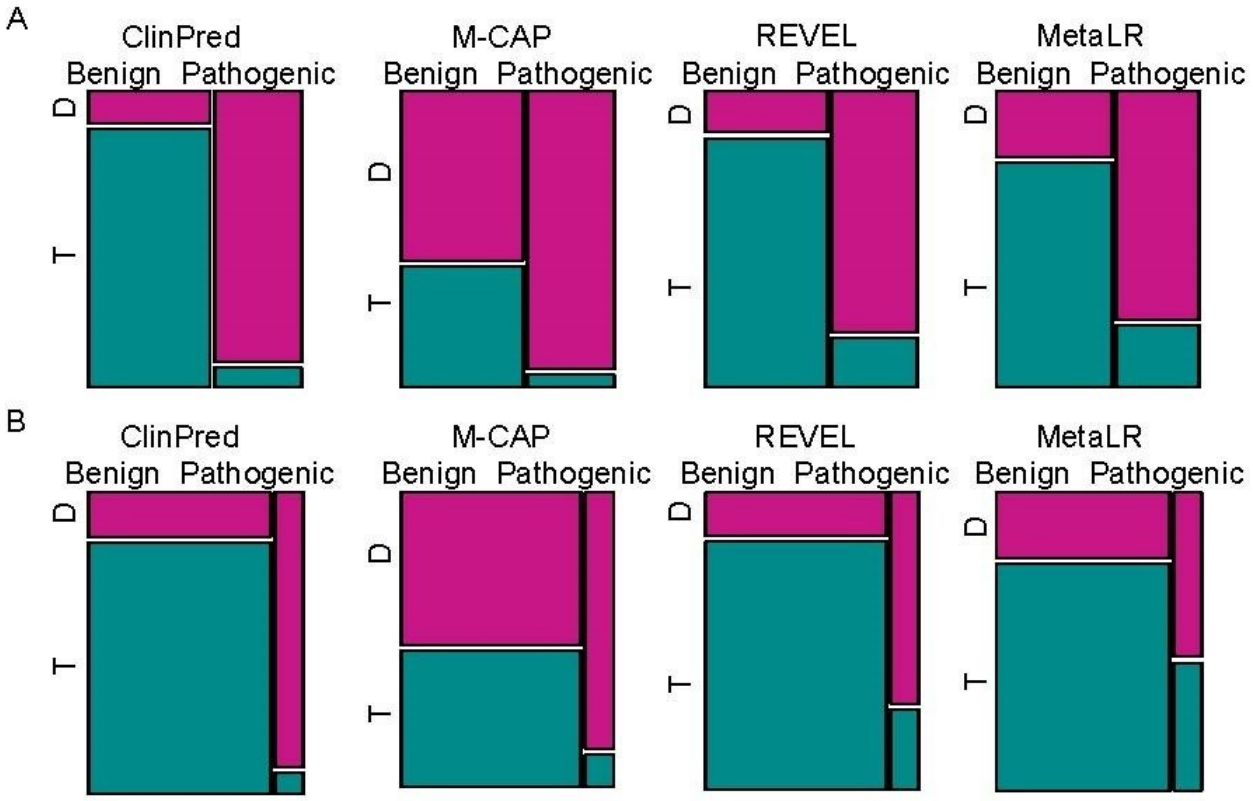


Figure 5. 9. Comparison of ClinPred with categorical predictions available from M-CAP, REVEL, and MetaLR in AF<0.01.

REVEL and ClinPred scores lower than 0.5 are defined as tolerant and greater than 0.5 as damaging. We show proportions of benign and pathogenic variants that were classified as Tolerated (T, Green) and Damaging (D, Pink). ClinPred had the best performance in finding as many pathogenic variants possible while minimizing the number of benign variants that are predicted as damaging both in ClinVarTest with AF<0.01 (A) and MouseVariSNP with AF<0.01 (B). Adapted from Alirezaie et al., The American Journal of Human Genetics, 2017.

5.2.5. Investigating generalizability of ClinPred to different disease mechanisms

Further, we examined if our algorithm's performance differs between mutations resulting in gain or loss of function, in either rare disease or cancer. Similarly, to the first and second test datasets, we calculated AUC and sensitivity for GainFunction, LossFunction, TSG and Oncogene test data. As demonstrated in Figure 5.10, ClinPred performance remained robust across all four test-datasets. Moreover, ClinPred retains the highest sensitivity to predict pathogenic variants among other tools (Figure 5.10).

Since the DoCM database consists of only pathogenic variants, we could only compile sensitivity scores based on the categorical predictions provided by the tools. (Table 5.5). ClinPred could successfully predict pathogenic variants in cancer, achieving a sensitivity score equal to 94.02 percent.

Table 5. 5. Overview of performance of ClinPred in comparison to categorical scores of other tools in DoCM test.

Adapted from Alirezaie et al., The American Journal of Human Genetics, 2017.

	NA/Pathogenic	TPR sensitivity	FNR
cforest	0	0.89	0.10
xgboost	0	0.91	0.08
ClinPred	0	0.94	0.05
REVEL	0	0.83	0.16
M-CAP	12	0.95	0.04
MetaLR	0	0.67	0.32
Fathmm_mkl	0	0.97	0.02

NA/pathogenic: Number of pathogenic variants with missing data
TPR: True positive rate
FNR: False negative rate

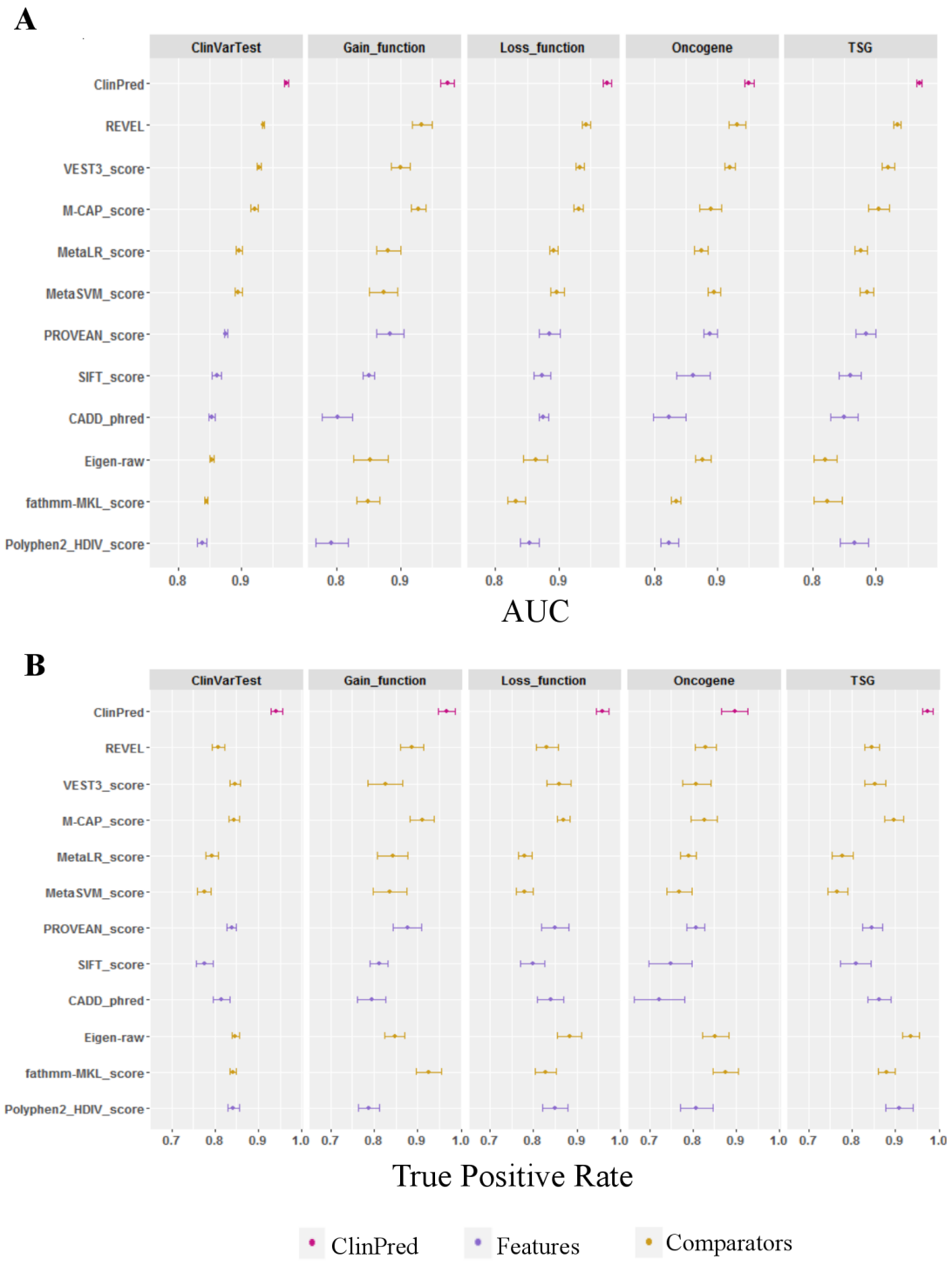


Figure 5. 10. AUC and sensitivity score were compared in five datasets.

Error bars show data for 5-fold cross validation. We observed that AUC is in agreement for all these datasets regardless of type of the variants. Our method yield state of the art sensitivity in most of the datasets we analyzed. Adapted from Alirezaie et al., The American Journal of Human Genetics, 2017.

5.2.6. Application of ClinPred to patient data

Finally we evaluated the performance of ClinPred in comparison to commonly used predictors, SIFT, Polyphen2, CADD as well as the recent ensemble predictors, MetaSVM, MetaLR, REVEL, M-CAP and VAAST Variant Prioritizer (VVP) (Flygare, et al. 2018) in 31 exomes from the FORGE Canada and Care4Rare Canada projects. To compare categorical scores, variants were categorized as pathogenic if they were predicted as pathogenic/probably pathogenic (Polyphen2) or damaging (SIFT, MetaSVM, MetaLR, M-CAP). Since CADD authors did not provide a categorical score, we defined pathogenic variants according to different CADD_PHRED score cutoffs (more than 10, 15 and 20). We considered any score higher than 0.5 as pathogenic in ClinPred and REVEL, and any score higher than 50 as pathogenic in VVP.

After typical quality filtering (Beaulieu, et al. 2014) a patient's exome on average harbored 433 non-synonymous variants (AF<0.05 in ExAC). There were 25 different nonsynonymous variants with strong supporting evidence for being causative in these samples. All studies have been published in peer-reviewed journals or are in press. In this analysis, we defined the sensitivity of a predictor as the number of known causative variants that were predicted as pathogenic, divided by 25 (the total number of known causative variants). Although each exome likely contains other pathogenic variants, in addition to those that cause the disease, we aimed to identify the prediction tools that selected the highest number of the 25 known disease variants, while discarding the highest proportion of the remaining variants.

