Unraveling the genetics of cancer using whole-exome sequencing

by

Jian Zhang (Carrot-Zhang)

M.Sc. Western University 2010

B.Sc. Sichuan University 2008

Department of Human Genetics

McGill University

Montreal, Quebec, Canada

February 2016

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

Doctor of Philosophy

DEDICATION

To a little boy who has a lot of trains

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	Allelic imbalance
ATRT	Atypical teratoid/rhabdoid tumor
BAF	B-allele frequency
BWA	Burrows-Wheeler Aligner
CCDS	Consensus coding sequence
ChIP-seq	Chromatin immunoprecipitation-sequencing
CNV	Copy number variants
COSMIC	Catalogue of Somatic Mutations in Cancer
dbSNP	Single Nucleotide Polymorphism Database
ER	Estrogen receptor
EVS	Exome variant server
FFPE	Formalin-fixed and paraffin-embedded
GATK	Genome analysis toolkit package
Gb	Gigabase
GBM	Glioblastoma
GWAS	Genome-wide association studies
HER2	Human epidermal growth factor receptor 2
HGSC	High-grade serous ovarian cancer
IGV	Integrative Genomics Viewer
IHC	Immunohistochemistry
Indels	Small insertions or deletions

LOH	Loss of heterozygosity
NGS	Next-generation sequencing
PARP	Poly ADP-ribose polymerase
PR	Progesterone receptor
ROC	Receiver operating characteristic
RPKM	Reads Per Kilobase of exon model per Million mapped reads
SCCOHT	Small cell carcinoma of the ovary, hypercalcemic type
SIFT	Sorting Intolerant from Tolerant
SNP	Single-nucleotide polymorphisms
SNV	Single-nucleotide variants
TCGA	The cancer genome atlas
WES	Whole-exome sequencing
WGS	Whole-genome sequencing

Abstract

Hereditary cancer studies have shed light on the understanding of cancer genetics. Inherited components are particularly important in breast cancer and ovarian cancer. Whole-exome sequencing (WES) that targets the protein-coding region of the genome has emerged as a powerful method to reveal genetic alterations in cancer. This thesis work focuses on uncovering the genetic basis underlying familial breast cancer and ovarian cancer. Using WES as the primary investigative tool, deleterious germ-line mutations in *SMARCA4* were identified in all the small cell carcinoma of the ovary, hypercalcemic type (SCCOHT) families that we studied. Genomic analysis revealed that SCCOHT is universally characterized by the complete loss of SMARCA4. By performing WES analysis on French-Canadian, *BRCA1* and *BRCA2*-negative breast cancer families followed by validation of the candidate gene in additional cases, *RECQL* was discovered as a new breast cancer susceptibility gene. These findings indicate the central roles of the SWI/SNF chromatin-remodeling complex and the DNA repair pathway in driving the development of SCCOHT and breast cancer, respectively, and provided novel targets for the therapeutic development of these diseases. Furthermore, motivated by accurately identifying somatic mutations with low allelic-fraction from WES data, I developed a variant caller using tumor sample and matched normal sample, plus a user-defined control panel of non-cancer samples. In comparison to other existing tools, I showed superior

performance of this algorithm, especially for calling variants from low-quality cancer samples such as formalin-fixed and paraffin-embedded (FFPE) samples.

Abrégé

Les études sur les cancer héréditaires ont aidé à la compréhension de la génétique du cancer. Les composantes héréditaires sont particulièrement importantes dans les cancers du sein et de l'ovaire. La technique du séquençage de l'exome (WES), qui cible la région codant pour la protéine dans le génome, a émergé comme une méthode puissante pour révéler des altérations génétiques liées au cancer. Le travail de doctorat présenté dans cette thèse porte sur la découverte de la base génétique sous-jacente du cancer du sein et de l'ovaire familial. En utilisant la technique WES comme outil d'enquête primaire, nous avons identifié les mutations délétères dans SMARCA4 présent dans tous les cas de carcinomes à petites cellules familiales de l'ovaire, de type hypercalcémie (SCCOHT) que nous avons étudié. L'analyse génomique a révélé que SCCOHT est universellement caractérisé par la perte totale de SMARCA4. Nous avons également effectué une analyse WES sur des familles québécoises avec cancer du sein sans mutations dans les gènes BRCA1 et BRCA2. Le criblage du gène candidate par séquençage de type Sanger dans des cas québécois supplémentaire a permis d'identifier RECQL comme un nouveau gène de susceptibilité au cancer du sein. Nos résultats indiquent l'importance des complexes de remodelage chromatique SWI/SNF dans le développement du SCCOHT et de la voie de réparation de l'ADN dans le développement du cancer du sein. L'identification de ces nouveaux gènes fournit des nouvelles cibles thérapeutiques de ces maladies. Dû l'importance des mutations somatiques à

faible fraction allélique dans les cancers, nous avons développé un outil appelé « variant caller ». Cet outil utilise les donnés WES d’un échantillon tumoral, un échantillon apparié sain ainsi qu’un panneau de commande définie par l'utilisateur d'échantillons non-cancéreux. En comparaison avec d'autres outils existants, nous avons démontré une performance supérieure de notre algorithme, en particulier pour les échantillons de cancer de faible qualité tels que des échantillons formaline fixé et paraffine intégrée « as formalin-fixed and paraffin-embedded (FFPE) ».

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor Dr. Jacek Majewski. His insightful vision for research has been inspiring me. The discussion environment he provided has always been calm and relaxing. His mentorship has shown me what it takes to succeed in science. I would also like to thank my co-supervisor Dr. William Foulkes, who has provided me with support and guidance throughout this thesis work. I thank him for being an excellent advisor, and for trusting me handling his precious samples.

I thank the former and current members of the Majewski lab, my friends, Claudia Kleinman, Zibo Wang, Jeremy Schwartzenbruber, Amandine Bemmo, Martine Tetreault, Simon Papillon, Emilie Lalonde, Yuhao Shi, Somayyeh Fahiminiya, Najmeh Alirezaie, Hamid Nikbakht, Javad Nadaf, Eric Bareke, Matt Osmond and Rui Li, who contributed to so many aspects of my PhD experience. I am also thankful for the collaboration and help given by the members of the Foulkes lab: Leora Witkowski, Barbara Rivera, Mona Wu, Ester Castellsague, Nelly Sabbaghian, Rachel Silva-Smith, and Francois Plourde. I give many thanks to Nancy Hamel for all the career advice. I thank Hamid, Rui and Simon for the helpful discussions about LoLoPicker. I thank Nancy, Martine and Simon for proofreading this thesis.

I am thankful for the advice from Dr. Mark Basik and Dr. Patricia Tonin. Particularly, I am also indebted to Dr. Aimee Ryan and Ross MacKay for their support and dedication. The Department of Human Genetics has simply been a wonderful place. I give thanks to my colleagues in The McGill University and Génome Québec Innovation Centre: Pierre Lepage and Alexandre Montpetit for the sequencing service; Hoai-Thu Vo for the administrative support; Genevieve Dancausse for the technical support; Dr. Simon Gravel, Dr. Rob Sladek, Bing Ge, Xiaojian Shao and Louis Letourneau for all the discussions.

I would like to thank all the families for their unfailing support of this thesis work. Many thanks go to our collaborators and their group members: Dr. Yasser Riazalhosseini, Dr. Mohammad R. Akbari, Dr. Nada Jabado, Dr. Steffen Albrecht and Dr. Martin Hasselblatt.

Last but not the least, I owe special appreciation to my dearest family, to my parents who always believe in me and give me the freedom to follow my dream; to my husband who sacrifices so much to support my career; to my son who has always been my source of strength. I love them from the bottom of my heart.

Originality and Significance

This thesis work addresses the issues in efficiently studying cancer genetics on the genomic level. In Chapter 2, the genetic basis of SCCOHT is studied for the first time. The work in Chapter 3 is among the first efforts to study population-specific breast cancer alleles using WES. A new breast cancer gene is successfully identified. The work in Chapter 4 innovatively studies the progression of ovarian cancer by performing WES on multiple tumor samples from a single patient. The work in Chapter 5 developed a novel algorithm to improve the accuracy of detecting somatic mutations. This thesis work significantly contributes to the understanding of genetic alteration in SCCOHT and breast cancer, which provides new opportunities for the development of novel targeted therapies. SCCOHT and breast cancer, which provides new opportunities for genetic testing and the development of novel targeted therapies.

Contribution of The Authors

Chapter 2. Genomic Characterization Revealed *SMARCA4* Alterations as the Major Genetic Cause of Malignant Small Cell Carcinoma of the Ovary, Hypercalcemic Type: Jian Carrot-Zhang performed the bioinformatics analysis and contributed to the novel findings. Leora Witkowski collected samples and performed the Sanger sequencing analysis. Somayyeh Fahiminiya aided the bioinformatics analysis. Javad Nadaf performed the copy number change analysis. Steffen Albrecht and Jocelyne Arseneau helped with pathological examination. Nancy Hamel oversaw the experiments. Eva Tomiak, David Grynspan, Emmanouil Saloustros, Caltherine Gilpin, Rouzan G. Karabakhtsian, Elizabeth A. Eeilly, Frederick R. Ueland, Anna Margiolaki, Kitty Pavlakis, Sharon M. Castellino, Janez Lamovec, Lawrence M. Roth, Thomas M. Ulbright, Tracey A. Bender, Vassilis Georgoulias, Michel Longy, Ryan S. Lee, Charles W. M. Roberts, Nada Jabado, and Martin Hasselblatt provided samples and critical input. Barbara Rivera, Ester Castellsague and Mona Wu aided in experiments. Rachel Silva-Smith, Francois Plourde, Avi Saskin and Helen J. Mackay aided in sample collection. Madeleine Arseneault aided in sample preparation. Inga Nagel and Reiner Siebert performed FISH analysis. W Glenn McCluggage and Blaise A. Clarke provide samples and were the reference pathologists. Yasser Riazalhosseini oversaw the preparation of formalin-fixed, paraffin-embedded samples for whole-exome sequencing. Jacek Majewski oversaw all

bioinformatics analysis. William D. Foulkes designed the project and oversaw all aspects of the project.

Chapter 3. WES study of familial breast cancer cases based on the French-Canadian founder population identifies *RECQL* as a new breast cancer susceptibility gene: Jian Carrot-Zhang performed the data analysis of WES for the French-Canadian study and contributed to the novel findings. Barbara Rivera and Nancy Hamel performed sample genotyping for the French-Canadian study. Javad Nadaf performed the copy number change analysis. Sylvie Giroux and Shiyu Zhang contributed to sample genotyping. Patricia N. Tonin and Francois Rousseau contributed to data collection and provided study subjects for the French-Canadian study. Jacek Majewski oversaw the bioinformatics analysis for the French-Canadian study. William D. Foulkes contributed to study design and data interpretation, provided study subjects for the French-Canadian study. Cezary Cybulski, Wojciech Kluzniak, Aniruddh Kashyap, Dominika Wokolorczyk, Tomasz Huzarski, Jacek Gronwald, Tomasz Byrski, Marek Szwiec, Anna Jakubowska, Helena Rudnicka, Marcin Lener, Bartlomiej Masojc, Bohdan Gorski, Tadeusz Debniak and Jan Lubinski contributed to this study from the Polish side. Steven A. Narod contributed to study design and data interpretation for both the Polish and French-Canadian studies, provided study subjects for the French-Canadian study. Mohammad R. Akbari contributed to the study design, the whole-exome and Sanger sequencing analysis, and interpretation of the data for both the Polish and French-Canadian studies and also drafted the manuscript.

Chapter 4. WES analysis of primary, metastatic and recurrent ovarian carcinomas in a *BRCA1*-positive patient: Jian Carrot-Zhang performed the WES and Sanger sequencing analysis. Yuhao Shi performed the copy number change analysis. Emilie Lalonde aided the bioinformatics analysis. Lili Li and Luca Cavallone collected samples and performed the experiments. Alex Ferenczy and Walter H. Gotlieb provided samples and performed the clinical analysis. Pierre Lepage performed the Sanger sequencing. Bing Ge aided the PeakPicker analysis. William D. Foulkes and Jacek Majewski designed the project and oversaw all aspects of the project.

Chapter 5. LoLoPicker — Detecting Low Allelic-Fraction Variants in Low-Quality Cancer Samples from Whole-exome Sequencing Data: Jian Carrot-Zhang designed the project, developed and implemented the algorithm. Jacek Majewski designed the project and oversaw the algorithm development.

Thesis Format

This thesis is in the traditional style and consists of a general introduction and literature review, four chapters of main contributions, and a general discussion. Chapter 2, 3, and 4 contain published results. Permissions are obtained for reproducing published materials. A version of Chapter 5 is under preparation for publication. The format of the thesis respects the McGill University Guidelines for Thesis Preparation.

Chapter 1. General Introduction and Literature Reviews

Cancer is a genetic disease that is characterized by unrestrained proliferations of cells as a result of DNA sequence alterations (Stratton, Campbell & Futreal 2009). Although copy number variants (CNVs) and other large-scale chromosomal rearrangements could be observed in tumor cells for over a hundred years, only in 1982 was the first point mutation identified in *HRAS* in bladder cancer (Rowley 1973, Reddy et al. 1982, Tabin et al. 1982). This discovery launched a new era of searching for genes that are associated with various human cancers. In fact, many of the known cancer genes were uncovered by investigating inherited forms of cancer, thus emphasizing the importance of studying hereditary cancers (Foulkes 2008). Now with the completion of the Human Genome Project and the advance of massively parallel and high-throughput sequencing technologies, the identification of cancer genes has been significantly accelerated.

This thesis first focuses on discovering cancer genes through the study of inherited cancers, using whole-exome sequencing (WES) – a high-throughput sequencing technology that targets all the known coding exons in the genome (known as exome) – as the primary investigative tool. The utility of WES in comprehensively studying somatic alterations will then be discussed. This

introduction will first describe inherited susceptibility to cancer, in particular, an overview of hereditary ovarian and breast cancers and their previously identified susceptibility genes. The introduction will then move on to a review of the WES technology, its data analysis aspect and issues in applying WES to cancer research. Finally, I will discuss our current understanding of cancer evolution.

1.1 Inherited Susceptibility in Cancer

1.1.1 Overview of Hereditary Cancer

Hereditary cancer is caused by a genetic alteration affecting cell function that is passed from generation to generation in a family. Individuals with family history of, for example, breast cancer, pancreatic cancer, colorectal cancer and ovarian cancer, have an approximately two-fold increased cancer risk, and the risk is further increased if multiple first-degree relatives are affected, or if the patients are diagnosed at a young age (Pharoah et al. 1997, Ford et al. 1998, Stratton et al. 1998, Johns, Houlston 2001, Permuth-Wey, Egan 2009). Hereditary cancer makes up about 5% of all cancers. Family history is present in higher proportion in prostate, breast and colorectal cancer, whereas in some cancer types, for example bone and salivary gland cancer, family history is rarely present (Hemminki, Sundquist & Bermejo 2008).

1.1.2 Genetic basis of Hereditary Cancer

The genetic bases of hereditary cancers have been explained by the “two-hit” model (Nordling 1953, Armitage, Doll 1957, Knudson 1971). In general, the initiation and development of tumor cells require a multitude of genetic changes, including the biallelic inactivation resulting in loss-of-function of a tumor-

suppressor gene. In hereditary cancers, one germ-line inactivation of the tumor suppressor gene is inherited and a second, somatic mutation inactivates the wild-type allele in the tumor cell. In sporadic cases, both inactivations are somatic (Knudson 1971). Therefore, inherited cancers are more likely to be early-onset than sporadic cancers. The somatic hit is not necessarily a point mutation. Large-scale deletion, translocation or mitotic recombination resulting in loss of heterozygosity (LOH) at the position of the tumor-suppressor genes, and subsequently the absence of the wild-type protein, are frequently observed in the tumors (Foulkes 2008, Cavenee et al. 1983).

The “two-hits” model was later revisited by Kinzler and Vogelstein. They described the initiation of tumor as a multi-step process, after discovering DNA repair genes in hereditary colorectal cancer (Kinzler, Vogelstein 1996, Kinzler, Vogelstein 1997). The original model offers strong support to explain, for example, the role of *RB1* susceptibility to retinoblastoma and that of *APC* in familial adenomatous polyposis, by which the germ-line mutation hits a cell growth gene that directly control cell proliferation, cell-cycle checkpoints and cell death. The rate-limiting step triggering the tumorigenesis is the second hit targeting the same gene. By contrast, some DNA repair genes, such as mismatch-repair genes *MLH1* and *MSH2* in Lynch syndrome (hereditary nonpolyposis colorectal cancer), and homologous recombination-repair genes *BRCA1* and *BRCA2* in hereditary breast and ovarian cancers do not promote tumor initiation directly, but rather are important to maintain genome integrity.

Biallelic inactivation of these genes allows accumulation of new mutations and results in genetic instability; further molecular events (for example, the targeting of growth-regulating genes) are required to trigger the malignant transformation (Kinzler, Vogelstein 1996, Kinzler, Vogelstein 1997).

1.1.3 Cancer Susceptibility Genes

Numerous cancer susceptibility genes have been identified. The observation of recurrent deletion on chromosome 13 in retinoblastoma tumors led to the discovery of the first tumor-suppressor gene, *RB1* (Knudson et al. 1976, Friend et al. 1986). Linkage analysis, which links the disease phenotype to polymorphic DNA markers so the causal allele can be localized, has identified a few rare but high-penetrance genes in common cancers (Foulkes 2008). However, the limitation of linkage studies is that large families with several affected generations are required to be able to identify new cancer susceptibility genes. Moreover, moderate to low-penetrance alleles may not be penetrant enough to cause a noticeable familial history of cancer, and therefore cannot be detected. Genome-wide association studies (GWAS) revealed many common, low-risk alleles by testing the association of thousands of single-nucleotide polymorphisms (SNPs) with cancer cases. Those alleles are combined to confer a range of susceptibility in the population (Houlston, Peto 2004).