As demonstrated in Figure 5.11, sensitivity scores among tools ranged from 44 to 100 percent, with the highest achieved by CADD and VVP for homozygous genotype (hom-VVP). Although CADD and hom-VVP identified all the causative variants as pathogenic, this came with the cost of low specificity: on average 94, 75, 60 and 50% of non-synonymous variants per exome

were predicted as pathogenic using different hom-VVP and CADD_PHRED cutoffs (more than 10, 15 and 20 respectively). While hom-VVP predicted all pathogenic variants as deleterious, VVP for heterozygous genotype (het-VVP) missed three heterozygote variants. ClinPred predicted 24/25=96% of the causative variants as pathogenic (Figure 5.12). The only variant missed by ClinPred had a marginal score of 0.449 and was found in late onset patients with compound heterozygote variants in the same gene - one frameshift and the other nonsynonymous.(Hoch, et al. 2017)

Further, we ranked the variants based on their ClinPred scores in any of these clinical exomes to investigate application of ClinPred in the filtering process. The median ranking of causative variants was 10 with the causative variant ranked as the first one in three cases. 34.5%, 52%, 66%, 83% of true positives were ranked on the top 5, 10, 15 and 25 variant ranks respectively. Only 18% of the causative variants ranked over 25; these variants were mostly in compound-heterozygote condition with other frameshift variant or top ranked heterozygote one.

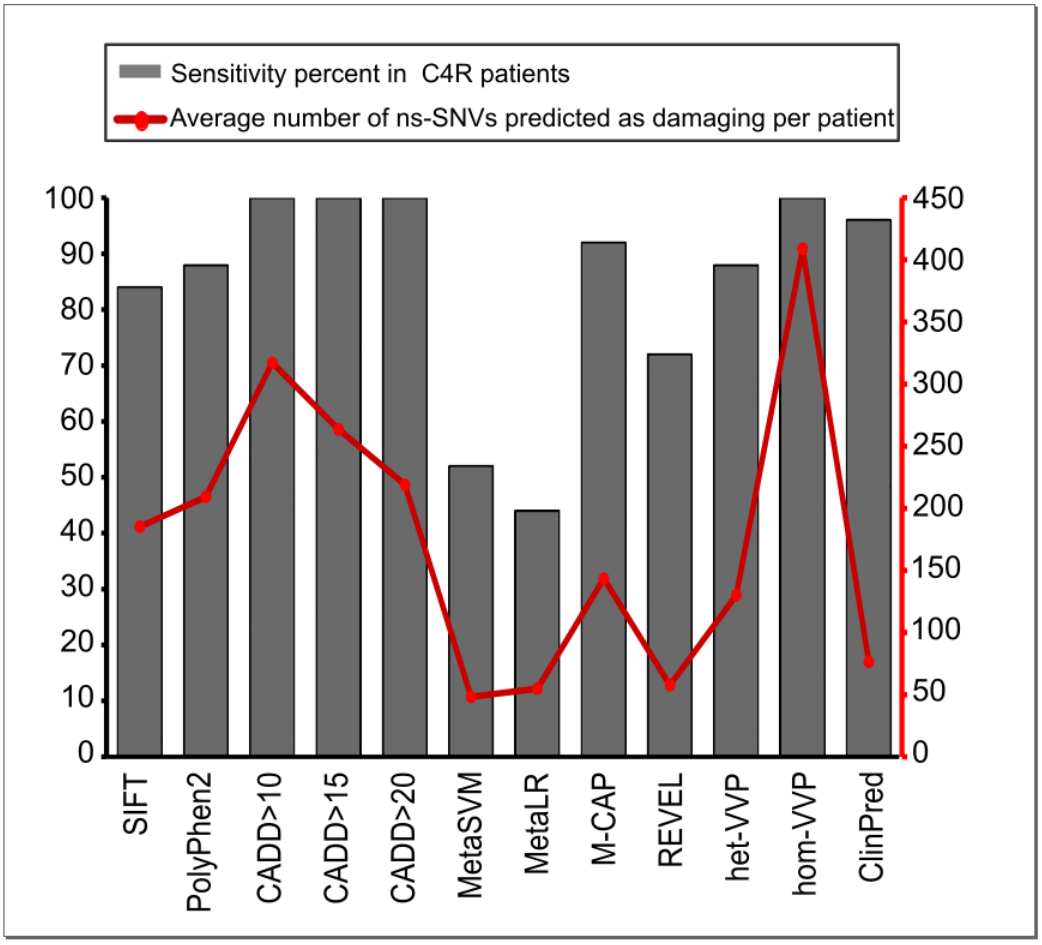


Figure 5. 11. Illustration of performance of ClinPred as compared to other tools on Care4Rare Canada project samples.

ClinPred reduced the number of nonsynonymous variants predicted as pathogenic and retained high sensitivity. Adapted from Alirezaie et al., The American Journal of Human Genetics, 2017.

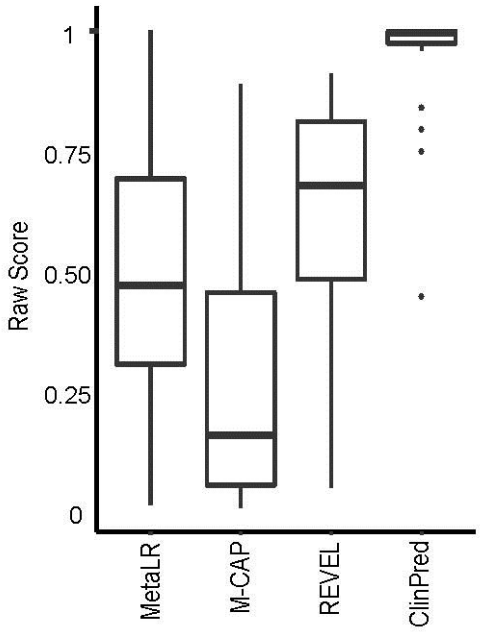


Figure 5. 12. Comparison of raw scores of MetaLR, M-CAP, REVEL and ClinPred for FORGE Canada and Care4Rare Canada projects cases

Adapted from Alirezaie et al., The American Journal of Human Genetics, 2017.

5.2.7. Assessing concordance between functional assay and computational prediction scores

Since ACMG guideline suggest the result of well-established in vitro or in vivo functional study as evidence for variant interpretation, we examined how our algorithm and other computational prediction methods' performance matches functional assay data. A recent study on large-scale functional classification of BRCA1 variants provides an excellent opportunity for such comparison. (Findlay, et al. 2018) While most of the computational methods had the ability to predict LOF variants in the BRCA1 dataset as pathogenic/deleterious (sensitivity ranged from 92.6 to 100 %), their performance was poor in predicting functional variants as benign (specificity ranged from 0.1 to 46% with the lowest in M-CAP and the best in MetaSVM). ClinPred predicted 97.5% of LOF variants as pathogenic and 32% of functional variants as benign (Figure 5.13).

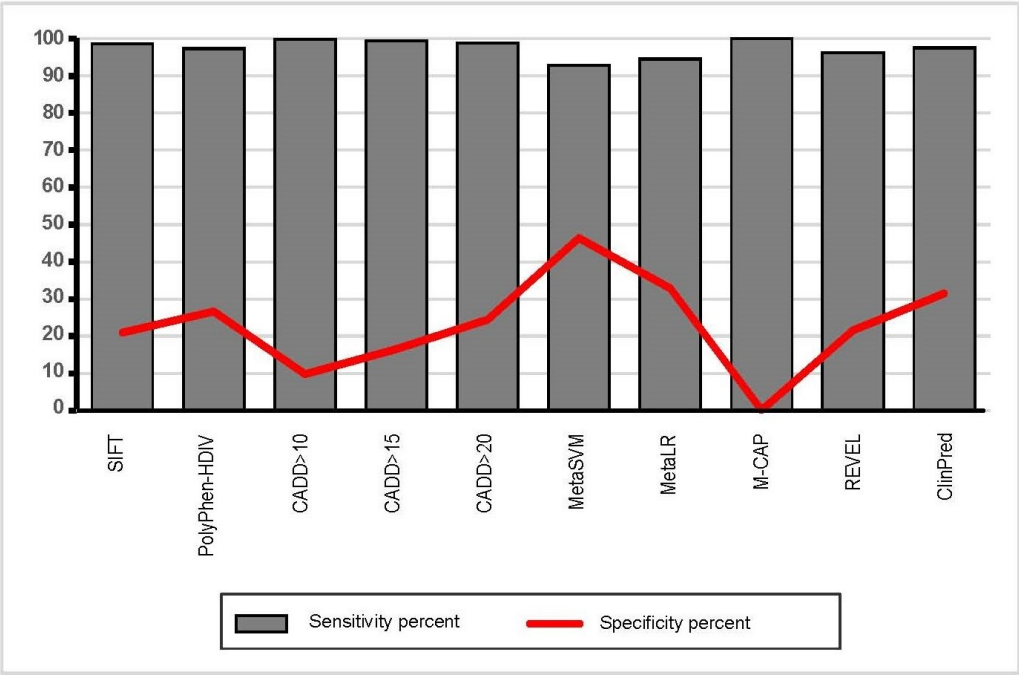


Figure 5. 13. Illustration of performance of ClinPred in comparison to other tools in BRCA1 dataset.

Sensitivity and specificity of each tool were compared based on the categorical scores of each tool. Adapted from Alirezaie et al., The American Journal of Human Genetics, 2017.

5.3. Conclusion

In this study, we used an improved supervised machine learning approach to create ClinPred, a method to efficiently distinguish clinically pathogenic from neutral variants. The first improvement concerns the choice of the most accurate training dataset: we train our predictor on clinically significant variants based on the joint consensus recommendation for the interpretation of sequence variants by the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP). Secondly, we apply two different machine-learning algorithms and include a wide range of recent supervised and unsupervised methods as predictive features. Finally, we identify allele frequency as one of our key predictive features, which improves performance both in the presence and in absence of a predetermined frequency threshold for inclusion of variants in the analysis.

Compared to other methods, ClinPred showed highest sensitivity, improved specificity, and obtained the best performance according to various performance metrics. We also found that our predictor maintains consistently superior performance across different genetic models and pathogenic mechanisms – for example dominant versus recessive or oncogene versus tumor suppressor classifications.

In our illustration of real-life utility with respect to clinical data, the FORGE Canada and Care4Rare Canada cases were selected from studies published after mid-2015 to avoid overlap with any of the training data. Applying our predictor as one of the selection criteria for pathogenicity would reduce the list of an average 443 non-synonymous in individual's exome to an average of 70 variants to be further manually followed up. In most cases, the entire list of selected variants - prioritized by prediction scores - would not have to be examined, as 83% of causative variants ranked within the top 25 candidates. Out of the 25 distinct disease-causing

variants, ClinPred misclassified only 1 causative variant as benign. This variant had a marginal score of 0.449 and was found in late onset case with compound heterozygote variants in the same gene - one frameshift and the other nonsynonymous (Hoch, et al. 2017). While this illustrates the pitfall of using classifiers such as ClinPred for purely automated filtering, it also suggests to a possible alternative multi-stage protocol, where patients can first be scanned using the currently optimized strict score cutoff and, if no definite disease cause is identified, the criteria can be relaxed and re-applied. The only approaches that succeeded in identifying all of the 25 pathogenic variants in this dataset were CADD and hom-VVP, but this came at the cost of specificity, and as a result, their application could only narrow down the candidate list to an average of 216 and 409 variants per case for CADD and hom-VVP respectively. While increasing the threshold of any predictor score results in higher sensitivity, it will jeopardize specificity. In the clinical domain, a test achieving 95% sensitivity with high specificity is generally favorable. Across our test data, ClinPred was able to achieve 95% sensitivity with the best specificity among other tools.

The relatively low specificity of computational predictions in the BRCA1 dataset may be at least partly due to the limited sensitivity of the in vitro assays used in that study. Functional scores in the BRCA1 dataset were measured based on cellular fitness in a haploid human cell line, which may not fully reflect the function in the complete organism. Such discrepancy between in vitro and in vivo BRCA1 mutant homologous recombination activity has previously been demonstrated (Drost, et al. 2011; Millot, et al. 2012). Conversely, computational predictions may be overestimating pathogenicity. At this point, we conclude that the relative strengths of computational predictions and in vitro functional assays warrant further investigation.

Our results also demonstrate the value of combining different methods that likely provide complementary information as a result of their divergent algorithms and training datasets. The

importance of diversifying training datasets is illustrated by comparing Polyphen2 HVAR and HDIV scores, where their difference in performance owes to the fact they applied different training dataset in spite of the same algorithm. On the other hand, illustrating difference in methodologies, DANN and CADD shared the same training data with different algorithm resulting in disjoint performance.

In our design, we took great care to avoid type-I circularity, (Grimm, et al. 2015) a problem that occurs in supervised machine learning when the training data directly or indirectly overlap with test data. Although in our model we eliminate any such overlap to prevent over-fitting, comparison against other tools is not completely free of bias. First, it is practically impossible to remove from our dataset the neutral variants that were used to train other ensemble tools, such as large number of variants present in the ExAC database. Second, M-CAP, VEST 3 and REVEL were trained on private pathogenic datasets which we were unable to access, and which may influence comparison of performance in their favor.

Finally, we provide pre-computed ClinPred scores for all possible human variants through our website (http://hubs.hpc.mcgill.ca/~alirezai/ClinPred) to facilitate its use in general practice.

Chapter 6: General discussion

In the past years, exome sequencing has become one of the most commonly used next-generation sequencing technique. After significant success of whole exome sequencing in the research area and identifying ~160 new disease genes annually, diagnostic and clinical genetics fields started to use this technique (Boycott, et al. 2017).

In this thesis, I first reported a large-scale NGS study on FPC patients and demonstrated how using WES as the primary investigative tool led to discovery of novel candidate genes in this lethal disease. Identifying germline mutations in pancreatic cancer patients not only will facilitate management of patients with a positive family history, but also help to screen them at an early stage. Knowledge of predisposition genes can also predict response to specific therapies and guide to better treatment. In Chapter 4, by performing WES in multiple TNBC tumors, I showed how WES in combination with other sequencing techniques, identified more than two-fold amplification in RAD21 gene on chr8q in drug-resistant TNBCs. This finding suggests that amplified RAD21 may be both a marker and a target to overcome drug resistance in TNBCs.

While WES is successful in both research and clinics, the significant challenge, which is clinical translation of exome sequencing and efficient variant interpretation, remains. In chapter five, I addressed this important problem in clinical genomics—how to identify the most likely pathogenic variant(s) responsible for a disease phenotype when analyzing variant data obtained by whole exome sequencing for diagnostics. I developed ClinPred, a method to predict clinically related variants. ClinPred showed superior accuracy for predicting pathogenicity, achieving the highest area under the curve (AUC) score and increasing both the specificity and sensitivity in different test datasets. It also obtained the best performance according to various other metrics.

Moreover, ClinPred performance remained robust with respect to disease type (cancer or rare disease) and mechanism (gain or loss of function). I provide pre-computed ClinPred scores for all possible human missense variants in the exome to facilitate its use by the community.

6.1. New pancreatic cancer genes

6.1.1. NEK1

NEK1 [NIMA (Never in Mitosis Gene A)-Related Kinase 1] is involved in DNA damage checkpoint control. Its protein product is a dual serine-threonine and tyrosine kinase required for proper DNA damage repair. Dysfunctional NEK1 fails to activate Chk1 and Chk2 kinases, and consequently G_1/S or M-phase checkpoints fail to stop in response to DNA damage, resulting in unstable chromosomes (Chen, et al. 2011b). Besides its role in DNA damage checkpoint control, NEK1 is involved in centrosomal function and is suggested to be a tumor suppressor (Chen, et al. 2011a).

In my findings, two of 93 high-risk PC families had a novel NEK1:p.Ala563Tyrfs*36 variant, which was not seen either in public databases such as gnomAD or in our in-house control database. Interestingly one of the families had Greek origin (family 17) and the other one was of English and Scottish descent (family 78). As a result, it is unlikely that this variant is ethnic-specific. It was concluded that this mutation is most likely a "hot spot" in NEK1. Segregation of the variant was observed in the English-Scottish family, but segregation analysis showed only partial segregation in the Greek family. Since the only affected patient with the wild-type NEK1 variant was the oldest diagnosed in this family (75 years of age), phenocopy is most likely the cause of PC in this patient. In the third family (89), the p.Asn648Lys variant segregated in the two PC-affected siblings.

One of the interesting facts about NEK1 that makes it more appealing is its chromosomal location. NEK1 maps to chromosomal region 4q33. Previous linkage analysis findings on a study of a family with 9 PC-affected members suggested locus (4q32-34) as the PC susceptibility

chromosomal region (Eberle, et al. 2002). Further, Pogue-Geile et al. sequenced candidate genes in this region and identified PALLD (P239S) as the causative mutation in this family (Pogue-Geile, et al. 2006). Further studies could not confirm the role of PALLD in other families with pancreatic cancer, suggesting it cannot be a common PC susceptibility gene (Klein, et al. 2009; Salaria, et al. 2007).

Even though Pogue-Geile et al did not detect a NEK1 variant in the aforementioned family, the role of NEK1 cannot be ruled out, as they could not detect large genomic structural changes at this locus. Therefore, there is a possibility that NEK1 was the true underlying PC predisposition gene even in that family.

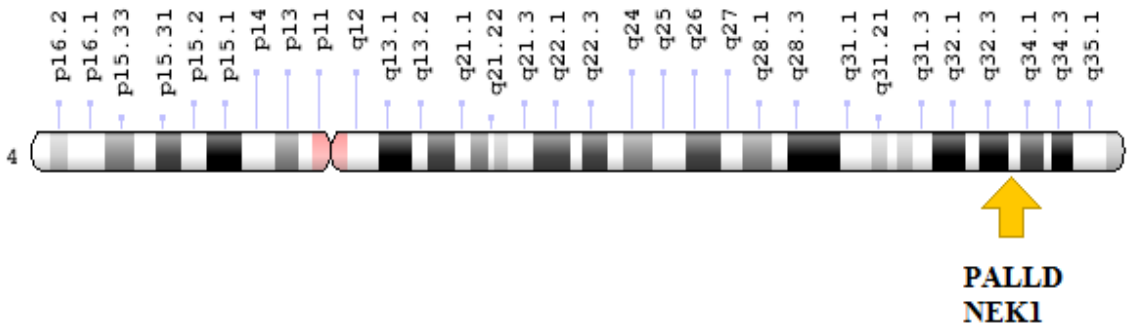


Figure 6. 1. Chromosomal location of PALLD and NEK1

6.1.2. FAN1

FAN1 (FANCD2/FANCI-associated nuclease 1) is the next strongest candidate gene. FAN1 plays a role in DNA interstrand cross-link repair and acts as a tumor suppressor by preventing genomic instability. Interestingly, recent research on more than 176 families with hereditary colorectal cancer reported germline mutations in the FAN1 gene as a putative colon cancer susceptibility gene (Segui, et al. 2015).

In our study, I found FAN1 mutation in three families: a PTV in FAN1 (p.Arg710*) in Family 42 as well as a missense variant (p.Met50Arg) in two other families. The missense variant (p.Met50Arg) was located in a very conserved region within the RAD18-like ubiquitin-binding (UBZ) domain. This area is essential for FAN1 localization to sites of DNA damage. This variant also showed complete co-segregation with pancreatic cancer in tested family members. Although I could not find LOH in the analysis of the two tumors from carriers of this variant, this finding is consistent with Segui et al.'s report, in which somatic inactivation of the wild-type allele was not detected in carriers of germline FAN1 mutations with colon cancer (Segui, et al. 2015). Unfortunately, DNA from other family members was not available to determine segregation or LOH of the p.Arg710* variant.

6.1.3. RHNO1

The third top ranking candidate gene is RHNO1 (Rad9-Hus1-Rad1 Interacting Nuclear Orphan 1). RHNO1 plays an important role in DNA damage response (DDR) signalling in S phase (Cotta-Ramusino, et al. 2011). In this study, I observed two different PTVs in RHNO1: the p.Arg84* variant in family 43 and the p.Arg113* variant in family 18.

Although we did not see segregation of the p.Arg84* variant in the second family member, we found LOH of the wild-type allele in the proband tumor. Segregation of the p.Arg113* variant was observed in both PC-affected family members. A third family with RHINO1 mutation was a carrier of the p.Leu16Val variant. This variant was predicted to be pathogenic by multiple in silico software methods and segregated in the family. This variant was previously reported in COSMIC database in thyroid cancer (COSM4146987).

6.1.4. Other Candidates

In this study, I found other noteworthy variants that can be potentially important in pancreatic cancer development. Most of these variants were located in genes implicated in other hereditary cancer syndromes. In particular, FANCG and FANCL genes (Fanconi Anemia genes) are important since HDR gene association with PC was reported in several studies (Couch, et al. 2005; Slavin, et al. 2018; van der Heijden, et al. 2004). Roberts et al also found multiple Fanconi anemia PTVs in analysing 638 WGS familial pancreatic cancer patients. In their study, they found four FANCG carriers besides PALB2 (FANCN), FANCC, and FANCM (Roberts, et al. 2016). Another publication by Witkiewicz et al. surveyed 109 micro-dissected pancreatic ductal adenocarcinoma cases; reported multiple Fanconi anemia genes were mutated or deleted in relatively high frequency in these samples (Witkiewicz, et al. 2015). A non-synonymous variant previously associated with Fanconi Anemia in FANCG was previously described in a cell line derived from early onset PC and demonstrated LOH (van der Heijden, et al. 2003). Therefore, I propose further investigation of the role of Fanconi anemia genes as pancreatic cancer susceptibility genes.

The other noteworthy gene is POLQ. This gene plays a key role in the microhomology-mediated end-joining of double stranded DNA breaks and has been suggested as a potential target for synthetic lethality in HDR-deficient tumors (Ceccaldi, et al. 2015; Mateos-Gomez, et al. 2015). In this study, one PTV and three predicted-pathogenic missense variants in five samples were found. However, the variants did not segregate with PC in the four families that were tested.

BLM (involved in hereditary breast cancer), and BARD1 (involved in hereditary breast and ovarian cancers) are other mutated genes worth mentioning and following up in larger-scale studies.

6.1.5. Limitations of this study

One of the main limitations of the study is the application of the filter-based candidate gene approach with special consideration on DNA repair genes, although other genes not in the DNA repair list may be involved in PC. In addition, large genomic deletions and rearrangements were not detected; therefore, the possibility of such variants in known and candidate PC predisposition genes cannot be excluded. Importantly, I had to create fixed criteria for defining pathogenic variants based on three available pathogenicity prediction methods, therefore some candidates that were not followed these criteria might be missed.

Another challenge in identifying genetic predisposition factors in hereditary PC is the phenomenon of phenocopies. Indeed most of the FPC patients were of the same age at onset as the sporadic ones (Brune, et al. 2010). Lack of segregation of the PALB2:c.3256C>T (p.Arg1086*) and ATM:c.1931C>A (p.Ser644*) PTVs with PC-affected relatives was reported in

previous studies. Therefore, I did not exclude any candidate gene due to partial segregation (Grant, et al. 2013).

Since there are other mechanisms of somatic loss of the wild-type allele, similarly to segregation, LOH was used as a criterion for prioritization rather than exclusion. I did not even observe LOH in the BRCA2 (p.Thr1566Aspfs*9), which is a known FPC susceptibility gene, in one sample with BRCA2 germline mutation, suggesting other mechanisms of wild-type allele silencing in the tumour. In addition, there is a possibility that haploinsufficiency is sufficient for tumorigenesis.