Furthermore, the candidate-gene approach, which tests the association of a specific variant with the disease, has identified several moderate-penetrance alleles in common cancers. Genes are selected based on the prior knowledge of their functions and their etiological roles in the disease (Tabor, Risch & Myers 2002). Candidate-gene experiments can be conducted for a large cohort of case and control samples at the population level. However, most of the candidate-gene studies allow for a small number of genes being examined at one time, and the selection of genes is limited by the given hypothesis. Thus, a hypothesis-free but cost-effective sequencing alternative is needed. The main focus of this thesis is the identification of ovarian cancer and breast cancer genes using high-throughput sequencing technologies. What is known about the genetic basis of ovarian cancer and breast cancer is discussed in the next sections.

1.1.4 Inherited Susceptibility in Ovarian Cancer

1.1.4.1 Classification of ovarian cancer

Ovarian cancer is the fourth leading cause of female cancer death in the developed world, accounting for nearly 4% of all female cancers (Bray et al. 2013). Based on the histogenesis of origin, ovarian cancer can be broadly classified as epithelial and non-epithelial (Scully 1987, Kaku et al. 2003). Epithelial ovarian tumors are further clustered into histological groups as serous,

mucinous, endometrioid, clear cell, transitional, squamous cell, mixed-epithelial and undifferentiated tumors (Kaku et al. 2003). Non-epithelial ovarian tumors are further clustered into germ-cell, sex cord-stromal and miscellaneous tumors (Kaku et al. 2003, Witkowski et al. 2013). Moreover, based on the morphological and molecular studies, epithelial ovarian tumors are divided into two types (Figure 1.1). Low-grade types of tumors associated with low proliferative activity and slow progression have been characterized by mutations in *BRAF, KRAS*, and *PTEN*. High-grade types of tumors associated with aggressive activity and rapid progression have been characterized by mutations in *TP53*, *BRCA1*, *BRCA2* and *NF1* (Ho et al. 2004). In spite of this, the notion of ovarian tumor classification continues to change. Now it is believed that high-grade serous ovarian cancer (HGSC) are derived from the fallopian tube, whereas low-grade serous ovarian cancer is considered to be of ovarian origin (Vaughan et al. 2011). More and more subtypes of ovarian cancer are likely to be identified by comprehensive genomic analysis (The Cancer Genome Atlas Research Network 2011, Jayson et al. 2014).

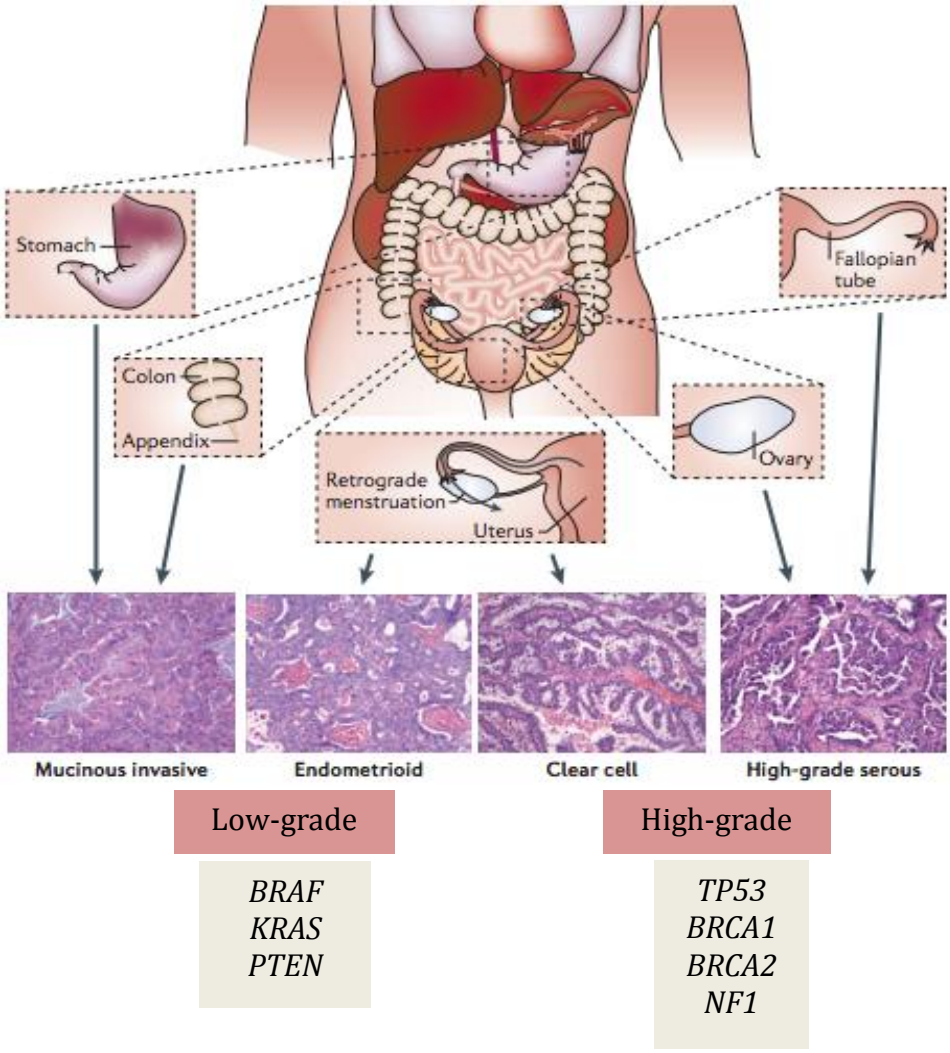


Figure 1.1: **Classification of ovarian cancer based on the origins and molecular profiles.**

Ovarian cancers are derived from different tissues. High-grade serous ovarian cancers are probably derived from the surface of the ovary and/or the fallopian tube, which is universally characterized by *TP53* mutations, and in some cases, associated with *BRCA1*, *BRCA2* and *NF1* mutations. Adapted with permission from Vaughan et al., *Nature Reviews Cancer*, 2011.

1.1.4.2 Genomic profile of epithelial ovarian cancer

Invasive epithelial ovarian cancer displays several distinct genomic and transcriptional features across different subtypes, suggesting that ovarian tumors should be considered differently at the molecular level (Vaughan et al. 2011, Köbel et al. 2008). In HGSC, the mutational spectrum is relatively simple – only a few recurrent mutations have been observed in HGSC – but the genomic landscape is remarkably complicated. Genomic analysis showed that *TP53* is prevalently mutated in 96% of all HGSC cases, followed by *BRCA1* and *BRCA2* mutations in 22% of all cases. Other genes, such as *NF1*, *RB1* and *CDK12*, have only been found mutated in only 2-6% of all cases. Meanwhile, at least 50% of the HGSC tumors showed recurrent CNVs, including amplifications on chromosome 3q, 8q and 20q, and deletions on chromosome 17, 18 and 19 (The Cancer Genome Atlas Research Network 2011).

On the other hand, *TP53* and the *BRCA* genes are rarely mutated in non-HGSC ovarian tumors. Many of the driving mutations in endometrioid, clear cell and mucinous ovarian tumors have been found in genes in the RAS signaling pathway, including *PIK3CA*, *KRAS*, and *PTEN* (Obata et al. 1998, Jones et al. 2010). Interestingly, the number of CNVs in the clear cell cancer is similar to that seen in low-grade serous ovarian cancer but much lower than in HGSC (Kuo et al. 2010). Moreover, alterations in *ARID1A* have been related to aberrant regulation of chromatin remodeling, suggesting that epigenetic change may play

an important role in the tumorigenesis of clear cell ovarian cancer (Jones et al. 2010).

1.1.4.3 Susceptibility genes in hereditary ovarian cancer

Approximately 7% of ovarian cancer patients have at least one affected, first-degree female relative (Whittemore 1994). Germ-line mutations in *BRCA1* and *BRCA2* are the most frequently observed inherited risk, accounting for the majority of hereditary ovarian cancers (The Cancer Genome Atlas Research Network 2011, Foulkes et al. 2007). In fact, *BRCA1* was discovered from families with the incidence of both breast and ovarian cancer (Miki et al. 1994). Other genes in the homologous recombination pathway, such as *RAD51C*, *RAD51D*, and *BRIP1* confer susceptibility to ovarian cancer as well (Meindl et al. 2010, Walsh et al. 2011, Loveday et al. 2011) (Figure 1.2). Mismatch repair genes *MSH2*, *MSH6* and *MLH1*, together with *TP53* explain another small proportion of hereditary ovarian cancer. Most of these genes were found in non-HGSC subtypes (Walsh et al. 2011, Song et al. 2014).

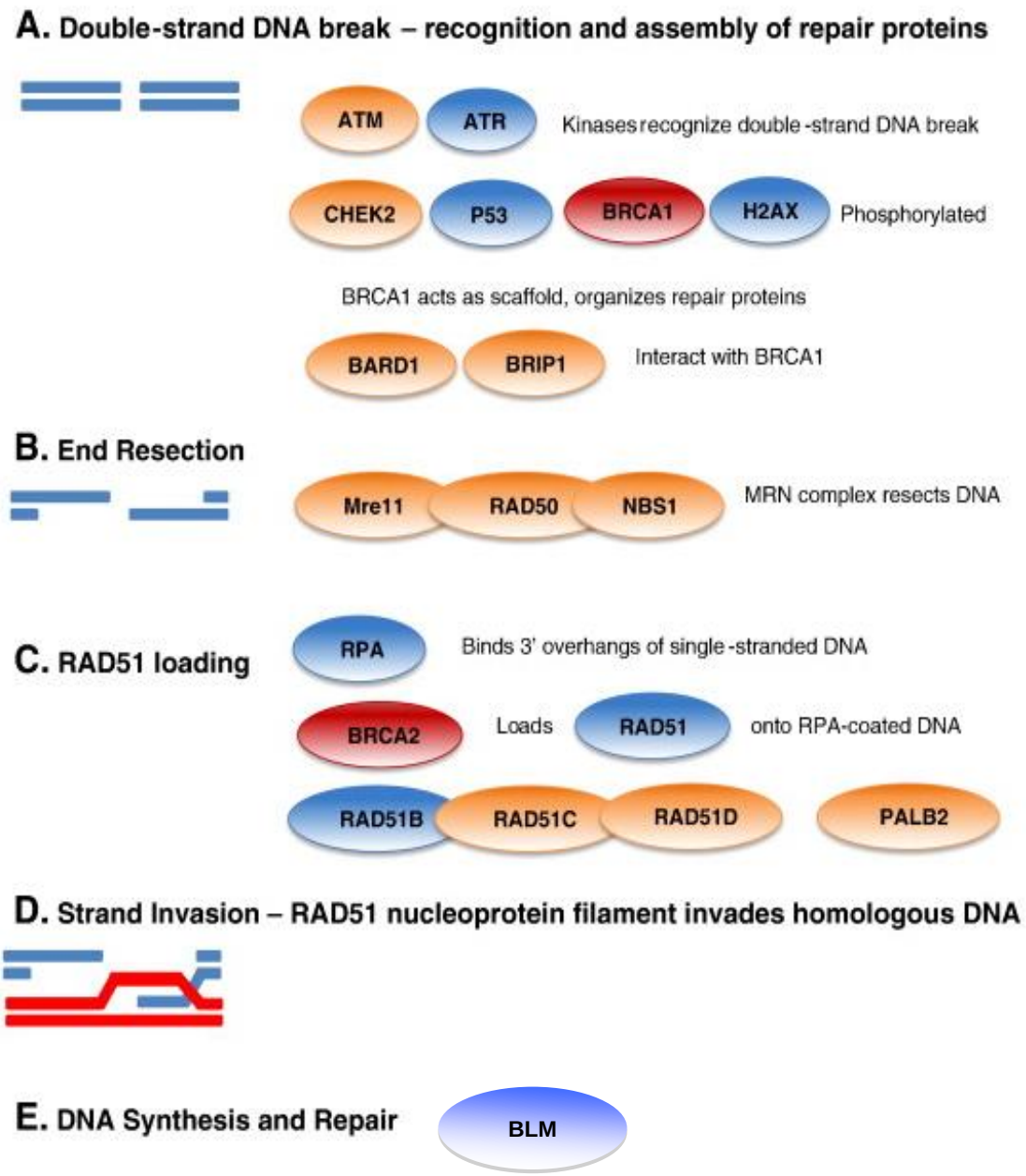


Figure 1.2: Homologous recombination repair proteins of double-strand DNA breaks implicated in hereditary breast and ovarian cancer risk.

The protein products of genes implicated in hereditary breast and ovarian cancer susceptibility are indicated in red (*BRCA1* and *BRCA2*) and orange (Other homologous recombination genes linked to cancer risk). A. Double-strand DNA break is recognized by of ATM and ATR and other repair proteins are assembled. B. DNA is resected by the MRN complex, consisting of MRE11,

RAD50 and NBS1. C. RAD51 is loaded. D. The RAD51 invades the homologous DNA strand. E. DNA repair by the DNA helicases. The homologous DNA strand provides a template for high-fidelity DNA synthesis and repair. Adapted with permission from Christine S. Walsh, *Gynecologic oncology*, 2015.

1.1.4.4 Genetic basis of rare ovarian cancers

The genetic basis of non-epithelial and miscellaneous ovarian tumors has not been well characterized yet, probably because those tumors are very rare. Germ-cell tumor and sex cord-stromal tumors together account for only 10% of all ovarian cancers. Miscellaneous ovarian tumors are even rarer, accounting for less than 1% of all ovarian cancer cases (Scully 1987, Quirk, Natarajan 2005). Unlike epithelial ovarian tumors, germ-cell tumors also occur in the testicle and outside the gonads. Somatic mutations in *KIT*, a gene that is critical for the development of germ cells, have been commonly observed in testicular, ovarian and intracranial germ-cell tumors (Kemmer et al. 2004, Cheng et al. 2011, Wang, Lai 2014). However, no germ-line mutations of *KIT* have been found in familial germ-cell tumors. Germ-line mutations in the histone demethylase gene *JMJD1C* have been found in the germ-cell tumor of the brain (Wang, Lai 2014, Rapley et al. 2004). In some non-epithelial ovarian tumors, particularly Sertoli-Leydig cell tumors from the sex cord-stromal tumor category harbor germ-line or somatic mutations in *DICER1*, a member of RNase III endonuclease that cleaves precursor miRNA into mature miRNA (Heravi-Moussavi et al. 2012, Rio Frio et al. 2011).

Perhaps the only well-characterized rare type of ovarian tumor is the granulosa cell tumor of the ovary- the most common subtype of sex cord-stromal tumors. Granulosa cell tumors are universally caused by a somatic mutation (p.C134W) in the transcription factor-encoding gene *FOXL2* (Wang, Lai 2014, Shah et al. 2009a, Jamieson et al. 2010, Caburet et al. 2015). Nevertheless, genetic causes in most of the ovarian tumor subtypes have not been identified yet. Specifically, prior to the work presented in this thesis, nothing was known about the molecular events underlying the development of the rare small cell carcinoma of the ovary, hypercalcemic type (SCCOHT), the most common type of undifferentiated ovarian tumor in women under age 40 (Clement 2005). Although rare, a few familial cases of SCCOHT have been reported, providing a great resource to study the genetic alterations in both inherited and non-inherited forms of this disease (Clement 2005, Lamovec, Bracko & Cerar 1995, Longy et al. 1996, Martinez-Borges et al. 2009).

1.1.4.5 Clinical management of hereditary ovarian cancer

Genetic testing for patients with increased risk of ovarian cancer offers effective strategies for the prevention and precise treatment of the disease. First, the identification of predisposing mutation carriers may significantly reduce the ovarian cancer incidence. For example, the provision of prophylactic salpingo-oophorectomy, a surgery that removes the ovary and the fallopian tube, has

been associated with 70%-85% reduction in risk of ovarian cancer among *BRCA1* mutation carriers (Domchek et al. 2010). Secondly, the identification of ovarian cancer susceptibility genes has provided several potential targets for treatment. The notion that genes from the homologous recombination DNA repair pathway are frequently mutated in hereditary ovarian cancer allowed the development of poly ADP-ribose polymerase (PARP) inhibitors, which block the PARP enzymes in repairing double-strand DNA breaks, and therefore induce cell death in the absence of DNA repair proteins (Lord, Ashworth 2012). Notably, a new approach that using the combination of platinum-based drugs and PARP inhibitors has improved clinical outcome in recurrent, platinum-sensitive HGSC patients (Luvero, Milani & Ledermann 2014).

Finally, the identification of dominantly inherited cancer susceptibility genes may assist to the classification of diagnostically problematic ovarian tumors. Given the rarity of non-epithelial ovarian tumors and miscellaneous ovarian tumors, pathological misdiagnosis is not uncommon. As an example, SCCOHT is morphologically overlapped and therefore often mixed with granulosa cell and many other ovarian tumors (Clement 2005). Now, thanks to the identification of the *FOXL2* mutations as a predisposing genetic event, granulosa cell tumors have a specific marker for diagnosis (Stewart et al. 2013). These advances show how molecular testing of causal mutations in ovarian tumors may significantly improve the prevention, diagnosis and treatment in ovarian cancers.

1.1.5 Inherited Susceptibility in Breast Cancer

1.1.5.1 Overview of breast cancer

Breast cancer is the most common and second leading cause of death in female cancers (Siegel et al. 2014). The major barrier to the improvement in breast cancer treatment is that the breast tumors are known to be highly heterogeneous, both pathologically and molecularly (Rakha et al. 2009). Gene-expression analysis using microarrays were able to distinguish various breast cancer subtypes. Most of the invasive breast cancers are hormone receptor-positive, characterized by overexpression of the estrogen receptor (ER) or the progesterone receptor (PR) (Perou et al. 2000). ER or PR-positive tumors are usually correlated with a luminal phenotype while hormone receptor-negative tumors are in general associated with a poor prognosis (Schnitt 2010). The later group is further divided into HER2 and basal-like subtypes. The HER2 group is defined by overexpression of human epidermal growth factor receptor 2 (HER2), whereas the basal-like subtype is defined by a 'basal' phenotype as described by pathologists (Schnitt 2010, Foulkes, Smith & Reis-Filho 2010). Basal-like tumors often lack over-expression of ER and HER2, but over-express cytokeratin 5 and epidermal growth factor receptor (EGFR) (Foulkes, Smith & Reis-Filho 2010). In term of treatment response, luminal tumors usually show better outcome in response to endocrine therapy. By contrast, basal-like tumors are sensitive to anthracycline or platinum-based chemotherapy (Badve et al. 2011).