The other challenges in identifying causative genes in hereditary PC are genetic heterogeneity and variable penetrance of disease-causing alleles. In the study, multiple predicted-pathogenic variants in DNA repair genes in a single individual were observed. This indicates the possibility of their "additive" haploinsufficiency effect as well as variable penetrance of disease-causing alleles. This double heterozygosity of pathogenic variants in different cancer predisposition genes has been reported before in breast cancer (Bell, et al. 2002; Sokolenko, et al. 2014).

Although confirmatory evidence for the candidate genes could not be provided, based on available family information, segregation and LOH analysis, I suggest FAN1, NEK1 and RHNO1 as the strongest candidates for further validation using additional FPC samples.

6.2. Resistance to neoadjuvant therapy in TNBC

Triple negative breast cancer (TNBC) is one of the more aggressive types of breast cancers. Unfortunately, there is no specific molecular target therapy available for patients diagnosed with this type of disease. Therefore, cytotoxic chemotherapy is the only standard and effective treatment for these patients. Prognosis in TNBC patients is highly correlated to the response to chemotherapy. In order to provide better treatment to TNBC patients, the molecular mechanisms of response and resistance to chemotherapy should first be understood. Therefore, in our study, extensive molecular analysis of pre- and post-chemotherapy tumor samples was performed to identify mechanisms or markers of resistance and/or sensitivity to chemotherapy.

The underlying hypothesis of the TNBC project was that chemotherapy either would induce chemo-resistance or lead to selection of a subclone that was already resistant to chemotherapy. Detailed analysis of array CGH, RNAseq and whole exome sequencing (WES) was performed on tumor samples obtained from 29 patients. The results suggested that there are two major driving factors involved in response to chemotherapy. 1: Immune response 2: the amplification of chromosome 8q23, specifically a segment containing 3 genes (RAD21, EIF3H and UTP23) whose expression strongly correlates with gene amplification. In particular, RAD21 has been previously reported to be associated with chemo-resistance in the MDA-MB-231 triple negative breast-cancer cell line (Xu, et al. 2011). Interestingly chemo-resistant tumors without chromosome 8 amplification were strongly enriched for high RNA levels of immune response genes.

WES analysis in these samples emphasized the genomic heterogeneity of TNBC tumors. In fact, the only truly shared somatic mutation was in TP53. This finding is concordant with a recently published article (Kim, et al. 2018). I observed shifts in somatic variants, with a minority

of variants being gained or lost during chemotherapy, but with no specific pattern. The presence of the "gained" / "lost" variants in the plasma samples detected by ddPCR suggests heterogeneity may be responsible for some of the changes in AF in detected variants. Similarly, in a study by Balko et al., who performed NGS sequencing using a 196-gene panel in 74 post-chemotherapy residual TNBCs, no significant increase in somatic mutations after chemotherapy was reported (Balko, et al. 2014).

Together with the array CGH data, the whole exome sequencing data suggests that the DNA of TNBCs remains stable during chemotherapy when the response to chemotherapy is incomplete. These findings are similar to those of a recent article that investigated chemo-resistance in TNBC by single-cell sequencing, indicating that chemo resistance exists in non-responders and is adaptively selected during chemotherapy (Kim, et al. 2018)

6.2.1. Is RAD21 the key?

Rad21 (double-strand-break repair protein rad21 homolog, also known as SCC1) is involved in homologous recombination–mediated double-strand break (DSB) repair and chromosome cohesion during the cell cycle. As well, it plays a role in cell cycle regulation and apoptosis (Ahn, et al. 2017). Together with SMC3, SMC1 and SCC3/SA, RAD21/SCC1 is one of the four subunits of the cohesion complex which is responsible for the cohesion of sister chromatids following DNA replication. Over the last decade, new roles for the cohesion complex have emerged. Multiple studies have shown cohesion-associated genes play roles in tissue-specific gene transcription (Rhodes, et al. 2011), cell proliferation and maintenance of pluripotency and act as potential drivers in tumors' genomic instability and progression (Yadav, et al. 2013). Ahn et al. suggested Rad21 interacts with mutant p53 to promote growth in ovarian cancer cells (Ahn, et al. 2017).

Schmidt et al reported that Rad21 co-localizes with estrogen receptors and suggested cohesion participates as an integral component of transcriptional regulatory networks and is required for efficient estrogen-dependent G0/G1–S phase transition in breast cancer cells (Schmidt, et al. 2010). Moreover, cohesion-associated genes' overexpression has been reported in multiple cancers and cancer cell lines (Deb, et al. 2014; Yadav, et al. 2013). In a study by Yadav et al., SMC1 was overexpressed in TNBC cell lines as compared to normal epithelial cancer cells.

Emerging evidence has also shown the involvement of RAD21 in chemo-resistance or response to chemotherapy. Nakashima et al. proposed the role of RAD21 in chemo-resistance to gemcitabine in biliary tract cancer (Nakashima, et al. 2015). Overexpression of RAD21 is also associated with aggressive colorectal carcinomas, especially in KRAS mutant tumors (Deb, et al. 2014).

In breast cancer, RAD21 has also been studied in multiple reports. Overexpression of both SMC1 and RAD21 was reported in the MDA-MB-453 cell line, and Atienza et al. showed RAD21 suppression increased cytotoxicity of etoposide and bleomycin in human breast cancer cells (Atienza, et al. 2005; Jeong, et al. 2012).

In this study, RAD21 located on chromosome 8q was amplified more than two fold in five of nine drug-resistant TNBCs, both in pre- and post-chemotherapy samples based on aCGH data. This gene was also amplified in 16-28% of breast cancer cases in TCGA.

In light of the above facts, I conclude that amplified RAD21 may be both a marker and a target to overcome drug resistance in TNBCs. This finding raises the possibility of developing novel cancer therapeutic strategies to overcome chemo-resistance.

6.2.2. Immune response and resistance

One of the factors that help cancer cells to relapse is their ability to escape from the immune system. Several studies have shown the role of the immune system in promoting death and response to therapies. In TNBCs, a study on the effect of four different chemotherapy drugs on TNBC cell lines showed inhibitory effects of chemotherapy on anti-tumor immunity (Samanta, et al. 2018).

Expression of programmed death-ligand 1 (PD-L1) is frequently reported in cancer cells and correlated with a poor clinical outcome (Powles, et al. 2014). Prior studies indicate the particular therapeutic effectiveness of anti-PD1 and anti-PDL1 antibodies in some types of cancers, such as melanoma and renal carcinoma. However, immunotherapy was not effective in all types of cancers (Samanta, et al. 2018). PD-L1 is a ligand of PD-1 (programmed cell death-1) and is commonly expressed on the surface of dendritic cells or macrophages. PD-1/PD-L1 plays a role in suppressing the immune system by inhibiting cytotoxic T cells and deactivating them (Chen and Mellman 2013). Although several studies associate PD-L1 with drug resistance and poor prognosis (Mori, et al. 2017; Zhang, et al. 2016), the value of PD-L1 as a biomarker in TNBC has been controversial. For example, a high level of PD-L1 has been correlated with tumor-infiltrating lymphocytes (TILs) and better response to neoadjuvant chemotherapy (Mittendorf, et al. 2014; Schalper, et al. 2014; Wimberly, et al. 2015).

TILs have been identified as predictors of response to neoadjuvant chemotherapy in all breast cancer subtypes (Denkert, et al. 2018), and their relation to pathological complete response in patients with TNBC has been reported in many studies (Herrero-Vicent, et al. 2017).

Indeed, in our study's results, overexpressed PD-L1 tumors that have high TILs had a good prognosis, while high PD-L1 and low TILs were associated with poor prognosis. This result agrees

with a recent study by Mori et al, who analyzed 248 TNBC patients and reported a strong correlation of PD-L1 expression with TILs, whereas PD-L1 expression and TILs were not independent prognostic factors. In their result PD-L1-positive/TILs-low tumors were associated with a poor prognosis although PD-L1-positive/TILs-high tumors had the best prognosis (Mori, et al. 2017). The same trend was reported in non-small-cell lung cancer patients treated with neoadjuvanct chemotherapy (Zhang, et al. 2016). The only sample with high PD-L1+TILS and poor prognosis in our samples was Neo31. Interestingly, this sample has RAD21 amplification.

The results of this study suggest combination of chemotherapy and immunotherapy (anti-PD-1/PD-L1 monoclonal antibody therapies) may improve outcomes of TNBC patients, especially patients with PD-L1-positive/TILs-low tumors.

6.2.3. Limitations of this study

There are several limitations in this study. First, we could not have matched pre- and post-chemotherapy samples for all patients, and the sample size was limited. Second, patients received different types of chemotherapy; therefore, we could not differentiate the genomic changes related to different drugs and the impact of genomic changes on sensitivity to the different drugs. Third, whole genome sequencing or epigenetic studies were not performed; therefore, the role of epigenetic and non-exonic regions cannot be ruled out. Finally, TNBC has different subtypes that might respond to chemotherapy differently. Due to the small sample size, different responses in different TNBC subtypes could not be investigated.

6.3. New insight in pathogenicity prediction

One of the important problems in clinical genomics is how to identify the most likely pathogenic variant(s) responsible for a disease phenotype when analyzing variant data obtained by whole-genome or exome sequencing for diagnostics, specially the effect of non-synonymous variants. In this endeavor, clinicians typically examine a subset of the patient genome/exome variants that have been filtered by several criteria and/or prioritized by algorithms that aim to predict pathogenicity of the variant with respect to patient phenotypes. Prior clinical information (e.g. from databases where curated variants can be accessed, such as ClinVar) is commonly used as evidence. However, lacking this prior knowledge, previously unassessed variants need to be classified as of either benign, likely pathogenic, or unknown significance, with the aid of methods that predict the deleteriousness of variants. This process can be time consuming, and existing predictive measures of non-synonymous variant pathogenicity are not powerful enough alone to inform diagnostic decisions. Therefore, tools with a higher ability to distinguish between pathogenic and neutral variants will be beneficial for future precision medicine, and intense research is needed to increase these tools' reliability and utility.

In this study, I used an improved supervised machine learning approach to create ClinPred, a method to efficiently distinguish clinically pathogenic from neutral variants. I showed that choosing the most accurate training dataset, such as ClinVar, improves the ability of the predictor. In addition, I identified allele frequency as one of the key predictive features, which improves performance both in the presence and in absence of a predetermined frequency threshold for inclusion of variants in the analysis.

Although, traditionally, AF is employed to discard benign variants, it is unclear what threshold should be selected at which a variant is considered benign. Many investigators use a cutoff of 5%, which is the upper bound for carrier frequency of most common Mendelian diseases, such as cystic fibrosis. However, in view of rarity of many other phenotypes, researchers often select lower AF cutoffs (Kobayashi, et al. 2017). In designing ClinPred, I did not set any restrictions regarding the AFs of either pathogenic or benign variants, and I allowed the algorithms to learn the best use of this feature as a predictor. This contrasts with other methods, where the benign variants are often selected on the basis of certain AFs (Ioannidis, et al. 2016; Jagadeesh, et al. 2016). As examples, M-CAP considers any variant with a mean allele frequency ≤ 1% in ExAC and 1000 Genomes as benign, while REVEL selects variants with AFs between 0.1% and 1% across the seven study populations for their benign label. In my approach, I utilized AFs from the largest database available, gnomAD, as one of the predictor variables and allowed the model to learn the optimal parameters, without using a specific threshold. Some other existing methods have also incorporated AF in their approach. For example, MetaLR, MetaSVM and Eigen applied AF from 1000 Genome database in their model. M-CAP indirectly benefits from AF by using MetaLR, MetaSVM as their feature. The relatively lower level of success in using AF in those methods may be due to high missing values for AF in the smaller, less representative databases. As far as I know, Gavin and VVP are the only methods that use AFs from a large database; however, my method is different from theirs. Gavin applied AFs equal to 0.00346 in ExAC and CADD>15 as the fixed thresholds for defining variants as pathogenic (van der Velde, et al. 2017). VVP incorporated population variant frequencies from the WGS portion of gnomAD (15,496 whole genomes), but I incorporated not only gnomAD all exome AFs (123,136 exome sequences), but also AFs in 8 different populations available in gnomAD: African/African American, Latino,

Ashkenazi Jewish, East Asian, Finnish, Non-Finnish European, South Asian, and other ethnicities. A large part of the increase in performance of ClinPred is attributed to allowing the classifier to learn and optimize the use of AFs in making the distinction between pathogenic and benign variants.