1.1.5.2 Susceptibility genes in hereditary breast cancer

1.1.5.2(i) BRCA1 and BRCA2

Hereditary breast cancer accounts for approximately 10% of all breast cancer cases (Foulkes 2008). The two most important genes *BRCA1* and *BRCA2*, together explain about 30% of hereditary breast cancer cases and 15% of all cases with a familial relative risk (the ratio of the disease risk for the family member of an affected individual to that for the general population) (Couch, Nathanson & Offit 2014). As the first discovered breast cancer gene, *BRCA1* was cloned in 1994 from linked region on chromosome 17 using large families with cases of early-onset breast and ovarian cancers. Soon after that, *BRCA2* was linked to a region on chromosome 13 using families with a high incidence of male breast cancer. Rare, germ-line mutations in both genes are considered to be highly penetrant. Most of the breast cancer-associated variants in the *BRCA* genes are frame-shift insertions or deletions, or nonsense mutations creating a stop codon and result in truncated, non-functional BRCA proteins (Collins 1996). Missense mutations located in the RING finger and BRCT domains for *BRCA1*, or in the DNA binding domain for *BRCA2*, are also highly penetrant (Couch, Nathanson & Offit 2014). Moreover, the loss of the wild-type allele is consistently observed in the tumors. Nevertheless, *BRCA1* and *BRCA2* are rarely mutated in sporadic cases (Venkitaraman 2002).

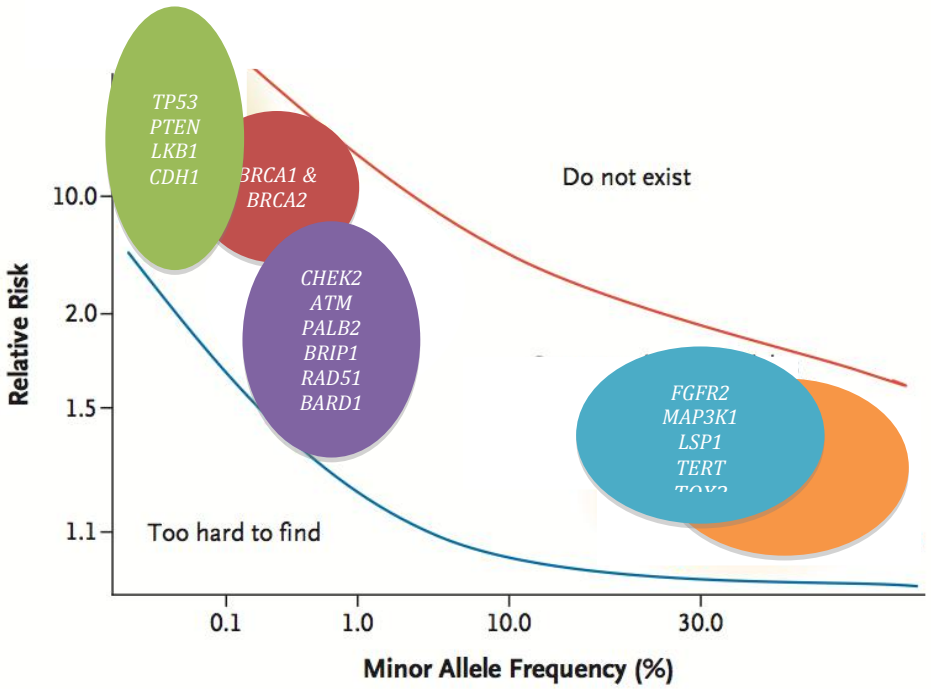
Both *BRCA1* and *BRCA2* have been implicated in a multitude of important processes, including DNA repair, cell-cycle checkpoint control, chromatin remodeling, and transcription. In particular, both proteins interact with RAD51 to form the RAD51 nuclear foci during the S phase and G2 arrest of cell cycle. The RAD51 loci localize to DNA damage regions and repair the double-strand breaks through homologous recombination. All evidence suggests that mutations in the *BRCA* genes can induce accumulated genomic instability (Venkitaraman 2002, Scully, Livingston 2000, Venkitaraman 2004). Interestingly, most breast tumors associated with *BRCA1* cluster with basal-like tumors based on the gene expression profiles (Sotiriou et al. 2003, Schnitt 2010). Finally, the discovery that *BRCA* genes are part of the homologous recombination pathway has resulted in the intensive research into PARP inhibitors as an effective drug for BRCA1/2-deficient patients (Lord, Ashworth 2012, Rouleau et al. 2010).

1.1.5.2(ii) Other high-penetrance to moderate-penetrance genes

Other high-penetrance breast cancer genes, including *TP53*, *PTEN*, *STK11* and *CDH1* account for less than 3% of all familial cases (Couch, Nathanson & Offit 2014, Ripperger et al. 2008) (Figure 1.3). They were mostly discovered by linkage analysis or candidate gene approach in several rare, cancer-susceptibility syndromes associated with high incidence of breast cancer, such as *TP53* in Li-Fraumeni syndrome, *PTEN* in Cowden syndrome, and *STK11* in Peutz-Jeghers

syndrome (Malkin et al. 1990, Liaw et al. 1997, Jenne et al. 1998). Many moderate-risk genes that increase the risk of cancer two to five times, such as *ATM*, *CHEK2*, and *PALB2* together explain another 4% familial cases (Couch, Nathanson & Offit 2014, Ripperger et al. 2008) (Figure 1.3). Most of these genes were discovered by candidate-gene approach, based on their known functions related to the *BRCA* genes and the homologous recombination DNA repair pathway. For example, *PALB2* was a Fanconi anemia gene interacting with *BRCA2* (Rahman et al. 2006). At present, more and more researchers start to believe that more genes found in this category (high to moderate-penetrance) will explain only a small proportion of hereditary breast cancer cases.

A



B

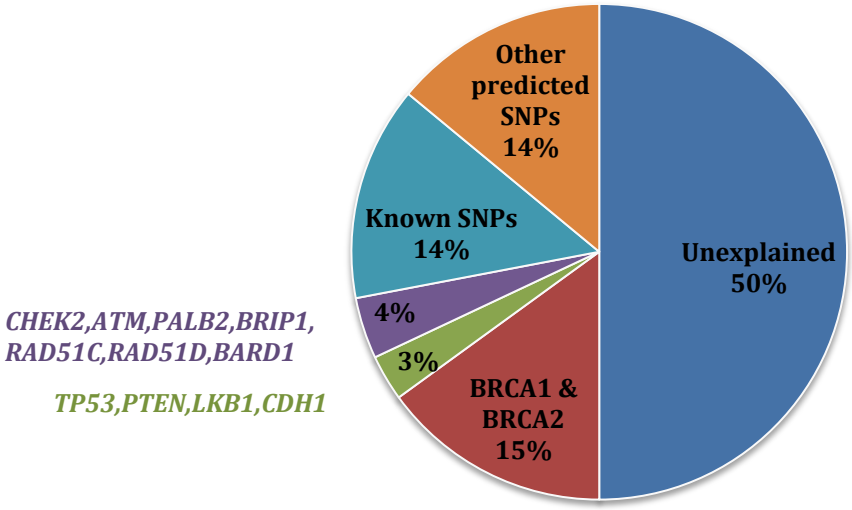


Figure 1.3: The genetic landscape of breast cancer.

A. Identified breast cancer susceptibility genes. The Y-axis is the relative risk for breast cancer associated with mutations in the gene. Genes with high-risk alleles are highlighted in green and red. The moderate-penetrance genes are highlighted in purple. There are probably many more genes in this class, but they can be identified only by studying affected persons in breast cancer families. The common, low-risk genes are shown in blue and orange. Adapted with permission from Foulkes, *New England Journal of Medicine*, 2008.
B. Pie chart shows the estimated percentage contribution of breast cancer alleles. Besides, *BRCA1* and *BRCA2* (purple), all the other high to moderate breast cancer genes (purple and green) are rare. Adapted with permission from Couch et al., *Science*, 2014.

1.1.5.2(iii) Other low-penetrance genes

GWAS has identified at least 94 loci associated with a small impact on breast cancer risk, and approximately 30% of all familial breast cancer cases can be explained by common variants conferring increased risk between 1.1 to 1.6 fold changes (Michailidou et al. 2013, Maxwell, Nathanson 2013, Michailidou et al. 2015). The underlying mechanisms by which these variants contribute to breast cancer might be related to gene-gene interactions or combined effects on gene transcription (Ripperger et al. 2008). For example, three SNPs, namely rs35054928, rs2981578 and rs45631563 in the *FGFR2* gene have been repeatedly identified in GWAS. Evidence suggested that rs35054928 affected the binding with FOXA1 and ERα, whereas rs2981578 affected the binding with E2F1. And three induced factors together enhance the expression of *FGFR2* (Meyer et al. 2013). Moreover, since the variants identified by GWAS are also

common in the healthy population, the clinical utility of them in predictive testing for breast cancer patients is controversial (Couch, Nathanson & Offit 2014, Ripperger et al. 2008).

1.1.5.3 Founder mutations in breast cancer genes

A founder mutation is defined as a specific mutation occurring in high frequency in an ethnically homogeneous as a consequence of a founder effect. A founder population typically results from the rapid growth of a small number of people with shared common ancestry. As an example, the French Canadian population of Quebec contains about six-million descendants of French settlers who colonized the Quebec region between 1608 and 1765, and is notable for the presence of numerous different founder mutations (Vézina et al. 2005, Laberge et al. 2005). With the founder effect, certain rare mutation associated with breast cancer may display higher frequency within the founder population than in ethnically mixed populations (Tonin et al. 1998, Narod, Foulkes 2004). In the Ashkenazi Jewish population, about 2-3% of individuals carry one of three specific founder mutations in *BRCA1* and *BRCA2*, namely 185delAG (c.68_69delAG), 5382insC (c.5266dupC) in *BRCA1* and 6174delT (c.5946delT) in *BRCA2*, and almost all familial breast cancer patients with Ashkenazi Jewish ancestors harbor one of the three mutations (Couch, Nathanson & Offit 2014, Narod, Foulkes 2004).

Founder mutations in breast cancer genes are also found in the French-Canadian population, confirmed through haplotype analysis (Vézina et al. 2005, Oros et al. 2006). In a study of 564 French-Canadian women with early-onset breast cancer, nine founder mutations found in *BRCA1*, *BRCA2*, *PALB2* and *CHEK2* account for 6% of all cases (Ghadirian et al. 2009). In another study of 82 French-Canadian breast cancer families, 45% of them carried a *BRCA1* or *BRCA2* mutation, but only six out of 19 of those mutations were recurrent mutations (Cavallone et al. 2010).

1.1.5.4 The hunt for additional breast cancer susceptibility genes

More than half of the familial breast cancer cases remain unexplained by any known breast cancer susceptibility genes. Mutations in all the identified high to moderate-risk breast cancer genes apart from *BRCA1* and *BRCA2* are rare. In French-Canadian breast cancer families, only one founder mutation in *PALB2* was found among 71 *BRCA1* and *BRCA2* negative families (Tischkowitz et al. 2013). Mutations in other genes, such as *TP53* and *CHEK2* were found in less than 2% of all families (Arcand et al. 2008, Novak et al. 2008). Most, however, do not harbor mutations in those five genes, or any other known breast cancer susceptibility genes. Therefore, further discovery efforts are warranted. It is possible that most of the remaining genetic contributions to this so-called missing heritability are due to the hundreds of common SNPs that remain to be identified

or fully characterized, but these cannot easily explain strong family histories spanning several generations.

It is likely that other moderate-to-high-risk variants will be uncovered through hereditary breast cancer studies. However, researchers have suggested that non-*BRCA1* or *BRCA2* related hereditary breast cancer cases are genetically heterogeneous. Therefore, in order to discover exceptionally rare breast cancer alleles, a careful study design aiming to reduce the underlying genetic heterogeneity is required (Hilbers et al. 2012, Hilbers et al. 2013). Recently, with the advance of WES, a group in Finland revealed that *FANCM* gene was significantly associated with hereditary, tripe-negative breast cancer patients in the Finish founder population (Kiiski et al. 2014). By focusing on sequencing the entire exonic regions, WES provides a unique opportunity to identify novel breast cancer susceptibility genes. Detailed introduction of WES follows in the next section.

1.2 Whole-Exome Sequencing and Data Analysis

1.2.1 Overview of Massively Parallel Sequencing

In 2005, a landmark publication described a method – now is referred as next-generation sequencing (NGS) – able to process massive pyrophosphate-based sequencing in parallel (Margulies et al. 2005). In contrast to Sanger sequencing that relies on the chain termination before reading the nucleotides, sequencing-by-synthesis technology enables massively parallel sequencing by reading the nucleotides as they are incorporated into growing DNA strands (Fuller et al. 2009).

Sequencing-by-synthesis technology has been adopted in all the commercialized NGS platforms (Fuller et al. 2009). The sequencing read length from Illumina HiSeq systems is around 100-bases, and one DNA fragment can be sequenced from both ends such that paired reads are generated. Similar to Sanger sequencing, a Phred-scaled base quality score is used in NGS to quantify the probability of observing an error at the base (Nielsen et al. 2011). In Illumina, an accurate base calling usually has a quality score greater than 30, which infers that the probability of base accuracy is more than 99.9% (Quail et al. 2012).

The development of NGS soon revolutionized the sequencing market. Now it is possible to identify the disease-related variants by comparing the DNA sequence of a patient to a normal genome, at an unprecedented speed and affordable cost. Although whole-genome sequencing of a large number of samples is not feasible yet, some strategies such as targeted re-sequencing of PCR amplicon and whole-exome sequencing have been developed to capture and enrich the genomic regions of interest so sequencing time and cost are further reduced (Ng et al. 2009, Mamanova et al. 2010).

1.2.2 Introduction to WES Technology

WES is an approach of targeted re-sequencing of all protein-coding regions in the human genome (Ng et al. 2009). Although the exome accounts for only 1% of the genome, it harbors most of the rare but highly penetrant, disease-causing mutations (Choi et al. 2009). Previous work using positional cloning of coding-only region has been highly successful in identifying mutations in monogenic diseases and hereditary cancers. Since WES is testing all exons simultaneously, it is considered to be an efficient strategy to search for variants (Bamshad et al. 2011). Therefore, WES is rapidly entering the field of genetic and clinical research before whole-genome sequencing (WGS) (Samuels et al. 2013) (Figure 1.4).

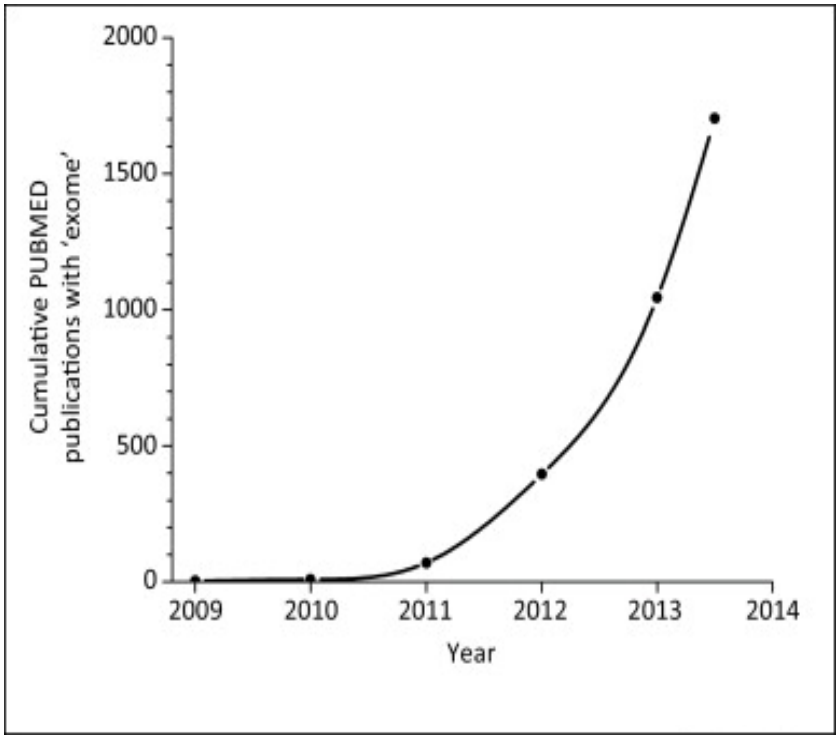


Figure 1.4: The rapid spread of WES application in the field of genetics. Results are produced by PUBMED search of papers using the term of 'exome'. Reproduced with permission from Sameuls et al., *Trends in Genetics*, 2013.

NGS technology requires several steps; among the most important is library preparation. Typical steps in the library preparation of WES include 1) shearing genomic DNA into small fragments; 2) attaching sequencing adapters to both ends of the fragment for the following sequence-by-synthesis initiation; 3) selecting targeted fragments based on hybridization against oligonucleotide microarrays or in-solution capture-probes; and 4) eluting and PCR amplifying the hybridized fragments to achieve an enriched library of the exome (Figure 1.5). As such, all sequencing throughput is focused on the enriched region, which

allows faster data processing and higher per-base coverage for exons, and finally enables more reliable variant calling in the downstream analysis (Haas, Katus & Meder 2011, Majewski et al. 2011).

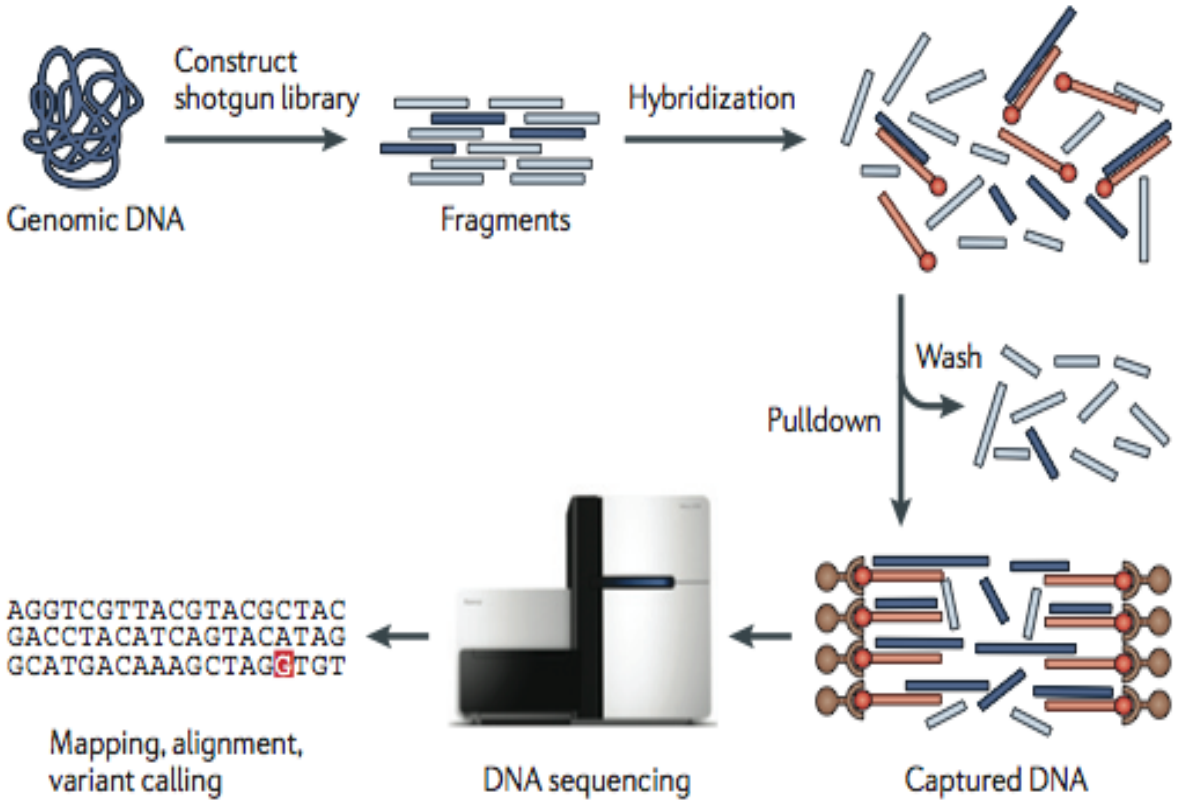


Figure 1.5: Procedure of WES.

Typical steps in WES include 1) DNA capture and PCR amplification to achieve an enriched library of the exome; 2) DNA sequencing using the next generation technology; 3) WES data analysis. Adapted with permission from Bamshad et al., *Nature Reviews Genetics*, 2011.

1.2.3 Analyzing WES Data

1.2.3.1 Sequence alignment

After receiving millions of short sequence reads generated by the sequencer, the first step is to reconstruct the genome or the exome. One way is to assemble the reads to long contiguous sequences sharing the same original template DNA. However, due to the limitation of read length, performing *de novo* assembly for millions of short reads is time-consuming and computational costly (Paszkiewicz, Studholme 2010).

Alternatively, the location of where the obtained sequences were generated from the genome can be determined, such that short reads can be aligned to a previously assembled reference genome from the same species (Flicek, Birney 2009). A mapping quality score is used to measure how confident that a read is not misplaced (Li, Ruan & Durbin 2008). In particular, by using the paired-end method, the length of the sequence to be aligned is increased, and therefore, significantly reduces the probability of multiple mapping locations. As the alignment approach is widely applied in current NGS data analysis, numerous algorithms have been developed to enable higher accuracy and faster processing (Li, Ruan & Durbin 2008, Langmead et al. 2009, Li, Durbin 2009). As an example, Burrows-Wheeler Aligner (BWA) first indexes the reference genome using an effective data compression algorithm called Burrows-Wheeler transform

to allow fast matching of the sequence reads to the reference sequence (Li, Durbin 2009).

Additionally, some algorithms focus on incorporating the effects of genomic alterations in the alignments to improve the identification of single-nucleotide variants (SNVs), small insertions or deletions (indels) and large-scale structural variants. For instance, IndelRealigner implemented in the Genome Analysis Toolkit package (GATK) can target a candidate indel region and locally re-align the nearby reads by minimizing the number of mismatching bases to recover more gap events (DePristo et al. 2011).

1.2.3.2 Point mutation and indel calling

Calling SNVs and small insertions or deletions (indels) is essentially the identification of sites with an alternative base rather than the reference base. An easy way to call variants is to count the number of alleles at each position, by which takes into account the base quality and mapping quality. A heterozygous variant is called when the percentage of the non-reference allele is between 20% and 80% (Nielsen et al. 2011). This method works for calling germ-line variants with high read-depth. More sophisticated algorithms like SAMtools package uses a probabilistic approach to identify the most likely genotype at each position using the base qualities of reads covering that position, such that statistical measurement is provided (Li 2011). To achieve high-quality variant calls, many

analytical pipelines, such as the GATK Best Practices and our in-house bioinformatics workflow are developed to suit various research projects (Auwera et al. 2013).

1.2.3.3 Variant calling errors

Sequencing error rate in NGS is overall higher than Sanger sequencing. In Illumina, the average error rate is between 0.2% to 0.8% (Quail et al. 2012, Kircher, Stenzel & Kelso 2009). Increased sequencing read-depth could reduce the chance of calling random errors. But for some specific site, systematic error can occur repeatedly at a frequency as high as a heterozygous variant. In the Illumina platforms, increased base-calling errors have been observed toward read ends, in high GC content and inverted repeat regions (Ledergerber, Dessimoz 2010, Nakamura et al. 2011). Meanwhile, mapping reads to the reference genome is confounded by repetitive regions, regions affected by gap events, and the presence of mismatched bases in the read (Trapnell, Salzberg 2009). Other sources of errors likely exist and may not be characterized yet. For instance, low-frequency, C>A/G>T artifacts were found particularly in WES because the DNA acoustic shearing process for WES can induce DNA oxidation (Costello et al. 2013). Contamination of foreign DNA in the sample can be another major source of artifacts, which again, should be corrected in variant calling (Flickinger et al. 2015).

1.2.3.4 Variant annotation

Variant annotation provides necessary information to identify which variant among other candidates is associated with the disease phenotype. First, SNVs are annotated by variant types, such as nonsense or missense mutations. Missense mutations can be further annotated by tools such as Sorting Intolerant from Tolerant (SIFT), PolyPhen-2, and MutationTaster to predict how likely the protein change is damaging (Ng, Henikoff 2003, Schwarz et al. 2010, Adzhubei, Jordan & Sunyaev 2013). Other information about the variant can also be useful for decision-making, for example, how conserved the residue is across species; what the allele frequency is from 1000 Genome project; whether the variant has been seen in diseases; and our knowledge about the affected gene. Some package like ANNOVAR can perform different types of annotations for a large list of candidate variants efficiently (Wang, Li & Hakonarson 2010). Other annotation databases, such as Online Mendelian Inheritance in Man (OMIM) can be used to interpret the association of human phenotypes to their causative genes (Amberger, Bocchini & Hamosh 2011).

1.2.3.5 Calling copy number changes

Calling CNVs or other structural variants, such as inversions and translocations using WES data has been challenging to researchers. Due to the fact that WES captures only 1% of the genome, the chance of sequencing the CNV breakpoints

is slim. Thus, detecting CNVs in WES can only be inferred by the read-depth method. Nevertheless, the accuracy of read-depth reflecting true variant is confounded by exome capturing bias, complexity of mapping homologous exons and experimental batch effect. Hence, additional methods for read-depth normalization are required. Some tools such as XHMM and FishingCNV have suggested analyzing a group of samples together, to which similar experimental protocols are used, such that batch-to-batch variations can be removed (Fromer et al. 2012, Shi, Majewski 2013). Recently, a novel method using “off-target” reads from the exome capture has been developed to allow uniformly distributed copy number information for CNV calling (Kuilman et al. 2015).

1.3 Application of Whole-Exome Sequencing Data in Cancer

1.3.1 Complexities of Sequencing Cancer Samples

Unlike germ-line samples, cancer samples can suffer from low quantity, quality and purity. As fresh frozen tumor sample from surgical resection is not always available, most of the tumor DNA for cancer research is obtained from diagnostic biopsy or formalin-fixed and paraffin-embedded (FFPE) sample. As the size of biopsy for diagnosis is minimized, the amount of DNA that can be extracted is limited (Meyerson, Gabriel & Getz 2010). Tumor samples are FFPE-preserved for easier slicing and staining in the microscopically histological analysis, but DNA is usually degraded and cross-linked with proteins (Gilbert et al. 2007). Finally, a cancer sample is a mixture of tumor and normal cells. Thus, extraction protocols with additional steps of purification, as well as a high efficient exome capture kit designed for low DNA inputs are needed. Meanwhile, an assessment of tumor content in the sample before sequencing is advisable where possible.

1.3.2 Calling Low allelic-fraction, Somatic Mutations

Low allelic-fraction variants are commonly observed in tumor samples. The ratio of DNA harboring the mutation in the tumor can be reduced by normal cell contamination during sample preparation or by local CNVs; for example in tumor cells where the wild type allele happens to have duplications in that region.

Moreover, populations of cancer cells are known to be heterogeneous, in that some cancer drivers may present only in a subclone of tumor cells at early stages of the disease (Meyerson, Gabriel & Getz 2010, Gundry, Vijg 2012). Subclonal variants cannot be overlooked because they may expand later in a changing environment, for instance by treatment, and therefore, result in therapy resistance or relapse (Stratton, Campbell & Futreal 2009, Mullighan et al. 2008). For this purpose, many variant callers focus on increasing the sensitivity of detecting low-fraction variants, especially SNVs, the simplest class of mutation (Wilm et al. 2012, Cibulskis et al. 2013, Gerstung, Papaemmanuil & Campbell 2013, Yost et al. 2013). However, given the fact that the rate of somatic, point variation is approximately one per million bases, any variant callers should control the false-positive rate as low as 10^{-6} errors per base (Meyerson, Gabriel & Getz 2010).

Several methods have been designed for calling somatic mutations using tumor and matched normal samples (Table 1.1). Among many tools based on Bayesian probabilistic models, MuTect is probably the most popular one (Wang et al. 2013). MuTect calculates the log-likelihood ratio of each variant, and a true event is called when it exceeds a threshold that is defined by the expected mutation rate and allowed false positive rate. The algorithm of MuTect significantly increases the sensitivity to detect low-fraction variants, but substantially decreases the specificity (Wang et al. 2013). For this reason, users need to apply

additional filtering steps to remove false positive calls, such as using a panel of normal samples to filter missed germ-line variants or false positive calls.

Some algorithms, including LoFreq, deepSNV, and Mutascope are designed to distinguish low allelic-fraction SNVs from errors by interpreting the probability as the frequency of the event at each specific position of the genome (Wilm et al. 2012, Yost et al. 2013, Gerstung et al. 2012). Typically, they test the probability of observing a certain number of reads with the altered base under a null model that defines background error distribution of that specific site, or under an alternative model allowing for true variants. LoFreq treats each read of sequencing at the site as a Bernoulli trial associated with the base quality score. Given a binomial distribution, each Bernoulli trial can have a distinct success probability and the *p*-value is computed from the sum of the probabilities obtained from each Bernoulli trial. Similarly, Mutascope estimates the error rate using the sequencing of matched normal sample followed by a binomial test and then classified germ-line or somatic mutations using a Fisher's exact test.

Based on the algorithm of deepSNV, Shearwater further improves the measurement of site-specific error rate by using the observed errors from a large cohort of control samples, and then models a beta-binomial distribution where the prior knowledge of whether the allele was reported in cancer database is incorporated to increase the sensitivity (Gerstung, Papaemmanuil & Campbell 2013). However, because deepSNV, Shearwater and Mutascope are designed

for targeted re-sequencing, in which reads covering specific amplicon have identical genomic starting positions, the way they filter amplicon-related false-positives may not be suitable for WES, in which random shotgun DNA fragments are sequenced.

Table 1.1: Summary of computational tools for detecting somatic mutations from tumor sample.

Program	Method	Application	Reference
MuTect	Bayesian probability with post-filtering	WES	(Cibulskis et al. 2013)
VarScan2	Fisher exact test and FDR correction	WES	(Koboldt et al. 2012)
LoFreq	Binomial probability and Bonferroni correction	WES	(Wilm et al. 2012)
JointSNVMix	Binomial mixture model	WES	(Roth et al. 2012)
SNVMix	Binomial mixture model	WES	(Larson et al. 2012)
SomaticSniper	Bayesian probability with post-filtering		(Larson et al. 2012)
Strelka	Bayesian probability with post-filtering	WES	(Saunders et al. 2012)
DeepSNV	Binomial probability and Bonferroni or FDR correction	TRS*	(Gerstung et al. 2012)
Shearwater	Binomial probability with prior knowledge	TRS	(Gerstung, Papaemmanuil & Campbell 2013)
Mutascope	Binomial probability and Fisher exact test	TRS	(Yost et al. 2013)

*TRS=Targeted re-sequencing.

1.3.3 Identification of LOH

As mentioned earlier, the ratio of the mutant allele count and total read count can be used to decide the heterozygous or homozygous status of the variant. An event of LOH is called at which a variant presents a ratio close to 50% in the normal sample and 100% in the tumor sample. However, those cut-offs become unclear when read-depth bias and normal cell contamination are present. In VarScan 2, Fisher's exact test is performed to determine if the ratio in the tumor is statically different than in the germ-line (Koboldt et al. 2012). Moreover, chromosomal rearrangements resulting in LOH of cancer-driving genes can be inferred by identifying allelic imbalance events at heterozygous sites. Algorithms such as ExomeCNV and ExomeAI segment the genome based on similar allelic ratios of nearby SNPs, so LOH region and region with allelic imbalance (AI) resulting from copy number changes can be revealed (Sathirapongsasuti et al. 2011, Nadaf, Majewski & Fahiminiya 2015). ExomeAI is also able to identify allelic-imbalanced regions when the matched normal sample is absent, and a control database of WES data is used to filter false positive calls.

1.3.4 WES to Study FFPE Samples

FFPE tissue blocks are commonly used in clinical laboratories, because they are easier for archiving. However, FFPE samples are not ideal for molecular testing,

due to DNA damage resulting from the fixation process. For example, formalin can induce DNA-protein crosslinking and subsequent DNA fragmentation. This affects the quantity and quality of the DNA yield and therefore results in smaller insert size (the distance between properly mapped forward and reverse read pairs) and greater coverage variability in sequencing (Spencer et al. 2013). Moreover, increased C to T or G to A transitions have been observed in FFPE samples (Spencer et al. 2013, Williams et al. 1999, Jasmine et al. 2012). Because fresh-frozen surgical biopsies are rare, robust WES of FFPE samples and methods developed specifically for analyzing WES data of FFPE samples are highly demanded (Spencer et al. 2013, Van Allen et al. 2014).

1.4 Cancer Evolution and Tumor Progression

1.4.1 Understanding The Dynamic Evolution of Cancer

Cancer is evolved from a single common ancestor cell with acquired somatic mutations conferring a growth advantage to the tumor microenvironment. The underlying driving force is genomic instability, which randomly creates genetic variants in the cancer genome (Nowell 1976). These variants are broadly divided into two groups. “Driver” mutations are acquired mutations giving advantages to the clonal expansion and are positively selected by the environmental pressure; whereas the remainders are “passenger” mutations, which are generated in the common ancestor cell by chance, but are not required during the tumor progression (Stratton, Campbell & Futreal 2009).

As a result of the genetic diversity and natural selection, the tumor cell population can be one dominant, “winning” clone, or most likely a mixture of multiple sub-clones. Each of the sub-clones may be capable of invading tissues, and giving rises to metastasis and relapse (Yates, Campbell 2012). Moreover, the clonal architecture can be re-shaped by a changed environment provided by, such as a distant organ or a cancer treatment. Some previous subclones carrying organ-specific or treatment-resistant variants may find themselves able to expand with better fitness in response to the changed environment (Yachida et al. 2010, Marusyk, Almendro & Polyak 2012). Thus, understanding the dynamic process of

cancer evolution is critical for the prediction of cancer metastasis and treatment outcome.

1.4.2 Cancer Metastasis and Relapse

Metastasis is the most deadly feature of cancer. Several models have been proposed to explain the progression patterns. In a linear model, the most advanced and aggressive clone evolves as dominant in the primary tumor, and then gives rise to the metastasis. In a parallel model, the metastatic cells disseminate early from the original tumor cells and evolve independently with the primary tumor (Klein 2009). In the self-seeding model, tumor cells can colonize not only distant sites, but also the primary tumor itself (Kim et al. 2009). Sequencing analysis in the paired, primary and metastatic tumors suggested that some variants with high frequencies in the metastatic stage were present as low frequent, subclonal variants in the primary tumor (Shah et al. 2009b, Ding et al. 2010). Previous work in pancreatic cancer suggested that additional, *de novo* mutations were required for metastasis (Yachida et al. 2010). Similarly, recurrent tumors usually respond poorly to the initial therapy, indicating the emergence of therapy-resistant variants either pre-existing before the treatment, or newly gained during the treatment (Marusyk, Almendro & Polyak 2012). Genomic study of drug response in acute myeloid leukaemia suggested that the relapse clone required additional mutations to expand, either from the major clone in the primary tumor, or the subclone that escaped from the first-line treatment (Ding et

al. 2010). Taken together, cancer heterogeneity is considered as the major barrier in treating patients with secondary cancer and relapse.