I also found that our predictor maintains consistently superior performance across different genetic models and pathogenic mechanisms—for example, dominant versus recessive or oncogene versus tumor suppressor classifications. However, it should be noted that this outcome is highly dependent on the types and proportions of variants that are currently present in the disease databases. The currently catalogued pathogenic variants are predominantly highly penetrant monogenic or oncogenic mutations, generally with severe disease phenotypes that are strongly selected against in human populations.

In applying our predictor to real samples (the FORGE Canada and Care4Rare Canada), our predictor reduced the list of an average 443 non-synonymous in patient exome to an average of 70 variants, to be further manually followed up. It is important that the prediction method that ranks the causative variant on the top level is the most favorable. To address this, we ranked the variants according to their ClinPred scores for 25 causative variants in the 31 clinical exomes available. The median of ranking of causative variants was 10. In most cases, the entire list of selected variants prioritized by prediction scores would not have to be examined, as 83% of causative variants ranked within the top 25 candidates.

In summary, we systematically compared both categorical prediction and raw scores of different commonly used methods under different AF cutoffs that might typically be used by researchers and clinicians to narrow down lists of variants. ClinPred outperformed all existing ensemble classifiers in distinguishing between disease-relevant pathogenic from neutral variants.

Our model generalizes well when applied to variants from various sources not included in its training dataset. It also has high performance both in rare diseases and in cancer. I provide pre-computed prediction scores for all possible variants in the human exome to facilitate interpretation of high throughput sequencing results.

Chapter 7: Conclusions and future directions

In the first part of this thesis, we undertook a large-scale WES study focused on germline variants in putative DNA repair genes aimed at identifying novel genetic causes of hereditary PC in 109 selected cases with increased risk of genetic PC predisposition from 93 families. By applying a filter-based candidate gene approach focused on 513 DNA repair genes, we found PTVs in 41 putative DNA repair genes among 36 families. We then prioritized our list of 41 candidate genes based on the evidence obtained from segregation analysis and LOH. Our top 17 candidate genes were those with more than one family with a PTV in that gene, genes with segregation of a predicted-pathogenic variant (PTV or nonsynonymous variant) in at least one family, and/or genes with LOH associated with at least one predicted-pathogenic variant. Our findings suggest that several novel DNA repair genes may have a role in hereditary PC and that genetic susceptibility of hereditary PC is highly heterogeneous. Amongst the set of 17 candidates, FAN1, NEK1 and RHNO1 were considered as the top ranked candidates, having variants present in three families and at least partial co-segregation of the variants with PC in two or more families. We propose these three genes for further validation using additional families with PC.

In the second part of this thesis, we investigated the mechanisms underlying resistance to chemotherapy in triple-negative breast cancer (TNBC), which accounts for a large proportion of all breast cancer mortality. Fresh frozen biopsies prior to and after standard anthracycline/taxane-based chemotherapy treatment were taken for whole exome sequencing, RNAseq and array CGH analysis. WES identified multiple single nucleotide variants (SNVs) in the samples, with few variants that were commonly shared apart from TP53 gene mutations. Although the degree of response to chemotherapy in RCB2/3 tumors was associated with more changes in somatic variants

in post-chemo compared to pre-chemo samples, no variant was associated with tumor response. We found amplification of the RAD21 gene on chromosome 8q23 on chemo-resistant tumors through copy number alteration analysis. RNA-seq analysis detected novel gene fusions in post-chemotherapy samples, including one involving the ABCB1 gene. Moreover, RNAseq analysis and pathological analysis of pre-chemotherapy tumors showed that the presence of immune-response genes was strongly associated with RCB0/1. We concluded that RAD21 gene amplification as well as immune response genes are predictive of response to chemotherapy in TNBCs in the neoadjuvant setting.

Furthermore, motivated by accurately identifying pathogenic variants from WES data, in the third part of the thesis, I addressed the current problems with pathogenicity prediction tools, especially the need for improved methods to predict clinically relevant variants. I developed an ensemble classifier called ClinPred for predicting disease relevance of missense SNVs, using a combination of two different machine learning algorithms and incorporating several popular pathogenicity predictors, along with population allele frequencies, as component features. Through rigorous testing, while avoiding problems common in machine learning, such as overfitting and circularity, I showed that ClinPred outperforms all existing ensemble classifiers in distinguishing between disease-relevant pathogenic variants and neutral variants. Our model can be applied to a range of diseases and has high performance both in rare disease and in cancer. We provide pre-computed prediction scores for all possible variants in the human exome to facilitate interpretation of high throughput sequencing results.

7.1. Challenges and future direction

Although whole-exome sequencing is a powerful tool for identifying disease-causing variants, some problems remain. WES is successful in detecting SNVs and small INDELs; however, it has difficulty identifying large INDELs. There are different approaches to detect genomic rearrangements, such as read depth (RD), allele frequency (AF), paired-end mapping, de novo assembly and split read (Alkodsi, et al. 2015; Tan, et al. 2014). WES CNV detection tools usually use the first two approaches, but RD-based approaches are not able to detect copy-neutral genomic rearrangements. In addition, WES is limited to finding rearrangements in exonic regions; therefore, whole genome sequencing is more reliable than WES in detecting CNVs.

In addition to its difficulty in detecting CNVs, WES analysis has limitations when it comes to variant detection in cancer tumours. Some cancer samples are heterogeneous and have subclonal structures (Navin, et al. 2010); therefore, detecting low frequency variants responsible for a specific subclone is challenging. Moreover, the molecular background of tumour samples in different patients with the same diagnosis can be very different. In our TNBC study, we noticed these samples were quite heterogeneous. Even though combining WES data, RNAseq data and aCGH increased our ability to explore drug resistance in these samples—emphasizing the power of using different tools to look at the data in different ways—access to more samples will provide the opportunity to investigate different subtypes individually and will definitely improve the results of the research.

Another problem in cancer research is that DNA extracted from cancer specimens is not pure and is contaminated with non-cancerous cells. Most studies in cancer genomics are limited by the availability of fresh tumor samples; however, formalin-fixed and paraffin-embedded (FFPE) specimens that are used routinely for diagnosis can be a source for more samples. The

problem in using FFPE samples is that they are often low quality (Gilbert, et al. 2007). Bioinformatics tools were developed to address this challenge and to detect low frequency variants; however, they still need to be improved. Analyzing FFPE data in an efficient way will provide more opportunities to draft large-scale studies and consequently will improve cancer research. In our study of identifying novel pancreatic cancer variants, WES data from tumor samples was helpful in prioritizing genes. Access to more cancer samples will enable setting up larger-scale studies and significantly benefit cancer research.

In both of our cancer studies, we were limited to coding regions, but some non-coding variants or epigenetic changes could have an impact on cancer development. A considerable number of cancer-related mutations have been identified in non-coding regions; however, these mutations can only be detected by WGS. As the technology continues to improve, the cost of WGS will drop. With further advancement in computing, WGS will become cheaper and more convenient. Therefore, it is expected that WGS will replace WES in the near future. In addition, by developing better CNV calling algorithms, WGS may eventually replace CNV tests (such as aCGH) currently being used. The lessons learned from analysis of WES data will directly apply to future NGS approaches, especially WGS.

In fact, for a long time, cancer was considered a disease of the genome. Although genetic mutations or indels were demonstrated as the main cause of hereditary cancers, a growing body of evidence indicated the role of epigenetics in many cancers. Genetic mutations that change epigenetic regulatory alterations have been linked to many types of cancers (Baylin and Jones 2011). DNA hypermethylation in tumor suppressor genes, which results in the silencing of these genes, and DNA hypomethylation in oncogenes were suggested as tumorigenic factors (Jiang, et al. 2009; van Doorn, et al. 2005). Our group determined that mutations in histone gene H3F3A

result in a defect in chromatin and causes pediatric glioblastoma multiforme (Schwartzentruber, et al. 2012). Besides the epigenetic changes involved in cancer development, a large number of resistance mechanisms are linked to the epigenetic domain. Therefore, epigenetics can be a predictor of treatment outcome (Hu and Baeg 2017). Although over the years, the technology has rapidly improved and the sequencing platforms have changed dramatically, we are still far from understanding the full picture of how genetic alterations interact with epigenetic regulations and lead to cancer development and resistance. Future improvement in next generation sequencing techniques and bioinformatics tools will help us understand this process better. In our TNBC study, we did not have enough DNA for epigenetic analysis. Therefore, we cannot rule out the effect of epigenetics in chemo-resistance.

Although the decrease in sequencing technology cost makes developing large-scale studies possible, analyzing and interpreting DNA-sequence data is still challenging. Generating data is only one part of the job, but interpreting the data and finding the causative variants still requires efficient algorithms. ClinPred is specifically designed to predict pathogenicity of variants that are causative for Mendelian diseases. As we begin to identify the variants responsible for less severe, polygenic, and complex traits, clinically relevant predictors will likely benefit from training on relevant subsets of disease databases. In particular, the use of allele frequency as a feature – even though we found it to be universally beneficial across the currently catalogued disease variants – should ideally be trained on sets of variants most relevant to different inheritance models and severity of diseases. In future developments, the predictive power of prediction models like ClinPred may be further enhanced by incorporating more components, such as specific population genotype frequency, penetrance, disease prevalence and human phenotype ontology (HPO) terms. Furthermore, progress in whole genome sequencing data will create the need to accurately predict

the effect of non-coding variants. The framework outlined here can help design future predictors for non-coding variants when appropriately large and reliable sources of pathogenic and benign variants become available.

References

Abdullah, L. N., and Chow, E. K.
2013 Mechanisms of chemoresistance in cancer stem cells. Clin Transl Med 2(1):3.

Adzhubei, I. A., et al.
2010 A method and server for predicting damaging missense mutations. Nat Methods 7(4):248-9.

Ahn, J. H., et al.
2017 Mutant p53 stimulates cell invasion through an interaction with Rad21 in human ovarian cancer cells. Sci Rep 7(1):9076.

Ainscough, B. J., et al.
2016 DoCM: a database of curated mutations in cancer. Nat Methods 13(10):806-7.

Al-Sukhni, W., et al.
2012 Screening for pancreatic cancer in a high-risk cohort: an eight-year experience. J Gastrointest Surg 16(4):771-83.

Alkodsi, A., et al.
2015 Comparative analysis of methods for identifying somatic copy number alterations from deep sequencing data. Brief Bioinform 16(2):242-54.

Althuis, M. D., et al.
2005 Global trends in breast cancer incidence and mortality 1973-1997. Int J Epidemiol 34(2):405-12.