1.4.3 Studying Tumor Progression

To understand the mechanism underlying the tumor progression, many studies applied NGS with multi-sampling strategies by comparing the genomes of tumor samples collected at different locations, different time points, or both, depending on the objective of the study (Yates, Campbell 2012). Remarkably, cancer heterogeneity exists among different types of tumors, different patients, different tumors within a single patient, as well as different cells within a single tumor (Marusyk, Almendro & Polyak 2012). Considering that NGS randomly sequences DNA fragments representing the genome of an individual cell within a tumor sample, variants from a subclone can be covered. To determine whether a variant is subclonal, the fraction of reads supporting a variant can be used to reflect the allele frequency of the variant in the tumor cell population (Yates, Campbell 2012). However, the ratio of mutant reads should be normalized by local CNVs and the level of normal cell contamination. Several subclonal variants contributed to the metastasis and therapy-resistance have been successfully identified in breast cancer, pancreatic cancer, and acute myeloid leukemia (Yachida et al. 2010, Ding et al. 2010). These variants can be present at low frequency at the diagnostic stage, but highly enriched at the relapse stage (Ding et al. 2012). Therefore, accurate identification of low allelic-fraction variants is

warranted. Finally, a study of renal carcinomas demonstrated the capability of WES in the tumor progression analysis (Gerlinger et al. 2012). It should be noticed that non-genetic factors, such as cancer stem cells, might also contribute to the cancer heterogeneity and tumor progression through self-renewal (Hanahan, Weinberg 2011). However, this thesis is mainly focusing on understanding the development of cancer from the genetic aspect.

1.5 Rationale and Objectives of Study

Recent advances in WES have revolutionized the study of Mendelian or monogenic disease. Advantages of WES include a “hypothesis-free” analysis for gene discovery with higher coverage in a shorter time frame. Hereditary cancer is a genetic disorder, and the application of WES in inherited forms of cancers has proved adept at revealing causal genetic events in a cost-effective manner. Moreover, genes identified from familial cancers often explain non-hereditary cancers. These discoveries in cancer genetics have informed the implementation of genetic testing, targeted therapy and precision cancer medicine.

Our group has developed a WES pipeline which automates the computational analysis of WES data and enables me to study a large number of samples concurrently. **The first objective of this thesis is to discover new cancer genes from cancer families, using WES as the primary investigative tool.** As somatic mutations arisen from hematopoietic stem cells can be mis-interpreted as germ-line variants from deep sequencing of blood-derived DNA, variants identified from blood samples will be validated in segregation analysis of the cancer families. In the first project, I focused on SCCOHT – a type of miscellaneous ovarian cancer, for which the genetic defect is unknown. A few SCCOHT families have been identified and my collaborators collected three of them, allowing me to study the inherited factors in SCCOHT. I aim to identify the

causal gene or genes underlying familial SCCOHTs. To better understand this disease, the landscape of the SCCOHT genome will be characterized using WES data from either familial or non-familial cases.

Using WES to uncover new breast cancer susceptibility genes is more challenging, because barring *BRCA1* and *BRCA2* mutations, moderate to high penetrant breast cancer alleles are exceptionally rare. This indicates that pre-selection of the breast cancer cases under study is needed to reduce the underlying heterogeneity. Thus, I hypothesize that by focusing on the French-Canadian population with known population-specific mutations in breast cancer susceptibility genes, I will have a better chance to identify novel genes with breast cancer associated alleles. This approach will allow me to identify genes for which multiple families share a common mutant allele, using a limited number of breast cancer families. As discussed above, I will likely find alleles conferring a relative risk for breast cancer of three to five, with a frequency of 0.001 among the general French Canadian population by WES.

Challenges remain in detecting somatic mutations. DNA derived from FFPE samples is commonly present in low quantity and quality. Tumor purity is difficult to be inferred from pathological examination. These issues profoundly affect the accuracy of variant calling. In particular, low allelic-fraction variants are commonly observed in tumor samples, owing to normal tissue contamination, local copy number change and the nature of cancer heterogeneity. When the

number or ratio of non-reference reads observed from WES data is low, true variants can be buried among artifacts. Furthermore, subclonal variants present at low frequency in the primary stage may significantly contribute to clonal expansion and therapy-resistance, thus emphasizing the importance of precisely detecting low-fraction variants.

The second objective of this thesis is to improve our knowledge of somatic mutations in the cancer exome. In an attempt to understand in somatic mutations contributing to cancer progression, multiple tumor samples collected in the diagnostic, metastatic and recurrent stages from a single patient will be analyzed using WES. Finally, current softwares are insufficient to identify low allelic-fraction SNVs on the whole-exome level, with high sensitivity and low false positive rate. However, low allelic-fraction variants can be important in driving cancer development and relapse. Therefore, I will develop a computational tool dedicated to accurately call low allelic-fraction, somatic SNVs from WES data of tumor samples. My goal is to reach high sensitivity and improved specificity using this method.

Chapter 2. Genomic Characterization Revealed *SMARCA4* Alterations as the Major Genetic Cause of Malignant Small Cell Carcinoma of the Ovary, Hypercalcemic Type

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

"Germline and Somatic mutations in *SMARCA4* characterize the small cell carcinoma of the ovary, hypercalcemic type" Leora Witkowski, Jian Carrot-Zhang (co-first author), Steffen Albrecht, Somayyeh Fahiminiya, Nancy Hamel, Eva Tomiak, David Grynspan, Emmanouil Saloustros, Javad Nadaf, Barbara Rivera, Catherine Gilpin, Ester Castellsagué, Rachel Silva-Smith, François Plourde, Mona Wu, Avi Saskin, Madeleine Arseneault, Rouzan G Karabakhtsian, Elizabeth A Reilly, Frederick R Ueland, Anna Margiolaki, Kitty Pavlakis, Sharon M Castellino, Janez Lamovec, Helen J Mackay, Lawrence M Roth, Thomas M Ulbright, Tracey A Bender, Vassilis Georgoulias, Michel Longy, Andrew Berchuck, Marc Tischkowitz, Inga Nagel, Reiner Siebert, Colin J R Stewart, Jocelyne Arseneau, W Glenn McCluggage, Blaise A Clarke, Yasser Riazalhosseini, Martin Hasselblatt, Jacek Majewski & William D Foulkes, *Nature Genetics*, 46, 438-443 (2014).

2.1 Introduction

SCCOHT is a rare type of miscellaneous ovarian tumor, featuring with early age of onset and the presence of hyperchromatic cells (Clement 2005). In a study of 150 SCCOHT cases, a mean diagnosis age of 23.9 years was reported, with 62% of them having hypercalcemia (Young, Oliva & Scully 1994). The long-term survival rate with an early stage of the disease was reported as 33%, suggesting that SCCOHT is also a highly aggressive tumor (Young, Oliva & Scully 1994). Before our work, no SCCOHT-specific markers had been available, and the genetic cause of SCCOHT had never been identified. Therefore, no effective treatments have been developed for SCCOHT. Although most patients underwent surgery, recurrence was found in 65% of 126 SCCOHT cases studied (Estel et al. 2011). A few SCCOHT families presenting autosomal dominant transmission have been identified (Lamovec, Bracko & Cerar 1995, Longy et al. 1996, Martinez-Borges et al. 2009). Using WES as the primary investigative tools, I studied the genetic cause underlying familial SCCOHTs, validated the candidate gene in additional SCCOHT cases, and finally characterized the landscape of the SCCOHT genomes using data obtained from WES (Figure 2.1).

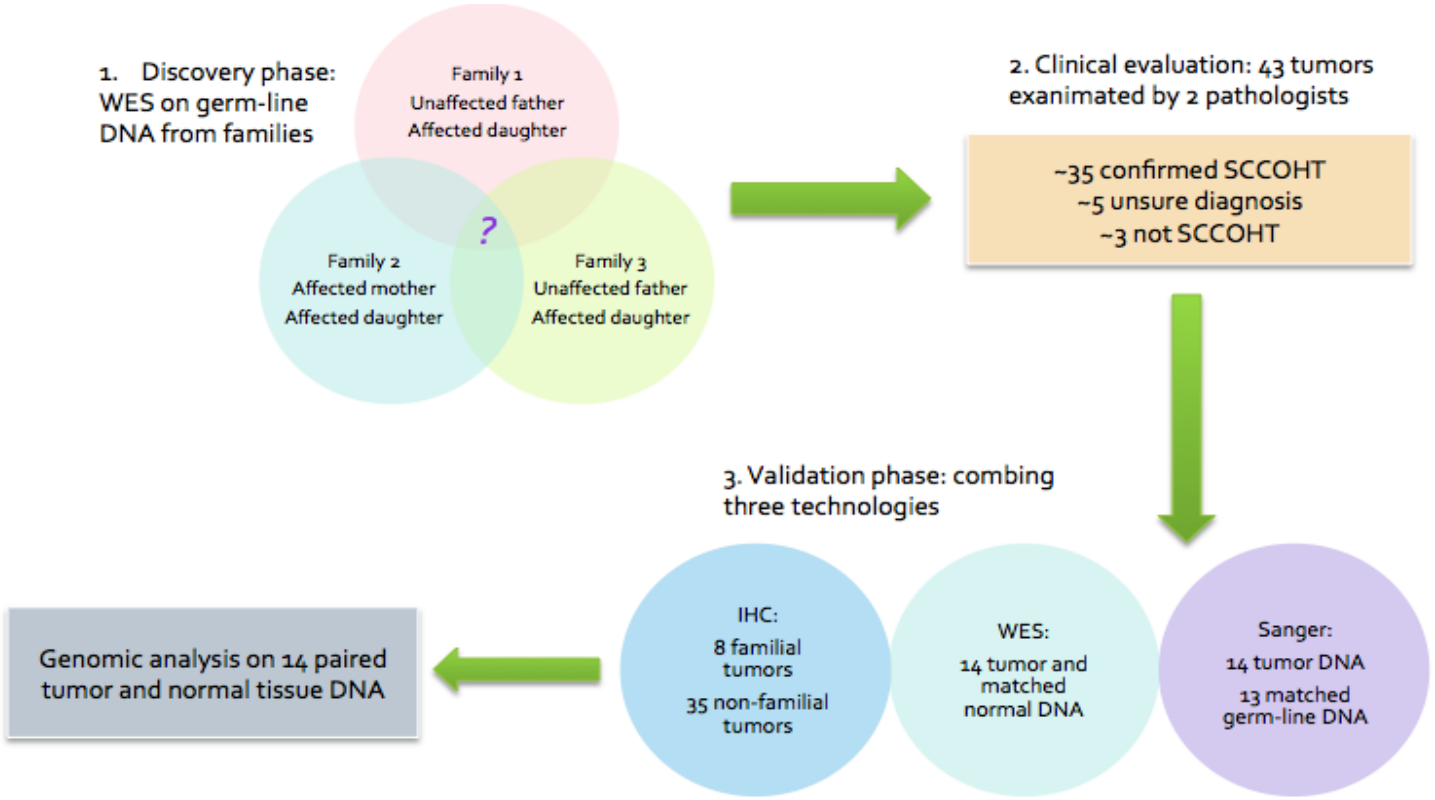


Figure 2.1: SCCOHT study design.

WES was originally used to detect variants in three SCCOHT families. As DNA was unavailable for the affected mothers from families 1 and 3, DNA from the unaffected fathers was used, in order to rule out germ-line variants unrelated to the disease. I looked for deleterious variants that were present in the affected daughter and were not present in the unaffected father. Adapted from Witkowski & Carrot-Zhang et al., *Nature Genetics*, 2014.

2.2 Materials and Methods

2.2.1 Library Preparation

Whole-exome library preparation, exome capture and sequencing were performed using our standard protocols at the McGill University and Génome Québec Innovation Centre. Blood-derived DNA (3 µg) from each subject underwent exome capture using Agilent SureSelect V4 kit following the manufacturer's protocols. The sequencing was subsequently performed using Illumina HiSeq 2000 with paired-end 100-bp reads. For FFPE tissue-derived DNA, 50 ng for each case was captured using the Nextera Rapid-Capture Exome kit and was sequenced on an Illumina HiSeq 2500 sequencer.

2.2.2 Pipeline of WES Data Analysis

WES data was processed using our in-house data analysis pipeline (Figure 2.2). Sequencing reads generated from the sequencer first underwent quality control using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Adaptor sequences and low-quality bases were removed from reads using the Fastx toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). High-quality trimmed reads were then aligned to the UCSC hg19 reference genome with BWA version 0.5.9 (Li, Durbin 2009). Only read pairs with both mates present were subsequently used. Indels were re-aligned using GATK IndelRealigner (McKenna et al. 2010). GATK was

also used to assess capture efficiency and coverage of consensus coding sequence (CCDS) bases (McKenna et al. 2010). Reads that were marked as PCR duplicates by Picard were removed (http://picard.sourceforge.net/). SNVs and indels were called by SAMtools mpileup and annotated by ANNOVAR (Li 2011, Wang, Li & Hakonarson 2010).

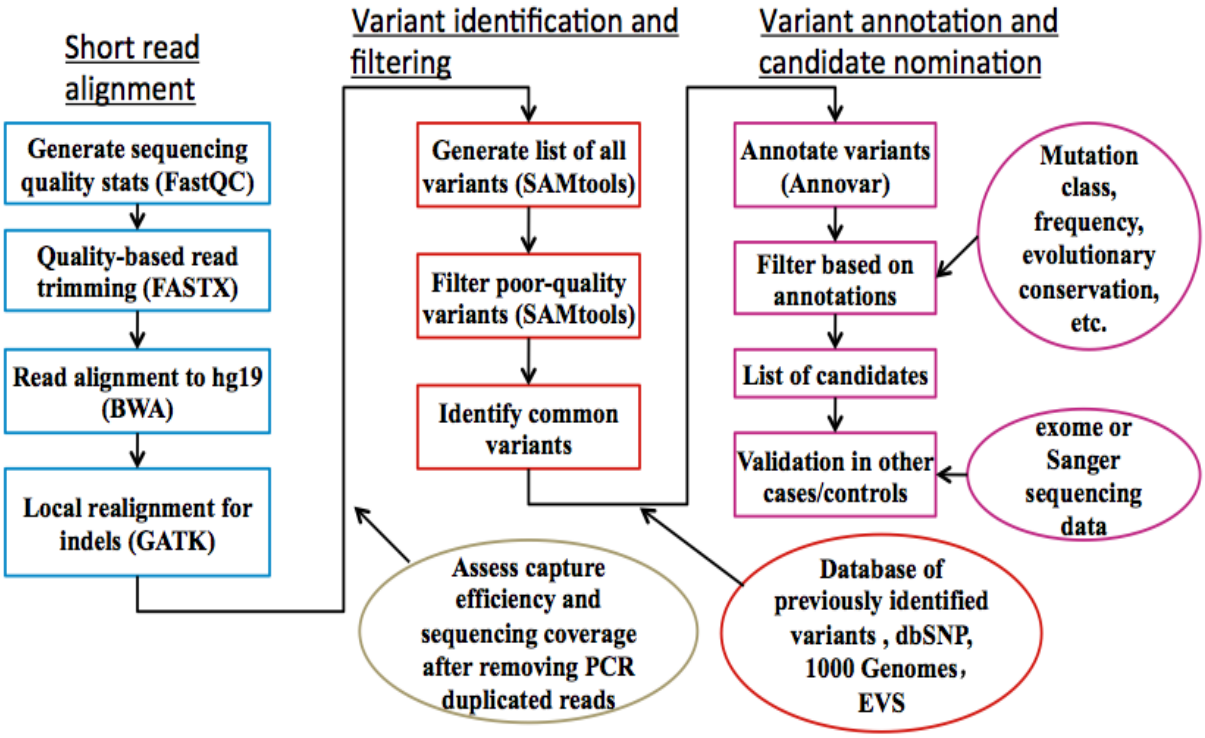


Figure 2.2: **Pipeline of WES data analysis used in this study.**

Three major steps are performed in this pipeline, including sequencing read alignment, variant calling and filtering, and variant annotation. 1000 Genomes = 1000 Genomes Project database; EVS = NHLBI Exome Sequencing Project Exome Variant Server; Mutation class = variants most likely to damage the protein (nonsense, splice-site, missense mutations and frame-shift indels).

2.2.3 Mutation Detection

To remove common variants and false positive calls in germ-line samples, only variants with a minimum of 5 reads supporting the alternative variant were retained. Then, only variants representing >5% or >15% of reads were called as SNVs or indels, respectively. Only heterozygous variants, defined as allelic-fraction between 20-80% were retained. The Integrative Genomics Viewer (IGV) was used for the manual visualization and examination of all variants called (Robinson et al. 2011). Variants that had been previously reported in Single Nucleotide Polymorphism Database (dbSNP), or NHLBI Exome Sequencing Project Exome Variant Server (EVS), or our in-house non-cancer (~1000 exomes sequenced in our group) database were further excluded for further analysis.

2.2.4 Genome-wide AI analysis

ExomeAI was performed to identify the region with allelic imbalance or LOH across the genome (Nadaf, Majewski & Fahiminiya 2015). All polymorphic variants with total read count greater than or equal to 10, mapping quality of greater than or equal to 30 and BAF between 0.05 and 0.95 were considered in the AI analysis, where BAF is the read count of non-reference base divided by the total read coverage at the variant site.