Amberger, J., et al.
2011 A new face and new challenges for Online Mendelian Inheritance in Man (OMIM(R)). Hum Mutat 32(5):564-7.

Amin, M. B., et al.
2017 The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. CA Cancer J Clin 67(2):93-99.

Amornsupak, K., et al.
2014 Cancer-associated fibroblasts induce high mobility group box 1 and contribute to resistance to doxorubicin in breast cancer cells. BMC Cancer 14:955.

Apweiler, R., et al.
2004 UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 32(Database issue):D115-9.

Atienza, J. M., et al.
2005 Suppression of RAD21 gene expression decreases cell growth and enhances cytotoxicity of etoposide and bleomycin in human breast cancer cells. Molecular Cancer Therapeutics 4(3):361-368.

Balaji, S. A., et al.
2016 Role of the Drug Transporter ABCC3 in Breast Cancer Chemoresistance. PLoS One 11(5):e0155013.

Balko, J. M., et al.
2014 Molecular profiling of the residual disease of triple-negative breast cancers after neoadjuvant chemotherapy identifies actionable therapeutic targets. Cancer Discov 4(2):232-45.

Bartsch, D. K., et al.

2012 Familial pancreatic cancer--current knowledge. Nat Rev Gastroenterol Hepatol 9(8):445-53.
Bates, R. C., and Mercurio, A. M.
2005 The epithelial-mesenchymal transition (EMT) and colorectal cancer progression. Cancer Biol Ther 4(4):365-70.
Baylin, S. B., and Jones, P. A.
2011 A decade of exploring the cancer epigenome - biological and translational implications. Nat Rev Cancer 11(10):726-34.
Beaulieu, C. L., et al.
2014 FORGE Canada Consortium: outcomes of a 2-year national rare-disease gene-discovery project. Am J Hum Genet 94(6):809-17.
Bell, C. J., et al.
2011 Carrier testing for severe childhood recessive diseases by next-generation sequencing. Sci Transl Med 3(65):65ra4.
Bell, D.W., et al.
2002 Selective Loss of Heterozygosity in Multiple Breast Cancers from a Carrier of Mutations in Both BRCA1 and BRCA2. Cancer Research 62(10):2741-2743.
Biesecker, L. G.
2010 Exome sequencing makes medical genomics a reality. Nat Genet 42(1):13-4.
Bonanno, L., A. , et al.
2014 Platinum drugs and DNA repair mechanisms in lung cancer. Anticancer Res 34(1):493-501.
Boycott, K. M., et al.
2017 International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. Am J Hum Genet 100(5):695-705.
Brune, K. A., et al.
2010 Importance of age of onset in pancreatic cancer kindreds. J Natl Cancer Inst 102(2):119-26.
Burstein, M. D., et al.
2015 Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. Clin Cancer Res 21(7):1688-98.
Byler, S., et al.
2014 Genetic and epigenetic aspects of breast cancer progression and therapy. Anticancer Res 34(3):1071-7.
Cancer Genome Atlas, Network
2012 Comprehensive molecular portraits of human breast tumours. Nature 490(7418):61-70.
Carbon, S., et al.
2009 AmiGO: online access to ontology and annotation data. Bioinformatics 25(2):288-289.
Carey, L. A., et al.
2007 The triple negative paradox: primary tumor chemosensitivity of breast cancer subtypes. Clin Cancer Res 13(8):2329-34.
Carey, L., et al.
2010 Triple-negative breast cancer: disease entity or title of convenience? Nat Rev Clin Oncol 7(12):683-92.

Carter, H., et al.
2013 Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics 14 Suppl 3:S3.
Ceccaldi, R., et al.
2015 Homologous-recombination-deficient tumours are dependent on Poltheta-mediated repair. Nature 518(7538):258-62.
Chen, D. R., et al.
2014 Mesenchymal stem cell-induced doxorubicin resistance in triple negative breast cancer. Biomed Res Int 2014:532161.
Chen, D. S., and I. Mellman
2013 Oncology meets immunology: the cancer-immunity cycle. Immunity 39(1):1-10.
Chen, Y., et al.
2011a Mutation of NIMA-related kinase 1 (NEK1) leads to chromosome instability. Mol Cancer 10(1):5.
Chen, Y., et al.
2011b Nek1 kinase functions in DNA damage response and checkpoint control through a pathway independent of ATM and ATR. Cell Cycle 10(4):655-63.
Childs, E. J., et al.
2015 Common variation at 2p13.3, 3q29, 7p13 and 17q25.1 associated with susceptibility to pancreatic cancer. Nat Genet 47(8):911-6.
Chlebowski, R. T., et al.
2010 Estrogen plus progestin and breast cancer incidence and mortality in postmenopausal women. JAMA 304(15):1684-92.
Chmielecki, J., and Meyerson, M.
2014 DNA sequencing of cancer: what have we learned? Annu Rev Med 65:63-79.
Cho, N.
2016 Molecular subtypes and imaging phenotypes of breast cancer. Ultrasonography 35(4):281-8.
Choi, M., et al.
2009 Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci U S A 106(45):19096-101.
Choi, Y., et al.
2012 Predicting the functional effect of amino acid substitutions and indels. PLoS One 7(10):e46688.
Chun, S., and Fay, J. C.
2009 Identification of deleterious mutations within three human genomes. Genome Res 19(9):1553-61.
Ciriello, G., et al.
2013 Emerging landscape of oncogenic signatures across human cancers. Nat Genet 45(10):1127-33.
Cline, M. S., and Karchin, R.
2011 Using bioinformatics to predict the functional impact of SNVs. Bioinformatics 27(4):441-8.
Comino-Mendez, I., et al.
2011 Exome sequencing identifies MAX mutations as a cause of hereditary pheochromocytoma. Nat Genet 43(7):663-7.

Consortium, 1000 Genomes Project, et al.
2010 A map of human genome variation from population-scale sequencing. *In* Nature. Pp. 1061-1073, Vol. 467.
Cooper, G. M., et al.
2005a Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 15(7):901-13.
Cooper, G.M., et al.
2005b Distribution and intensity of constraint in mammalian genomic sequence. Genome Research 15(7):901-913.
Cortazar, P., et al.
2014 Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis. Lancet 384(9938):164-72.
Cotta-Ramusino, C., et al.
2011 A DNA damage response screen identifies RHINO, a 9-1-1 and TopBP1 interacting protein required for ATR signaling. Science 332(6035):1313-7.
Cotterell, J.
2014 Exome sequencing reveals a potential mutational trajectory and treatments for a specific pancreatic cancer patient. Onco Targets Ther 7:655-62.
Couch, F. J., et al.
2005 Germ line Fanconi anemia complementation group C mutations and pancreatic cancer. Cancer Res 65(2):383-6.
Database of Single Nucleotide Polymorphisms (dbSNP). Bethesda (MD): National Center for Biotechnology Information, National Library of Medicine, (dbSNP Build ID: 135).
Davydov, E. V., et al.
2010 Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol 6(12):e1001025.
de Bono, J., et al.
2017 Phase I, Dose-Escalation, Two-Part Trial of the PARP Inhibitor Talazoparib in Patients with Advanced Germline BRCA1/2 Mutations and Selected Sporadic Cancers. Cancer Discov 7(6):620-629.
Deb, S., et al.
2014 RAD21 cohesin overexpression is a prognostic and predictive marker exacerbating poor prognosis in KRAS mutant colorectal carcinomas. Br J Cancer 110(6):1606-13.
Denkert, C., et al.
2015 Tumor-infiltrating lymphocytes and response to neoadjuvant chemotherapy with or without carboplatin in human epidermal growth factor receptor 2-positive and triple-negative primary breast cancers. J Clin Oncol 33(9):983-91.
Denkert, C., et al.
2018 Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: a pooled analysis of 3771 patients treated with neoadjuvant therapy. Lancet Oncol 19(1):40-50.
Dent, R., et al.
2007 Triple-negative breast cancer: clinical features and patterns of recurrence. Clin Cancer Res 13(15 Pt 1):4429-34.
DePristo, M. A., et al.

2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43(5):491-8.
Dieras, V., et al.
2007 [Trastuzumab (Herceptin) and breast cancer: mechanisms of resistance]. Bull Cancer 94(3):259-66.
Dietze, E. C., et al.
2015 Triple-negative breast cancer in African-American women: disparities versus biology. Nat Rev Cancer 15(4):248-54.
Dong, C., et al.
2015 Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet 24(8):2125-37.
Dorschner, M. O., et al.
2013 Actionable, pathogenic incidental findings in 1,000 participants' exomes. Am J Hum Genet 93(4):631-40.
Drost, R., et al.
2011 BRCA1 RING function is essential for tumor suppression but dispensable for therapy resistance. Cancer Cell 20(6):797-809.
Eberle, M. A., et al.
2002 A new susceptibility locus for autosomal dominant pancreatic cancer maps to chromosome 4q32-34. Am J Hum Genet 70(4):1044-8.
Eheman, C. R., et al.
2009 The changing incidence of in situ and invasive ductal and lobular breast carcinomas: United States, 1999-2004. Cancer Epidemiol Biomarkers Prev 18(6):1763-9.
Esteller, M., et al.
2000 Inactivation of the DNA-repair gene MGMT and the clinical response of gliomas to alkylating agents. N Engl J Med 343(19):1350-4.
Exome Aggregation Consortium (ExAC), Cambridge, MA (URL: http://exac.broadinstitute.org) [January, 2015].
Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA, http://evs.gs.washington.edu/EVS/.
Ferlay, J., et al.
2015 Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 136(5):E359-86.
Fernandez-Cuesta, L., et al.
2012 Prognostic and predictive value of TP53 mutations in node-positive breast cancer patients treated with anthracycline- or anthracycline/taxane-based adjuvant therapy: results from the BIG 02-98 phase III trial. Breast Cancer Res 14(3):R70.
Findlay, G.M., et al.
2018 Accurate functional classification of thousands of BRCA1 variants with saturation genome editing. bioRxiv.
Flickinger, M., et al.
2015 Correcting for Sample Contamination in Genotype Calling of DNA Sequence Data. Am J Hum Genet 97(2):284-90.
Flygare, S., et al.

2018 The VAAST Variant Prioritizer (VVP): ultrafast, easy to use whole genome variant prioritization tool. BMC Bioinformatics 19(1):57.
Forbes, S.A., et al.
2015 COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Research 43(Database issue):D805-11.
Foulkes, W. D.
2008 Inherited susceptibility to common cancers. N Engl J Med 359(20):2143-53.
Frebourg, T.
2014 The challenge for the next generation of medical geneticists. Hum Mutat 35(8):909-11.
Fu, W., et al.
2013 Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493(7431):216-20.
Gahl, W. A., et al.
2012 The National Institutes of Health Undiagnosed Diseases Program: insights into rare diseases. Genet Med 14(1):51-9.
Garber, M., et al.
2009 Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics 25(12):i54-62.
Giannakakou, P., et al.
1997 Paclitaxel-resistant human ovarian cancer cells have mutant beta-tubulins that exhibit impaired paclitaxel-driven polymerization. J Biol Chem 272(27):17118-25.
Gilbert, M. T., et al.
2007 The isolation of nucleic acids from fixed, paraffin-embedded tissues-which methods are useful when? PLoS One 2(6):e537.
Gilbert, W., and Maxam, A.
1977 A new method for sequencing DNAT Proc Natl Acad Sci U S A 74 (2): 560–4.
Global Burden of Disease Cancer, Collaboration, et al.
2017 Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-years for 32 Cancer Groups, 1990 to 2015: A Systematic Analysis for the Global Burden of Disease Study. JAMA Oncol 3(4):524-548.
Gonzalez-Perez, A., et al.
2011 Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet 88(4):440-9.
Gottesman, M. M., et al.
2002 Multidrug resistance in cancer: role of ATP-dependent transporters. Nat Rev Cancer 2(1):48-58.
Grant, R. C., et al.
2013 Exome sequencing identifies nonsegregating nonsense ATM and PALB2 variants in familial pancreatic cancer. Hum Genomics 7:11.
Grimm, D. G., et al.
2015 The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Hum Mutat 36(5):513-23.
Grossmann, V., et al.