2.3 Results

2.3.1 The Discovery of *SMARCA4* in Familial SCCOHT Cases

Using WES data of DNA from two individuals in each of the three SCCOHT families, I identified 221, 233, and 433 nonsense, missense and UTR variants from family one, family two and family three, respectively (Figure 2.3, Table 2.1). Those variants were either shared between affected mother and daughter (family two), or present in the affected daughter, but not in the unaffected father (family one and family three, where mother's DNA was not available). After excluding variants reported in 1000 genome with minor allele frequency less than 0.0005, 178, 168 and 335 variants were retained in each family, respectively. Six genes (*SMARCA4*, *LAMB2*, *CDYL*, *LGALS12*, *PRSS8*, *NEFH)* were commonly mutated in more than one family (Table 2.1, Table 2.2). Mutations in *CDYL*, *NEFH* and *PRSS8* were in UTRs. *SMARCA4* was the only gene mutated in affected women in all three families.

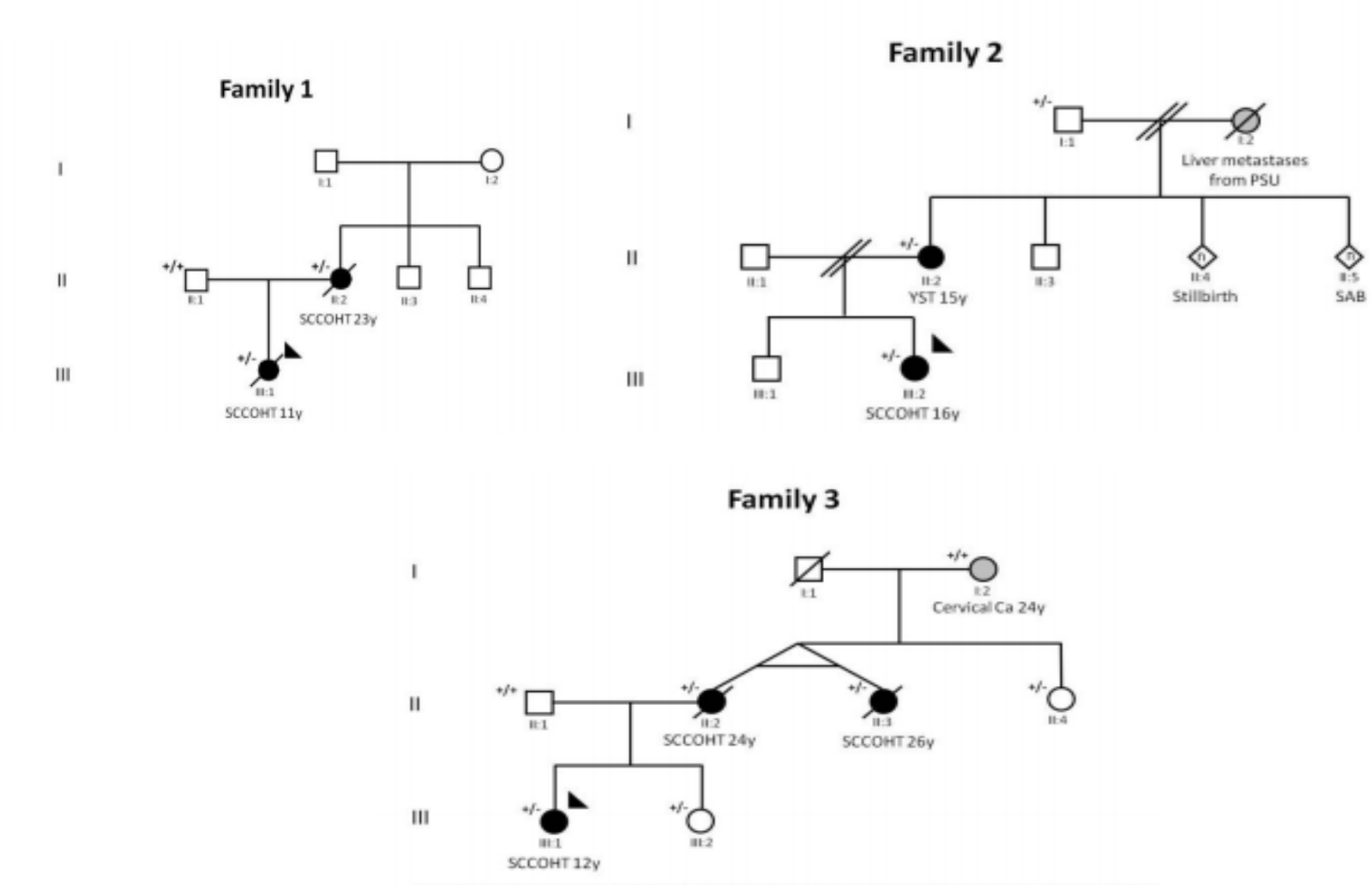


Figure 2.3: Pedigrees of three families studied by WES.

In family 1, members II:2 and III:1 correspond to samples FA1a and FA1b, respectively, in Table 2.2. In family 2, members II:2 and III:2 correspond to samples FA2a and FA2b, respectively, in Table 2.2. In family 3, members II:2 and III:1 correspond to samples FA3a and FA3b, respectively, in Table 2.2. SAB, spontaneous abortion; SCCOHT, small cell carcinoma of the ovary, hypercalcemic type; YST, yolk sac tumor; PSU, cancer, primary site unknown; +/+, wild-type for the familial *SMARCA4* mutation; +/− , heterozygous for the familial *SMARCA4* mutation. A diagonal line through a symbol indicates that the person is deceased. WES was carried out using DNA from individuals II:1 and III:1 (family 1); II:2 and III:2 (family 2); and II:1 and III:1 (family 3). Individual II:4 from family 3 is at high risk for SCCOHT and has consented to a preventive bilateral oophorectomy. Adapted from Witkowski & Carrot-Zhang, *Nature Genetics*, 2014.

Table 2.1: Variant prioritization steps in the analysis of combined exome data of 4 affected individuals.

Reproduced from Witkowski & Carrot-Zhang, *Nature Genetics*, 2014.

Variant prioritization	**FA1a**	**FA2a**	**FA2b**	**FA3a**
Nonsynonymous, splicing, coding indel and UTRS	221	233		433
After excluding variants reported in 1000 genome (MAF < 0.0005)	178	168		335
shared variants in Family 1 and 2	12*			
Shared variants in Family 1 , 2 and 3	3**			

* 12 variants corresponding to 6 genes : *SMARCA4, LAMB2, CDYL, LGALS12, PRSS8, NEFH* (mutations in *CDYL, NEFH, PRSS8* were in UTRs)

**3 variants corresponding to 1 gene: *SMARCA4*

Table 2.2: List of variants in genes mutated in affected members of at least two families.

Only variants with MAF < 0.0005 are listed. Adapted from Witkowski & Carrot-Zhang, *Nature Genetics*, 2014.

Genomic Position	Variation	Reference allele	Altenate allele	FA1b	FA2a	FA2b	FA3a	Protein Change	Gene	rsID	MAF from 1000genomes	EVS MAF
chr3: 49159227	nonsynonymous SNV	C	G	het	-	-	-	p.D1664H	*LAMB2*	.	0	0
chr3: 49167110	nonsynonymous SNV	G	A	-	het	Het	-	p.T482I	*LAMB2*	.	0	0
chr6: 4776911	5' UTR	G	C	-	het	Het	-		*CDYL*	.	0	0
chr6: 4935926	nonsynonymous SNV	T	C	het	-	-	-	p.I158T	*CDYL*	.	0	0
chr8: 10469846	nonsynonymous SNV	C	A	-	het	-	-	p.D588Y	*RP1L1*	rs200344135	0	0.000554
chr8: 10480144	nonsynonymous SNV	G	A	-	-	het	-	p.R190C	*RP1L1*	rs202110498	0.0005	0.001564
chr11: 63276410	stopgain SNV	C	T	-	het	het	-	p.R68X	*LGALS12*	rs141304527	0	7.70E-05
chr11: 63279237	stopgain SNV	C	A	het	-	-	-	p.S143X	*LGALS12*	.	0	0
chr11: 76867116	nonsynonymous SNV	G	A	-	het	het	-	p.R150Q	*MYO7A*	rs202245413	0	0.000241
chr11: 76891513	nonsynonymous SNV	G	A	-	-	-	het	p.E894K	*MYO7A*	.	0	7.90E-05
chr14: 65234378	splicing	C	T	-	het	het	-		*SPTB*	.	0	0
chr14: 65241845	nonsynonymous SNV	T	G	-	-	-	het	p.K1614Q	*SPTB*	.	0	0
chr14: 65262093	nonsynonymous SNV	C	T	-	het	-	-	p.D536N	*SPTB*	rs145675502	0.0005	0.000692
chr16: 2983169	nonsynonymous SNV	G	C	-	-	-	het	p.G278R	*FLYWCH1*	.	0	0
chr16: 2983203	nonsynonymous SNV	A	G	-	het	het	-	p.Y289C	*FLYWCH1*	.	0	0

chr16: 31146859	5' UTR	C	A	-	het	het	-		*PRSS8*	.	0	0
chr16: 31146928	5' UTR	A	G	het	-	-	-		*PRSS8*	.	0	0
chr17: 8076664	3' UTR	G	A	-	-	-	het		*TMEM107*	.	0	0
chr17: 8076835	3' UTR	T	C	-	-	het	-		*TMEM107*	rs201558321	0	0
chr17: 8076910	3' UTR	G	C	het	-	-	-		*TMEM107*	.	0	0
chr19: 11097152	stopgain SNV	C	T	-	het	het	-	p.Q215X	*SMARCA4*	.	0	0
chr19: 11132398	splicing	C	G	-	-	-	het		*SMARCA4*	.	0	0
chr19: 11145809	splicing	G	A	het	-	-	-		*SMARCA4*	.	0	0
chr20: 57766673	nonsynonymous SNV	C	T	-	-	-	het	p.S200F	*ZNF831*	rs201581384	0	0.003226
chr20: 57768704	nonsynonymous SNV	G	T	-	het	het	-	p.G877V	*ZNF831*	rs200405760	0	0.000652
chr22: 29885572	nonframeshift insertion	AGT	AGTTCCCTGAGAAGGCCAAGT	-	-	het	-	p.S654delinsFPEKAKS	*NEFH*	.	0	0
chr22: 29886718	3' UTR	C	G	het	-	-	-		*NEFH*	.	0	0.000155
chrX: 49031926	3' UTR	G	A	-	het	het	-		*PRICKLE3*	.	0	0
chrX: 49034411	nonsynonymous SNV	G	A	-	-	-	het	p.R296C	*PRICKLE3*	.	0	0

MAF = Minor allele frequency; het = heterozygous; rsID: ID of previously seen variants from dbSNP.

Therefore, I focused on mutations identified in the *SMARCA4* gene. In family one, a germ-line splice-site mutation, c.4071+1G>A, was found in the affected daughter, and was later confirmed in the affected mother by Sanger (Figure 2.3, Figure 2.4). The daughter's tumor showed LOH in the genome-wide AI analysis using WES data (Table 2.3). Additionally, Sanger sequencing identified a somatic mutation in the mother's tumor, c.1027delG, encoding p.Val343Cysfs*68. In family two, WES data showed that the mother and daughter carried an identical *SMARCA4* germ-line mutation, c.643C>T, encoding p.Gln215. The daughter's tumor carried a frameshift mutation, c.1687_1700delAACCTCACGGAGCT, encoding p.Asn563Glyfs*82 (Figure 2.4). The mother's tumor was unavailable. In family three, WES data showed that the daughter carried an intronic mutation 3 bp from an intron-exon junction, c.2617-3C>G, which was inherited from the mother (Figure 2.4). The tumor showed LOH for this mutation (Table 2.3). My colleague later confirmed that this mutation disrupted the splice site, causing a part of intron 18 retained in the cDNA leading to the introduction of a premature stop codon, and therefore resulted in a transcript subjected to nonsense-mediated decay (Witkowski et al. 2014).

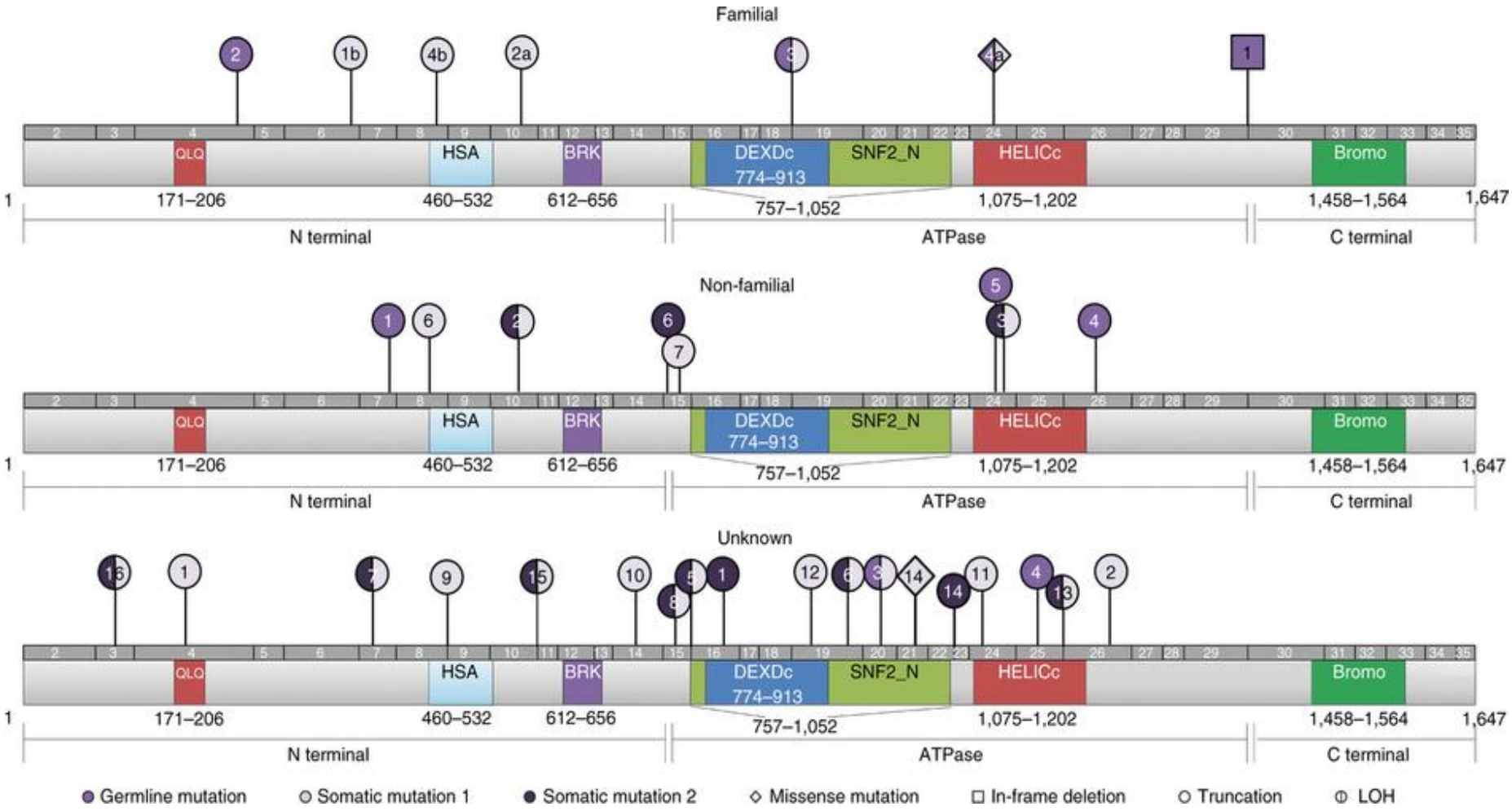


Figure 2.4: Location of mutations found in *SMARCA4* in SCCOHT.

Numbers in each lollipop correspond to the case numbers in Table 2.3. A representation of the primary structure of the protein is shown with exon locations noted above. For cases with unknown family history in which two mutations were found in tumors with no corresponding germline DNA, mutations were arbitrarily denoted as 'somatic mutation 1' and 'somatic mutation 2'. It is possible, however, that one of these mutations is a germline mutation. QLQ = Gln, Leu, Gln motif; HAS = helicase/SANT-associated domain; BRK = Brahma and Kismet domain; DEXDc = DEAD-like helicase superfamily domain; SNF2_N = SNF2 family N-terminal domain; HELICc = helicase superfamily C-terminal domain; Bromo = bromodomain. Reproduced from Witkowski & Carrot-Zhang et al., *Nature Genetics*, 2014.

Table 2.3: SCCOHT with at least one *SMARCA4* mutation.

Reused from Witkowski & Carrot-Zhang et al., *Nature Genetics*, 2014.

Sample ID	Germline Mutation		Somatic Mutation 1		Somatic Mutation 2		IHC
Familial							
FA1a	c.4071+1 G>A	Not Applicable	c.1027_1027delG	p.Val343Cysfs*68	Not Applicable		Loss
FA1b			LOH		Not Applicable		Loss
FA2a	c.643 C>T	p.Gln215*	Not Available		Not Applicable		Loss
FA2b			c.1687_1700delAACCTCACGGAGCT	p.Asn563Glyfs*82	Not Applicable		Loss
FA3a	c.2617-3 C>G	Not Applicable	Not Available		Not Applicable		Loss
FA3b			LOH		Not Applicable		Loss
FA4a	c.3239G>A	p.Gly1080Asp	LOH		Not Applicable		Normal
FA4b			c.1326_1326delC	p.Ser442Argfs*59	Not Applicable		Loss
Non-familial cases							
NF1	c.1224_1226delGCTinsAG	p.Leu409Glyfs*2	Not present		Not Applicable		Loss
NF2	c.1663 C>T	p.Gln555*	LOH		Not Applicable		Loss
NF3	Not Present		c.3496C>T	Gln1166*	LOH		Loss
NF4	c.3638_3638delA	p.Lys1213Argfs*3	Not present		Not Applicable		Loss
NF5	c.3480_3481insG	p.Leu1161Alafs*15	Not present		Not Applicable		Loss
NF6	Not Present		c.2129_2129delC	p.Lys711Serfs*63	c.1378C>T	Gln460*	Loss
NF7	Not Present		c.2245_2246insA	Met749Asnfs*75	Not Present		Loss
Unknown							
UN1	Not Available		c.561C>G	p.Thr187*	c.2362C>T	Gln788*	Loss
UN2	Not Available		c.3676C>T	p.Gln1226*	Not Present		Loss
UN3	c.2932C>T	p.Arg978*	LOH		Not Applicable		Loss

UN4	Not Present		c.3531_3531delC	p.Trp1178Glyfs*38	c.4687_4687 delG	p.Ile1564Serfs *32	Loss
UN5	Not Present		c.2275-1G>T	Not Applicable	LOH		Loss
UN6	Not Available		c.2838_2838delC	p.Phe947Leufs*3	LOH		Loss
UN7	c.1141 C>T	p.Arg381*	LOH		Not Applicable		Loss
UN8	Not Available		c.2190_2191insG	p.Tyr731Valfs*10	LOH		Loss
UN9	Not Available		c.1420+1 G>T	Not Applicable	Not Present		Normal
UN10	Not Available		c.2049_2049delC	p.Val684Trpfs*90	Not Present		Loss
UN11	Not Available		c.3244_3244delT	p. Phe1082Leufs*24	Not Present		Loss
UN12	Not Available		c.2766G>A	p.Trp922*	Not Present		Loss
UN13	Not Available		c.3546+1 G>T	Not Applicable	LOH		Loss
UN14	Not Available		c.2915T>C	p.Leu972Pro	c.3168+1G> C	Not Applicable	Loss
UN15	Not Available		c.1761+2T>A	Not Applicable	LOH		Loss
UN16	Not Available		c.233_237delinsACC	Ser78Tyrfs*3	LOH		Loss

Not Applicable = germline and somatic mutation already identified; Not Available = DNA/tissue sections not available; Not present = no mutation present; IHC=immunohistochemistry

2.3.2 Validation of *SMARCA4* in More SCCOHT Cases

In the second phase of this study, I analyzed WES data from 12 SCCOHT tumors. I first targeted on *SMARCA4* to validate that deleterious *SMARCA4* mutations are common in SCCOHTs. I found that *SMARCA4* mutations were present in all the 12 cases, including two cases showing bi-allelic alterations (Table 2.3). The locations of identified mutations were scattered along the gene, and nearly all of them were predicted to truncate part of the functional domains of the protein, suggesting the tumor suppressor role of *SMARCA4* in SCCOHT (Figure 2.4). Given the fact that the loss of *SMARCA4* protein was observed in the vast majority of SCCOHT cases studied, we suggest that *SMARCA4* gene is fully inactivated in SCCOHT cases, resulted from bi-allelic alterations or silencing epigenetic modification.

WES was carried out using FFPE samples, because fresh SCCOHT biopsies were not readily available. Recent developments in library preparation have achieved comparable sequencing depth for FFPE samples, allowing reliable variant calling similar to data from frozen samples (Table 2.4).

Table 2.4: Coverage of samples sequenced by WES.

Reproduced from Witkowski & Carrot-Zhang et al., *Nature Genetics*, 2014.

a) blood samples sequenced by WES.

Sample	Mean coverage	%CCDS bases >=5x coverage	%CCDS bases >=10x coverage	%CCDS bases >=20x coverage	%CCDS bases >=50x coverage
FA1(f)	118.04	96.5	95	90.9	73.3
FA1b	82.85	95.8	93.3	86.5	61.3
FA2a	123.34	96.5	95.2	91.7	75.6
FA2b	106.57	96.2	94.5	89.8	70
FA3(f)	134.01	96.6	95.5	92.6	79.1
FA3b	129.84	96.7	95.6	92.7	78.9

b) tumor samples sequenced by FFPE-WES.

Sample	Mean coverage	%CCDS bases >=5x coverage	%CCDS bases >=10x coverage	%CCDS bases >=20x coverage	%CCDS bases >=50x coverage
FA1b (T)	57	97	95	90	55
NF1	24	92	78	48	12
NF2	147	98	97	96	87
UN3	44	95	90	79	36
UN4	88	98	96	93	76
UN5	68	97	95	89	60
UN6	78	97	95	90	70
UN8	11	81	54	15	0
UN9	51	96	93	83	43
UN14	43	96	89	71	32
UN15	3	29	6	0	0
UN16	76	97	96	91	63

CCDS = Coverage of consensus coding sequence; FA1(f) and FA3(f) = fathers from family 1 and 3.

Although DNA from FFPE samples showed greater fragmentation and coverage variability, LOH was identified in all but one case, which only one *SMARCA4* mutation was found (Figure 2.5, Figure 2.6). The only case (UN9), which had a mono-allelic mutation (c.1420+1G>T) in *SMARCA4* but no LOH on chromosome 19p, showed weak staining of SMARCA4 from IHC (Figure 2.6).

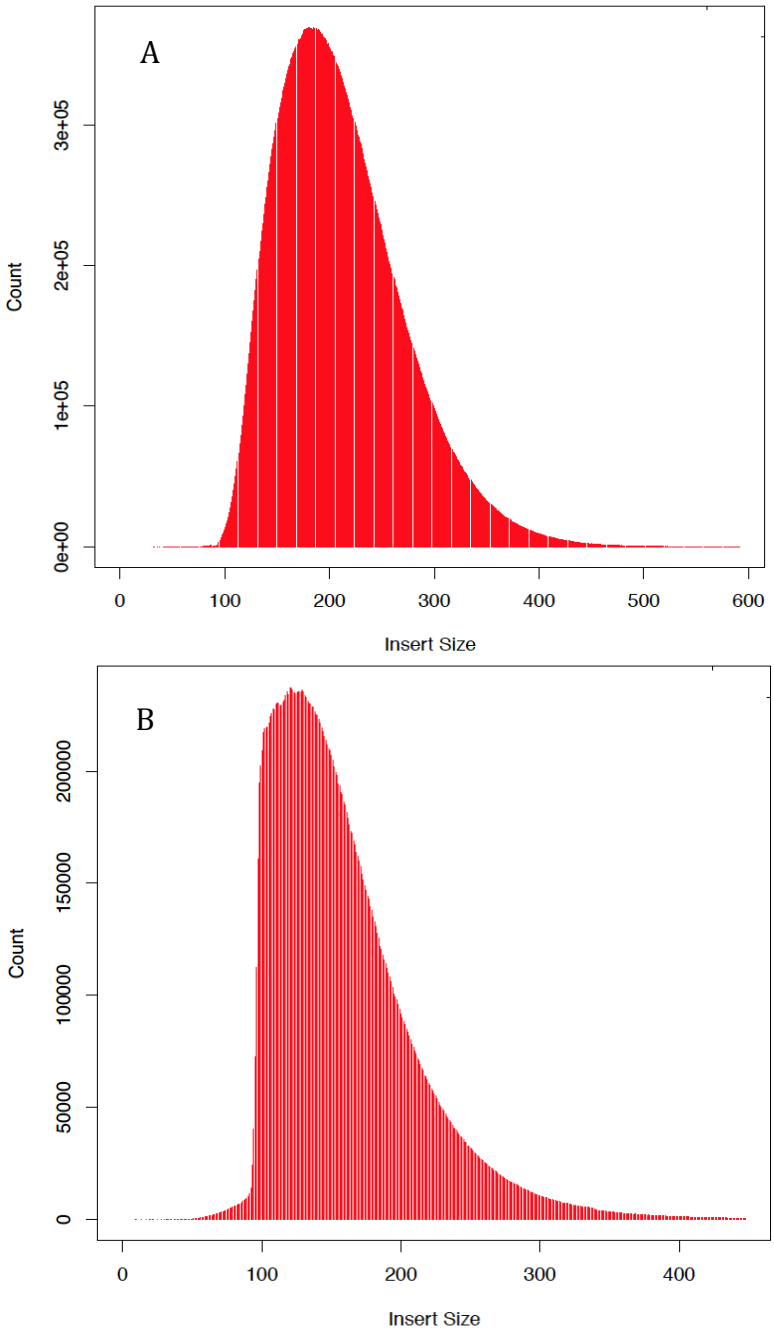


Figure 2.5: The distribution of library insert sizes for sequencing reads.

A) germ-line sample and B) FFPE sample. The distribution of insert size is determined from the distance between properly mapped forward and reverse read pairs. Insert sizes in FFPE samples tend to be smaller than high-quality samples, due to formalin-induced DNA fragmentation.

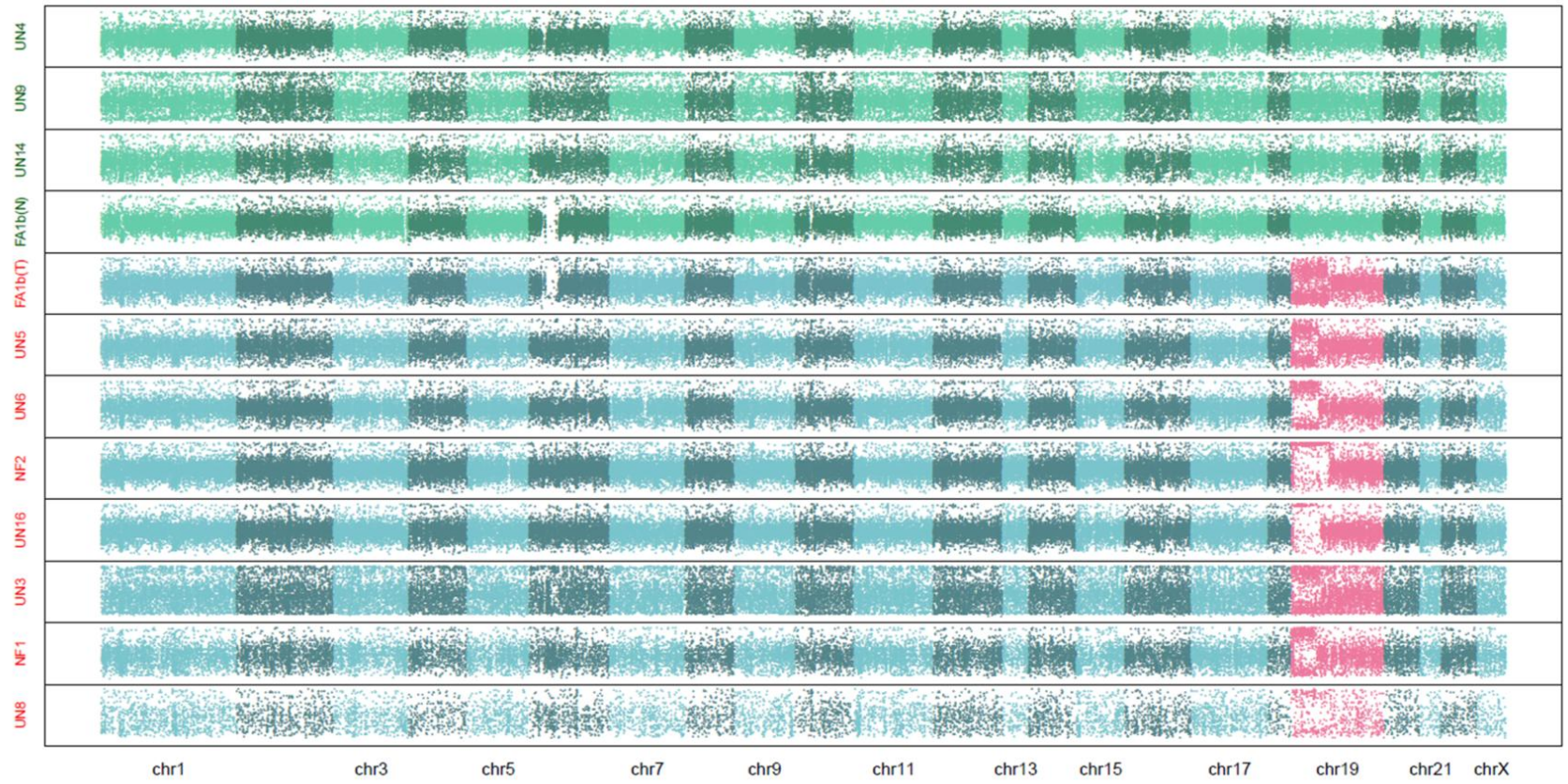


Figure 2.6: Genome-wide AI analysis on SCCOHT samples.

Region with LOH is known in pink. The top 3 rows show tumors with no LOH. Rows 4 and 5 show the matched normal and tumor sample, where row 4 is matched normal DNA and row 5 is tumor DNA with LOH on chr19p. chr19p is the region where *SMARCA4* is located. Reproduced from Witkowski & Carrot-Zhang et al., *Nature Genetics*, 2014.

Additionally, my colleagues performed Sanger sequencing and immunohistochemistry (IHC), targeting *SMARCA4* or SMARCA4 protein in more SCCOHT cases. In total, 23 of 26 cases showed at least one germ-line or somatic mutation in *SMARCA4*, and 38 of 43 tumors showed loss of SMARCA4 expression (Table 2.3, Figure 2.6). Of the five remaining cases, three cases were revised to non-SCCOHT by two reference pathologists. A fourth family with SCCOHT was also analyzed using Sanger sequencing. A missense mutation, c.3239G>A (encoding p.Gly1080Asp), was found in the affected mother and daughter. The mother's tumor showed LOH and retained SMARCA4 staining, whereas the daughter's tumor had a c.1326delC mutation (encoding p.Ser442Argfs*59) and showed loss of SMARCA4 staining.

2.3.3 Mutational Profiles of SCCOHTs

When we published our first paper with 12 SCCOHT tumors undergoing WES, most of the matched normal tissues were not available to identify germ-line variants. In an attempt to reveal all the possible, cancer-driving genes in SCCOHT tumors, a combined list of all the variants identified in the tumors were intersected with a list of more than 600 cancer-related genes obtained through FoundationOne (http://www.foundationone.com/genelist1.php), and previous studies in a pediatric central nervous system tumor, namely atypical teratoid/rhabdoid tumor (ATRT), as well as ovarian serous cystadenocarcinoma (top 100 mutated genes indicated obtained from the cBioPortal) (The Cancer

Genome Atlas Research Network 2011, Lee et al. 2012, Cerami et al. 2012) (Table 2.5). Genes with variants found in ATRTs were included because those tumors were universally characterized by deleterious *SMARCB1* mutations (Lee et al. 2012). Germ-line mutations in *SMARCA4* have also been reported in two ATRT cases (Schneppenheim et al. 2010, Hasselblatt et al. 2011). We therefore, suspected some genetic similarities between SCCOHT and ATRT.

Table 2.5: List of genes used for mutational profiling.

Colors correlate to list from which gene was chosen. Reproduced from Witkowski & Carrot-Zhang et al., *Nature Genetics*, 2014.

ABHD2	BPIFB3	CNTNAP2	FANCL	IFNA1	MAP2K4	NXPH2	PANX3	RNF43	TNRC6A
ABL1	BRAF	COL16A1	FAT2	IFNA13	MAP3K1	ODAM	PAX5	RPTOR	TOP1
ACADL	BRCA1	COL19A1	FBXW7	IGF1R	MAPK10	ODF1	PBRM1	RUNX1	TP53
ACSM2B	BRCA2	COL27A1	FCRLB	IGHMBP2	MAS1L	ODZ4	PCDH17	SALL2	TP63
ADD3	BRD7	COLEC11	FGF10	IKBKE	MBD5	OGDHL	PCDH9	SCN7A	TP73
ADH1B	BRIP1	CP	FGF14	IKZF1	MCART1	OLFM3	PCMTD1	SEC14L4	TRABD
AFF3	BTK	CREBBP	FGF19	IL36A	MCL1	OR10A7	PDE1C	SEC16A	TRAK1
AHCYL2	BTN1A1	CRKL	FGF23	IL7R	MDM2	OR10G7	PDGFRA	SEPT2	TRHDE
AKAP3	BTN2A1	CRLF2	FGF3	INHBA	MDM4	OR10H1	PDGFRB	SERPINF1	TRIO
AKD1	C10ORF113	CSF1R	FGF4	IRF4	MED12	OR10H5	PDK1	SETD2	TRPC4AP
AKT1	C10orf137	CSH1	FGF6	IRS2	MED25	OR10J3	PHF2	SF3B1	TRPM1
AKT2	C10orf140	CSN3	FGFR1	IRX1	MEF2B	OR10T2	PIK3CA	SFXN3	TSC1
AKT3	C11orf42	CT47B1	FGFR2	IRX3	MEN1	OR10X1	PIK3CG	SH3BP4	TSC2
ALK	C11orf63	CTCF	FGFR3	ITGA2	MET	OR11G2	PIK3R1	SHB	TSHR
AMPD2	C14orf101	CTNNA1	FGFR4	ITGA7	METTL2A	OR1L8	PIK3R2	SHPRH	TSNAXIP1
ANKRD17	C16orf62	CTNNB1	FLT1	ITGAV	MFSD6	OR2A12	PKHD1	SLC37A2	TUBB4Q
ANKS4B	C16orf89	CXCR3	FLT3	JAK1	MITF	OR2F2	PKLR	SLC4A11	UBQLN3
ANP32C	C18orf1	CYLC1	FLT4	JAK2	MKI67	OR2M2	PLA2G7	SLC8A1	UCK1
AP3B1	C19orf26	CYP4A11	FMN2	JAK3	MLH1	OR2T34	PLA2R1	SMAD2	UGT2B10
APAF1	C21orf29	CYP4F22	FOXH1	JHDM1D	MLL	OR2V2	PLAT	SMAD4	UGT2B4
APBA2	C2orf78	DAD1	FOXK2	JMJD7-PLA2G4B	MLL3	OR2W1	PLCB2	SMARCA4	USHBP1
APC	C5orf40	DAXX	FOXL2	JUN	MPL	OR4C13	PLEKHM3	SMARCB1	VENTX
APOL5	C9ORF171	DDR2	GABRA6	JUP	MPO	OR4C15	PLIN4	SMARCC2	VHL