2011 Targeted next-generation sequencing detects point mutations, insertions, deletions and balanced chromosomal rearrangements as well as identifies novel leukemia-specific fusion genes in a single procedure. Leukemia 25(4):671-80.

Guay, D., et al.
2008 The strand separation and nuclease activities associated with YB-1 are dispensable for cisplatin resistance but overexpression of YB-1 in MCF7 and MDA-MB-231 breast tumor cells generates several chemoresistance signatures. Int J Biochem Cell Biol 40(11):2492-507.

Gulko, B., et al.
2015 A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nat Genet 47(3):276-83.

Haber, M., et al.
2006 Association of high-level MRP1 expression with poor clinical outcome in a large prospective study of primary neuroblastoma. J Clin Oncol 24(10):1546-53.

Harding, K. E., and Robertson, N. P.
2014 Applications of next-generation whole exome sequencing. J Neurol 261(6):1244-6.

Harrison, S. M., et al.
2016 Using ClinVar as a Resource to Support Variant Interpretation. Curr Protoc Hum Genet 89:8 16 1-8 16 23.

Herrero-Vicent, C., et al.
2017 Predictive and prognostic impact of tumour-infiltrating lymphocytes in triple-negative breast cancer treated with neoadjuvant chemotherapy. Ecancermedicalscience 11:759.

Hoch, N. C., et al.
2017 XRCC1 mutation is associated with PARP1 hyperactivation and cerebellar ataxia. Nature 541(7635):87-91.

Holley, R. W., et al.
1965 Structure of a Ribonucleic Acid. Science 147(3664):1462-5.

Holohan, C., et al.
2013 Cancer drug resistance: an evolving paradigm. Nat Rev Cancer 13(10):714-26.

Hon, J. D., et al.
2016 Breast cancer molecular subtypes: from TNBC to QNBC. Am J Cancer Res 6(9):1864-1872.

Hruban, R. H., et al.
2010 Update on familial pancreatic cancer. Adv Surg 44:293-311.

Hu, Q., and Baeg, G. H.
2017 Role of epigenome in tumorigenesis and drug resistance. Food Chem Toxicol 109(Pt 1):663-668.

Ioannidis, N. M., et al.
2016 REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet 99(4):877-885.

Ionita-Laza, I., et al.
2016 A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet 48(2):214-20.

Jagadeesh, K. A., et al.

2016 M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet 48(12):1581-1586.
Jemal, A., et al.
2010 Cancer statistics, 2010. CA Cancer J Clin 60(5):277-300.
Jeong, H., et al.
2012 Epithelial-mesenchymal transition in breast cancer correlates with high histological grade and triple-negative phenotype. Histopathology 60(6B):E87-95.
Jiang, Y., et al.
2009 Aberrant DNA methylation is a dominant mechanism in MDS progression to AML. Blood 113(6):1315-25.
Jones, S., et al.
2009 Exomic sequencing identifies PALB2 as a pancreatic cancer susceptibility gene. Science 324(5924):217.
Karczewski, K. J., et al.
2017 The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Res 45(D1):D840-D845.
Kashiwagi, E., et al.
2011 Enhanced expression of nuclear factor I/B in oxaliplatin-resistant human cancer cell lines. Cancer Sci 102(2):382-6.
Kim, C., et al.
2018 Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. Cell 173(4):879-893 e13.
Kircher, M., et al.
2014 A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46(3):310-5.
Kleeff, J., et al.
2016 Pancreatic cancer. Nat Rev Dis Primers 2:16022.
Klein, A. P.
2013 Identifying people at a high risk of developing pancreatic cancer. Nat Rev Cancer 13(1):66-74.
Klein, A. P., et al.
2002 Evidence for a major gene influencing risk of pancreatic cancer. Genet Epidemiol 23(2):133-49.
Klein, A. P., et al.
2009 Absence of deleterious palladin mutations in patients with familial pancreatic cancer. Cancer Epidemiol Biomarkers Prev 18(4):1328-30.
Klein, A. P., et al.
2004 Prospective risk of pancreatic cancer in familial pancreatic cancer kindreds. Cancer Res 64(7):2634-8.
Kobayashi, Y., et al.
2017 Pathogenic variant burden in the ExAC database: an empirical approach to evaluating population data for clinical variant interpretation. Genome Med 9(1):13.
Kumar, P., S. et al.
2009 Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4(7):1073-81.
Landoure, G., et al.

2012 Exome sequencing identifies a novel TRPV4 mutation in a CMT2C family. Neurology 79(2):192-4.

Landrum, M. J., et al.
2016 ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res 44(D1):D862-8.

Landrum, M. J., et al.
2018 ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res 46(D1):D1062-D1067.

Landrum, M.J., et al.
2014 ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Research 42(Database issue):D980-5.

Langer, P., et al.
2009 Five years of prospective screening of high-risk individuals from families with familial pancreatic cancer. Gut 58(10):1410-8.

Ledergerber, C., and C. Dessimoz
2011 Base-calling for next-generation sequencing platforms. Brief Bioinform 12(5):489-97.

Lehmann, B. D., et al.
2011 Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J Clin Invest 121(7):2750-67.

Li, H., and Durbin, R.
2010 Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26(5):589-95.

Li, H., et al.
2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16):2078-9.

Li, Q., et al.
2014 Gene-specific function prediction for non-synonymous mutations in monogenic diabetes genes. PLoS One 9(8):e104452.

Liedtke, C., et al.
2008 Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. J Clin Oncol 26(8):1275-81.

Litman, T., et al.
2000 The multidrug-resistant phenotype associated with overexpression of the new ABC half-transporter, MXR (ABCG2). J Cell Sci 113 (Pt 11):2011-21.

Liu, X., et al.
2011 dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat 32(8):894-9.

Liu, Y., et al.
2017 ONGene: A literature-based database for human oncogenes. J Genet Genomics 44(2):119-121.

Loibl, S., et al.
2015 Neoadjuvant treatment of breast cancer--Clinical and research perspective. Breast 24 Suppl 2:S73-7.

Lopes, M. C., et al.

2012 A combined functional annotation score for non-synonymous variants. Hum Hered 73(1):47-51.

Lu, L., et al.
2012 Loss of E-cadherin in multidrug resistant breast cancer cell line MCF-7/Adr: possible implication in the enhanced invasive ability. Eur Rev Med Pharmacol Sci 16(9):1271-9.

Lu, Q., et al.
2015 A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. Sci Rep 5:10576.

Majewski, J., et al.
2011 What can exome sequencing do for you? J Med Genet 48(9):580-9.

Manolitsas, T. P., et al.
1997 No association of a 306-bp insertion polymorphism in the progesterone receptor gene with ovarian and breast cancer. Br J Cancer 75(9):1398-9.

Mardis, E. R., and Wilson, R. K.
2009 Cancer genome sequencing: a review. Hum Mol Genet 18(R2):R163-8.

Margulies, E. H., et al.
2003 Identification and characterization of multi-species conserved sequences. Genome Res 13(12):2507-18.

Massingham, T., and Goldman, N.
2012 All Your Base: a fast and accurate probabilistic approach to base calling. Genome Biol 13(2):R13.

Mateos-Gomez, P. A., et al.
2015 Mammalian polymerase theta promotes alternative NHEJ and suppresses recombination. Nature 518(7538):254-7.

McKenna, A., et al.
2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20(9):1297-303.

Meyerson, M., et al.
2010 Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 11(10):685-96.

Milanowska, K., et al.
2011 REPAIRtoire--a database of DNA repair pathways. Nucleic Acids Research 39(Database issue):D788-92.

Miller, K. D., et al.
2016 Cancer treatment and survivorship statistics, 2016. CA Cancer J Clin 66(4):271-89.

Millot, G. A., et al.
2012 A guide for functional analysis of BRCA1 variants of uncertain significance. Hum Mutat 33(11):1526-37.

Miosge, L. A., et al.
2015 Comparison of predicted and actual consequences of missense mutations. Proc Natl Acad Sci U S A 112(37):E5189-98.

Mittendorf, E.A., et al.
2014 PD-L1 Expression in Triple-Negative Breast Cancer. Cancer Immunology Research 2(4):361-370.

Moitra, K., et al.
2012 Differential gene and microRNA expression between etoposide resistant and etoposide sensitive MCF7 breast cancer cell lines. PLoS One 7(9):e45268.
Morey, M., et al.
2013 A glimpse into past, present, and future DNA sequencing. Mol Genet Metab 110(1-2):3-24.
Mori, H., et al.
2017 The combination of PD-L1 expression and decreased tumor-infiltrating lymphocytes is associated with a poor prognosis in triple-negative breast cancer. Oncotarget 8(9):15584-15592.
Nair, M. G., et al.
2016 beta3 integrin promotes chemoresistance to epirubicin in MDA-MB-231 through repression of the pro-apoptotic protein, BAD. Exp Cell Res 346(1):137-45.
Nakamura, K., et al.
2011 Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res 39(13):e90.
Nakashima, S., et al.
2015 BRCA/Fanconi anemia pathway implicates chemoresistance to gemcitabine in biliary tract cancer. Cancer Sci 106(5):584-91.
Navin, N., et al.
2010 Inferring tumor progression from genomic heterogeneity. Genome Res 20(1):68-80.
Ng, P. C., and Henikoff, S.
2003 SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31(13):3812-4.
Ng, S. B., et al.
2010 Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 42(1):30-5.
Niroula, A., and Vihinen, M.
2016 Variation Interpretation Predictors: Principles, Types, Performance, and Choice. Hum Mutat 37(6):579-97.
Nowell, P. C., and Hungerford, D. A.
1960 Chromosome studies on normal and leukemic human leukocytes. J Natl Cancer Inst 25:85-109.
O'Brien, C., et al.
2008 Functional genomics identifies ABCC3 as a mediator of taxane resistance in HER2-amplified breast cancer. Cancer Res 68(13):5380-9.
O'Fallon, B. D., et al.
2013 VarRanker: rapid prioritization of sequence variations associated with human disease. BMC Bioinformatics 14 Suppl 13:S1.
O'Rawe, J., et al.
2013 Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med 5(3):28.
Ostergaard, P., et al.
2011 Mutations in GATA2 cause primary lymphedema associated with a predisposition to acute myeloid leukemia (Emberger syndrome). Nat Genet 43(10):929-31.

Paik, S., et al.
2004 A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 351(27):2817-26.
Park, D. J., et al.
2012 Rare mutations in XRCC2 increase the risk of breast cancer. Am J Hum Genet 90(4):734-9.
Parsons, D. W., et al.
2008 An integrated genomic analysis of human glioblastoma multiforme. Science 321(5897):1807-12.
Pinxten, W., and Howard, H. C.
2014 Ethical issues raised by whole genome sequencing. Best Pract Res Clin Gastroenterol 28(2):269-79.
Pogue-Geile, K. L., et al.
2006 Palladin mutation causes familial pancreatic cancer and suggests a new cancer mechanism. PLoS Med 3(12):e516.
Pollard, K. S., et al.
2010 Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 20(1):110-21.
Powles, T., et al.
2014 MPDL3280A (anti-PD-L1) treatment leads to clinical activity in metastatic bladder cancer. Nature 515(7528):558-62.
Quang, D., et al.
2015 DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics 31(5):761-3.
Rabbani, B., M. Tekin, and N. Mahdieh
2014 The promise of whole-exome sequencing in medical genetics. J Hum Genet 59(1):5-15.
Rahib, L., et al.
2014 Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. Cancer Res 74(11):2913-21.
Reva, B., Y. Antipin, and C. Sander
2011 Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res 39(17):e118.
Rhodes, J.M., et al.
2011 Gene Regulation by Cohesin in Cancer: Is the Ring an Unexpected Party to Proliferation? Molecular Cancer Research 9(12):1587-1607.
Richards, S., et al.
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. Genet Med 17(5):405-24.
Roberts, N. J., et al.
2012 ATM mutations in patients with hereditary pancreatic cancer. Cancer Discov 2(1):41-6.
Roberts, N. J., et al.
2016 Whole Genome Sequencing Defines the Genetic Heterogeneity of Familial Pancreatic Cancer. Cancer Discov 6(2):166-75.

Robinson, J. T., et al.
2011 Integrative genomics viewer. Nat Biotechnol 29(1):24-6.
Romond, E. H., et al.
2005 Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. N Engl J Med 353(16):1673-84.
Rubinstein, W. S., and Weissman, S. M.
2008 Managing hereditary gastrointestinal cancer syndromes: the partnership between genetic counselors and gastroenterologists. Nat Clin Pract Gastroenterol Hepatol 5(10):569-82.
Rustgi, A. K.
2014 Familial pancreatic cancer: genetic advances. Genes Dev 28(1):1-7.
Saarinen, S., et al.
2011 Exome sequencing reveals germline NPAT mutation as a candidate risk factor for Hodgkin lymphoma. Blood 118(3):493-8.
Salaria, S. N., et al.
2007 Palladin is overexpressed in the non-neoplastic stroma of infiltrating ductal adenocarcinomas of the pancreas, but is only rarely overexpressed in neoplastic cells. Cancer Biol Ther 6(3):324-8.
Samanta, D., et al.
2018 Chemotherapy induces enrichment of CD47(+)/CD73(+)/PDL1(+) immune evasive triple-negative breast cancer cells. Proc Natl Acad Sci U S A 115(6):E1239-E1248.
Sanger, F
1958 Nobel lecture: the chemistry of insulin. .
Sanger, F., et al.
1977 DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74(12):5463-7.
Schaafsma, G. C., and Vihinen, M.
2015 VariSNP, a benchmark database for variations from dbSNP. Hum Mutat 36(2):161-6.
Schalper, K. A., et al.
2014 In Situ Tumor PD-L1 mRNA Expression Is Associated with Increased TILs and Better Outcome in Breast Carcinomas. Clinical Cancer Research 20(10):2773-2782.
Schmidt, D., et al.
2010 A CTCF-independent role for cohesin in tissue-specific transcription. Genome Res 20(5):578-88.
Schneider, R., et al.
2011 German national case collection for familial pancreatic cancer (FaPaCa): ten years experience. Fam Cancer 10(2):323-30.
Schwartzentruber, J., et al.
2012 Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma. Nature 482(7384):226-31.
Schwarz, J. M., et al.
2010 MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods 7(8):575-6.
Segui, N., et al.

2015 Germline Mutations in FAN1 Cause Hereditary Colorectal Cancer by Impairing DNA Repair. Gastroenterology 149(3):563-6.
Shah, S., et al.
2013 A recurrent germline PAX5 mutation confers susceptibility to pre-B cell acute lymphoblastic leukemia. Nat Genet 45(10):1226-1231.
Shen, H., et al.
2007 Comparative metabolic capabilities and inhibitory profiles of CYP2D6.1, CYP2D6.10, and CYP2D6.17. Drug Metab Dispos 35(8):1292-300.
Sherry, S. T., et al.
2001 dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29(1):308-11.
Shi, C., et al.
2009 Familial pancreatic cancer. Arch Pathol Lab Med 133(3):365-74.
Shihab, H. A., et al.
2014 Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. Hum Genomics 8:11.
Shihab, H. A., et al.
2015 An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics 31(10):1536-43.
Shou, J., et al.
2004 Mechanisms of tamoxifen resistance: increased estrogen receptor-HER2/neu cross-talk in ER/HER2-positive breast cancer. J Natl Cancer Inst 96(12):926-35.
Siepel, A., et al.
2005 Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15(8):1034-50.
Slamon, D. J., et al.
1987 Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science 235(4785):177-82.
Slavin, T. P., et al.
2018 The spectrum of genetic variants in hereditary pancreatic cancer includes Fanconi anemia genes. Fam Cancer 17(2):235-245.
Soden, S. E., et al.
2014 Effectiveness of exome and genome sequencing guided by acuity of illness for diagnosis of neurodevelopmental disorders. Sci Transl Med 6(265):265ra168.
Sokolenko, A. P., et al.
2014 Double heterozygotes among breast cancer patients analyzed for BRCA1, CHEK2, ATM, NBN/NBS1, and BLM germ-line mutations. Breast Cancer Research and Treatment 145(2):553-562.
Sotiriou, C., and Pusztai, L.
2009 Gene-expression signatures in breast cancer. N Engl J Med 360(8):790-800.
Srivastava, S., et al.
2014 Clinical whole exome sequencing in child neurology practice. Ann Neurol 76(4):473-83.
Stadler, Z. K., et al.
2014 Cancer genomics and inherited risk. J Clin Oncol 32(7):687-98.
Stavrovskaya, A. A.

2000 Cellular mechanisms of multidrug resistance of tumor cells. Biochemistry (Mosc) 65(1):95-106.

Stenson, P. D., et al.
2017 The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet 136(6):665-677.

Stenson, P. D., et al.
2014 The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet 133(1):1-9.

Stevens, K. N., et al.
2013 Genetic susceptibility to triple-negative breast cancer. Cancer Res 73(7):2025-30.

Symmans, W. F., et al.
2007 Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy. J Clin Oncol 25(28):4414-22.

Tan, Q., et al.
2015 Src/STAT3-dependent heme oxygenase-1 induction mediates chemoresistance of breast cancer cells to doxorubicin by promoting autophagy. Cancer Sci 106(8):1023-32.

Tan, R., et al.
2014 An evaluation of copy number variation detection tools from whole-exome sequencing data. Hum Mutat 35(7):899-907.

Testa, J. R., et al.
2011 Germline BAP1 mutations predispose to malignant mesothelioma. Nat Genet 43(10):1022-5.

Torre, L. A., et al.
2016 Global Cancer Incidence and Mortality Rates and Trends--An Update. Cancer Epidemiol Biomarkers Prev 25(1):16-27.

Trivers, K. F., et al.
2009 The epidemiology of triple-negative breast cancer, including race. Cancer Causes Control 20(7):1071-82.

van 't Veer, L. J., et al.
2002 Gene expression profiling predicts clinical outcome of breast cancer. Nature 415(6871):530-6.

van der Heijden, M. S., et al.
2004 Functional defects in the fanconi anemia pathway in pancreatic cancer cells. Am J Pathol 165(2):651-7.

van der Heijden, M. S., et al.
2003 Fanconi anemia gene mutations in young-onset pancreatic cancer. Cancer Res 63(10):2585-8.

van der Velde, K. J., et al.
2017 GAVIN: Gene-Aware Variant INterpretation for medical sequencing. Genome Biol 18(1):6.

van Doorn, R., et al.
2005 Epigenetic profiling of cutaneous T-cell lymphoma: promoter hypermethylation of multiple tumor suppressor genes including BCL7a, PTPRG, and p73. J Clin Oncol 23(17):3886-96.

Varela, I., et al.
2011 Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. Nature 469(7331):539-42.
Vihinen, M.
2013 Guidelines for reporting and using prediction tools for genetic variation analysis. Hum Mutat 34(2):275-82.
Wang, K., et al.
2011 Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer. Nat Genet 43(12):1219-23.
Wang, K., et al
2010 ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research 38(16):e164-e164.
Wang, L., and Wheeler, D. A.
2014 Genomic sequencing for cancer diagnosis and therapy. Annu Rev Med 65:33-48.
Wang, W., et al.
2007 PancPRO: risk assessment for individuals with a family history of pancreatic cancer. J Clin Oncol 25(11):1417-22.
Williams, E. S., and Hegde, M.
2013 Implementing genomic medicine in pathology. Adv Anat Pathol 20(4):238-44.
Wimberly, H., et al.
2015 PD-L1 Expression Correlates with Tumor-Infiltrating Lymphocytes and Response to Neoadjuvant Chemotherapy in Breast Cancer. Cancer Immunol Res 3(4):326-32.
Witkiewicz, A. K., et al.
2015 Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. Nat Commun 6:6744.
Witkowski, L., et al.
2014 Germline and somatic SMARCA4 mutations characterize small cell carcinoma of the ovary, hypercalcemic type. Nat Genet 46(5):438-43.
Witz, I. P.
2008 The selectin-selectin ligand axis in tumor progression. Cancer Metastasis Rev 27(1):19-30.
Wolfgang, C. L., et al.
2013 Recent progress in pancreatic cancer. CA Cancer J Clin 63(5):318-48.
Worm, J., et al.
2001 Methylation-dependent silencing of the reduced folate carrier gene in inherently methotrexate-resistant human breast cancer cells. J Biol Chem 276(43):39990-40000.
Worthey, E. A., et al.
2011 Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genet Med 13(3):255-62.
Xu, H., et al.
2011 Enhanced RAD21 cohesin expression confers poor prognosis and resistance to chemotherapy in high grade luminal, basal and HER2 breast cancers. Breast Cancer Res 13(1):R9.
Yabar, C. S., and Winter, J. M.
2016 Pancreatic Cancer: A Review. Gastroenterol Clin North Am 45(3):429-45.

Yachida, S., et al.
2010 Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature 467(7319):1114-7.
Yadav, S., et al.
2013 Role of SMC1 in overcoming drug resistance in triple negative breast cancer. PLoS One 8(5):e64338.
Yanase, K., et al.
2004 Gefitinib reverses breast cancer resistance protein-mediated drug resistance. Mol Cancer Ther 3(9):1119-25.
Yang, H., and Wang, K.
2015 Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nat Protoc 10(10):1556-66.
Yang, Y., et al.
2013 Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N Engl J Med 369(16):1502-11.
Yang, Y., et al.
2014 Molecular findings among patients referred for clinical whole-exome sequencing. JAMA 312(18):1870-9.
Yao, H., et al.
2017 Triple-negative breast cancer: is there a treatment on the horizon? Oncotarget 8(1):1913-1924.
Yao, Y. S., et al.
2015 miR-141 confers docetaxel chemoresistance of breast cancer cells via regulation of EIF4E expression. Oncol Rep 33(5):2504-12.
Zang, Z. J., et al.
2012 Exome sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes. Nat Genet 44(5):570-4.
Zhang, J., et al.
2013 Exome profiling of primary, metastatic and recurrent ovarian carcinomas in a BRCA1-positive patient. BMC Cancer 13:146.
Zhang, P., et al.
2016 Upregulation of programmed cell death ligand 1 promotes resistance response in non-small-cell lung cancer patients treated with neo-adjuvant chemotherapy. Cancer Sci 107(11):1563-1571.
Zhao, M., et al.
2016 TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. Nucleic Acids Res 44(D1):D1023-31.