AR	CABS1	DEFB118	GABRB1	KAT6A	MRE11A	OR4C46	PLUNC	SMC6	VMO1
ARAF	CACNA1B	DHDH	GABRB2	KAT6B	MRPL53	OR4F21	PLXNA2	SMO	VN1R5
ARFRP1	CACNA1C	DHTKD1	GABRG1	KCNB1	MSH2	OR4F5	PMS2	SNX33	VPS26A
ARGFX	CAMSAP1	DNAH8	GADD45GIP1	KCND1	MSH6	OR4K1	POLR3D	SOCS1	VSIG2
ARHGEF7	CARD11	DNAH9	GAS2L2	KDM5A	MSL3	OR4K13	POMC	SOX10	WDR6
ARID1A	CASP4	DNMT3A	GATA1	KDM5C	MTOR	OR4K2	POTEC	SOX2	WDR66
ARID2	CASP5	DOPEY2	GATA2	KDM6A	MUTYH	OR4L1	POTED	SPANXN3	WISP3
ARMC10	CATSPER1	DOT1L	GATA3	KDMB1	MVP	OR4S2	PPBP	SPEN	WNT
ARMCX5	CBFB	DPP10	GDA	KDR	MYC	OR51A4	PPP2R1A	SPOP	WT1
ASB11	CBL	EGF	GID4	KEAP1	MYCBP2	OR51G1	PPP2R5E	SPTA1	XPO1
ASB12	CBLC	EGFR	GNA11	KHSRP	MYCL1	OR52L1	PRDM1	SRC	ZBTB26
ASB2	CCND1	EIF3IP1	GNA13	KIAA0284	MYCN	OR52N5	PREX1	SSBP2	ZC3H8
ASB8	CCND2	ELOVL4	GNAQ	KIAA1486	MYD88	OR56A1	PRKAR1A	SSBP3	ZFP30
ASMTL	CCND3	EMSY	GNAS	KIAA2026	MYF5	OR5B17	PRKDC	ST3GAL3	ZG16B
ASXL1	CCNE1	EP300	GOLGA5	KIF2B	MYH7	OR5B2	PROKR2	STAG2	ZMYM3
ATM	CCT7	EPHA3	GPHB5	KIT	MYO6	OR5B21	PRRC2A	STAG3	ZMYND19
ATP2B2	CD79A	EPHA5	GPR101	KLHL6	MYO9B	OR5D16	PRRG2	STARD13	ZNF217
ATP8B3	CD79B	EPHB1	GPR119	KRAS	MYST4	OR5F1	PSORS1C2	STAT4	ZNF28
ATR	CDC73	EPS8L2	GPR124	KRT222	MYT1	OR5H2	PTCH1	STK11	ZNF295
ATRX	CDH1	ERBB2	GPR149	KRT72	NCKAP5	OR5L2	PTEN	STRBP	ZNF326
AURKA	CDK12	ERBB3	GPR65	KRTAP10-1	NCOA7	OR5M10	PTK2B	SUFU	ZNF429
AURKB	CDK4	ERBB4	GRIA3	KRTAP19-4	NF1	OR5M11	PTPN11	TAAR6	ZNF433
AVP	CDK6	ERG	GRIN2A	KRTAP27-1	NF2	OR5P2	PTPN7	TAS2R1	ZNF441
AXL	CDK8	ESR1	GSK3B	LAML	NFATC1	OR5R1	PYROXD1	TAS2R42	ZNF479
B3GALT4	CDKN1B	ESRP1	GUK1	LDB3	NFE2L2	OR6A2	RAD50	TAS2R50	ZNF546
B3GNT3	CDKN2A	ESRP2	GZMA	LEPREL1	NFKBIA	OR6B3	RAD51	TBC1D21	ZNF658
B4GALNT3	CDKN2B	EXTL3	HECW1	LIMCH1	NID2	OR6C75	RAF1	TBP	ZNF681
BAF200	CDKN2C	EZH2	HES1	LIMK1	NKX2-1	OR6C76	RARA	TBX1	ZNF70
BAP1	CEBPA	FABP3	HGF	LNX1	NLRP13	OR7C1	RASGRF2	TCF20	ZNF703

BARD1	CELSR2	FAM123B	HIBCH	LRP1B	NOTCH2	OR7D4	RASSF10	TET2	ZNF786
BAT2	CGREF1	FAM167B	HIST1H1A	LRRC15	NPM1	OR8B4	RB1	TFAP2D	ZNF804B
BAf180	CHEK1	FAM24A	HIST1H1C	LRRIQ3	NPPB	OR8D1	RBBP4	TGFBR2	ZNF813
BCL11B	CHEK2	FAM46C	HIST1H1T	LTK	NRAS	OR8H3	RBMXL1	THEG	
BCL2	CHIA	FAM47B	HIST1H2AM	LYPLA2P1	NRG2	OR8J1	RC3H1	TIGD5	
BCL2L2	CHRNA4	FAM75A6	HIST1H2BH	MACC1	NTNG1	OR9A2	RCAN3	TINAG	
BCL6	CHRND	FANCA	HIST1H3B	MAF	NTNG2	OR9G4	REG4	TJP2	
BCOR	CIC	FANCC	HIST1H4H	MAGI2	NTRK1	OVCH1	RET	TMED7-TICAM2	
BCORL1	CLDN17	FANCD2	HORMAD1	MAN1A2	NTRK2	P2RY10	RHBDF1	TMEM126A	
BLM	CLEC4F	FANCE	HRAS	MAP1A	NTRK3	PAGE2	RICTOR	TNF	
BPIFA1	CLSPN	FANCF	IDH1	MAP2K1	NUP93	PAK3	RNA-PolII	TNFAIP3	
BPIFB2	CNKSR3	FANCG	IDH2	MAP2K2	NVL	PALB2	RNF31	TNFRSF14	

Dark Blue = FoundationOne; Green = ATRT; Orange = TCGA; Purple = FoundationOne and TCGA; Light Blue = FoundationOne and ATRT; Pink = ATRT and TCGA; Red = FoundationOne, ATRT and TCGA.

Overall, a total of 35 genes overlapped, of which *SMARCA4* was the most recurrently mutated gene (Table 2.6). The other gene mutated in more than two samples is the proto-oncogene *RET*. Of note, mutations in this gene have been previously seen in our dataset of non-cancer samples. A splicing variant in *MYH7*, which is a member of the myosin superfamily of genes with mutations causing myopathies, was shared by two samples. Again, this gene has been mutated in our control dataset. Three known cancer genes have been observed, including *NF2*, *RAD50* and *IGF2*, which were mutated in individual samples. However, mutations in those three genes were only present in samples with *SMARCA4* mutations, suggesting that they may not be the major, cancer-driving events.

Table 2.6: Mutation Profile comparison results.

Reproduced from Witkowski & Carrot-Zhang et al., *Nature Genetics*, 2014.

Position	Gene	cDNA or Protein Change	Variation Type	FA1b(T)	NF1*	NF2	UN3	UN4	UN5	UN6	UN9	UN14	UN16	MAF 1000genomes	MAF EVS
chr10:124671222	FAM24A	p.M24I	Missense	het	-	-	-	-	-	-	-	-	-	0	0
chr10:129903826	MKI67	p.T1733I	Missense	-	het	-	-	-	-	-	-	-	-	0	0
chr10:43601928	RET	p.W324C	Missense	-	-	-	-	-	-	het	-	-	-	0	0
chr10:43615063	RET	p.Y826S	Missense	-	-	-	-	-	het	-	-	-	-	0	0
chr10:88441401	LDB3	p.A177V	Missense	-	-	-	het	-	-	-	-	-	-	0	0
chr11:123909467	OR10G7	p.M81K	Missense	-	-	-	-	-	-	-	het	-	-	0	0
chr11:65789055	CATSPER1	p.A535T	Missense	-	-	-	-	-	-	het	-	-	-	0	0
chr11:723327	EPS8L2	p.P477fs	Frameshift insertion	-	-	-	-	-	-	-	-	het	-	0	0
chr14:23900692	MYH7	T/C	Splicing	-	-	-	het	het	-	-	-	-	-	0	0
chr15:40591094	PLCB2	p.S252F	Missense	-	-	-	-	-	-	het	-	-	-	0	0
chr15:43814359	MAP1A	p.V230M	Missense	-	-	-	-	-	-	-	-	-	het	0	0
chr15:99250917	IGF1R	p.R74L	Missense	-	-	-	het	-	-	-	-	-	-	0	0
chr16:3843600	CREBBP	p.V335I	Missense	-	-	-	-	-	-	-	het	-	-	0	0
chr16:67860422	TSNAXIP1	p.N399S	Missense	-	-	-	het	-	-	-	-	-	-	0	0
chr19:11095960	SMARCA4	p.79_79del	nonframeshift deletion	-	-	-	-	-	-	-	-	-	hom	0	0
chr19:11102000	SMARCA4	G/T	Splicing	-	-	-	-	-	-	-	het	-	-	0	0
chr19:11106958	SMARCA4	p.Q555X	stopgain SNV	-	-	hom	-	-	-	-	-	-	-	0	0
chr19:11121123	SMARCA4	p.S730fs	frameshift insertion	-	-	-	-	-	-	-	-	-	-	0	0
chr19:11123624	SMARCA4	G/T	Splicing	-	-	-	-	-	het	-	-	-	-	0	0
chr19:11132619	SMARCA4	p.P946fs	frameshift deletion	-	-	-	-	-	-	het	-	-	-	0	0
chr19:11134249	SMARCA4	p.L972P	Missense	-	-	-	-	-	-	-	-	het	-	0	0
chr19:11134266	SMARCA4	p.R978X	stopgain SNV	-	-	-	het	-	-	-	-	-	-	0	0

chr19:11136185	SMARCA4	G/C	Splicing	-	-	-	-	-	-	-	-	het	-	0	0
chr19:11141554	SMARCA4	p.1178_1178del	frameshift deletion	-	-	-	-	het	-	-	-	-	-	0	0
chr19:11145809	SMARCA4	G/A	Splicing	het	-	-	-	-	-	-	-	-	-	0	0
chr19:11170479	SMARCA4	p.I1534fs	frameshift deletion	-	-	-	-	het	-	-	-	-	-	0	0
chr19:12126957	ZNF433	p.R242Q	Missense	-	-	-	-	-	-	-	het	-	-	0	0
chr19:14910161	OR7C1	p.A263E	Missense	-	-	-	-	-	-	-	het	-	-	0	0
chr19:42795291	CIC	p.T791P	Missense	-	-	-	-	-	-	-	het	-	-	0	0
chr1:158648254	SPTA1	p.R250H	Missense	het	-	-	-	-	-	-	-	-	-	0	0
chr20:3218583	SLC4A11	p.A43V	Missense	-	het	-	-	-	-	-	-	-	-	0	0
chr20:39746921	TOP1	p.M645I	Missense	-	-	-	-	het	-	-	-	-	-	0	0
chr20:57429476	GNAS	p.Q323H	Missense	-	-	-	-	-	-	-	het	-	-	0	0
chr20:57429479	GNAS	p.S387A	Missense	-	-	-	-	-	-	-	het	-	-	0	0
chr20:57429484	GNAS	p.R326P	Missense	-	-	-	-	-	-	-	het	-	-	0	0
chr20:57429488	GNAS	p.T390P	Missense	-	-	-	-	-	-	-	het	-	-	0	0
chr20:57429497	GNAS	p.R393G	Missense	-	-	-	-	-	-	-	het	-	-	0	0
chr20:57429539	GNAS	p.Q344H	Missense	-	-	-	-	-	-	-	het	-	-	0	0
chr22:30057203	NF2	p.G146C	Missense	-	-	het	-	-	-	-	-	-	-	0	0
chr2:149247048	MBD5	p.P1050A	Missense	-	-	-	-	-	-	het	-	-	-	0	0
chr2:74044083	C2orf78	p.D911E	Missense	-	-	het	-	-	-	-	-	-	-	0	0
chr4:74124265	ANKRD17	p.G41C	Missense	-	-	-	het	-	-	-	-	-	-	0	0
chr5:131953924	RAD50	p.I1109M	Missense	-	-	-	-	het	-	-	-	-	-	0	0
chr5:54398524	GZMA	p.Y5C	Missense	-	-	-	-	-	-	-	-	het	-	0	0
chr6:41903784	CCND3	p.A186V	Missense	-	-	-	-	-	-	het	-	-	-	0	0
chr6:46675732	PLA2G7	p.I346V	Missense	-	-	-	-	het	-	-	-	-	-	0	0
chr6:51890718	PKHD1	p.A1297V	Missense	-	-	-	-	-	-	-	het	-	-	0	0
chr6:76600958	MYO6	p.E961K	Missense	-	-	-	-	-	-	het	-	-	-	0	0
chr7:57188114	ZNF479	p.W336X	stopgain SNV	-	-	-	het	-	-	-	-	-	-	0	0
chr7:57188128	ZNF479	p.K332Q	Missense	-	-	-	het	-	-	-	-	-	-	0	0
chr8:37698761	GPR124	p.A969S	Missense	-	-	-	-	-	-	-	-	-	het	0	0
chr9:138709859	CAMSAP1	p.P1412L	Missense	-	-	het	-	-	-	-	-	-	-	0	0
chrX:129149954	BCORL1	p.D1069A	Missense	-	-	-	-	het	-	-	-	-	-	0	0

Red cells highlight mutations found in *SMARCA4*. Hom = Homozygous; Het = Heterozygous; MAF = Minor allele frequency

2.3.4 Genomic Analysis of Paired tumor and Normal Samples Revealed Recurrent, Somatic Aberration in SCCOHT

To characterize the landscape of somatic alterations in SCCOHT, I then performed WES analysis on 14 SCCOHT tumors and their matched normal tissues. Apart from *SMARCA4*, no other genes showed recurrent, somatic mutations in those cases. Mutation rate analysis revealed fewer mutations per megabase (Mb) in the coding regions in SCCOHT than in high-grade serous ovarian cancers (HGSCs) (Figure 2.7). Of note, the splicing variant of *MYH7* and one of the two *RET* variants discussed above were present in the germ-line samples of the carriers. The other *RET* variant was somatic. The only recurrent AI aberration identified in SCCOHT was on chr19p, surrounding *SMARCA4*. In contrast, HGSC, which is known for its aneuploidy, showed multiple recurrent AI aberrations across the genome. Taken together, our genomic analysis suggests that SCCOHTs have a remarkably simple genome, as the only, recurrent genomic abnormality is related to *SMARCA4*.

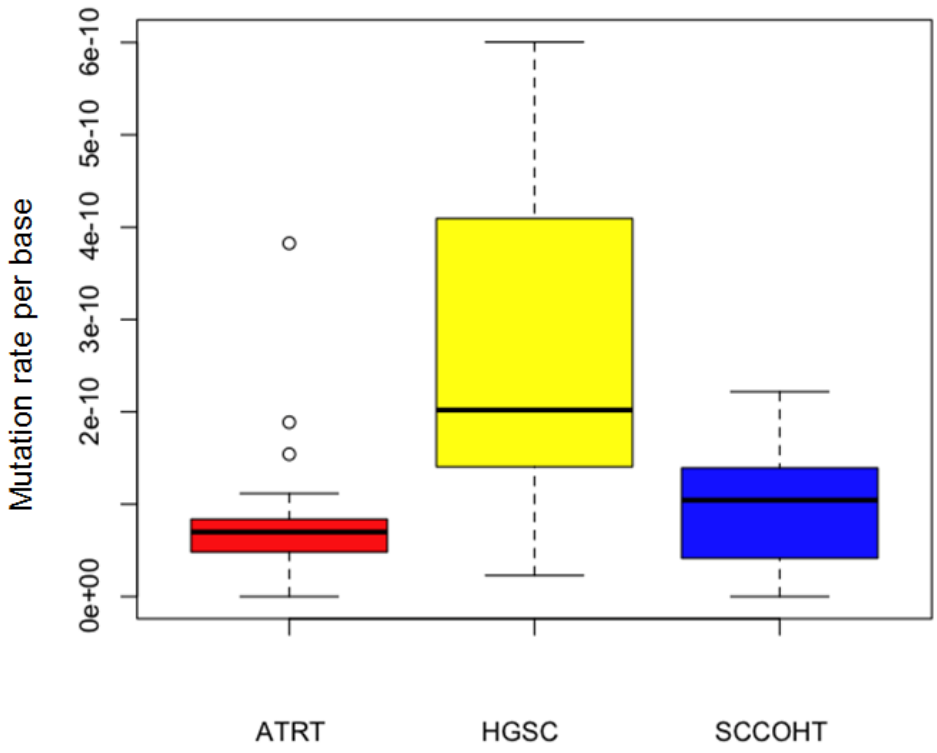


Figure 2.7: Mutation rate of SCCOHT, ATRT, and HGSC.

Mutation rate is calculated as the ratio of the number of somatic mutations and the total number of sites sequenced in the sample. Only sites with total coverage larger than 20X are used for the calculation. Besides the SCCOHT (blue) samples sequenced in this study (N=12), ATRT (red) samples obtained from our collaborators (N=20) and publically available HGSC (yellow) samples from The Cancer Genome Atlas (TCGA) consortium (N=16) were used.

2.4 Discussion

SMARCA4 encodes an ATPase subunit in the SWI/SNF chromatin-remodeling complex (Witkowski, Foulkes 2015). Deleterious mutations in *SMARCA4* and differences in SMARCA4 expression have been found in many cancer cell lines and primary cancers (Wong et al. 2000, Roberts, Orkin 2004, Sun et al. 2007, Kadoch et al. 2013). Other genes in the SWI/SNF chromatin-remodeling complex have been implicated in cancers, such as *SMARCB1* in schwannomatosis and *SMARCE1* in spinal meningiomas (Lee et al. 2012, Smith et al. 2012, Smith et al. 2013). Germ-line mutations in *SMARCA4* are also known to rarely predispose ATRTs (Schneppenheim et al. 2010, Hasselblatt et al. 2011, Witkowski et al. 2013). Similarly, rhabdoid tumor predisposition syndrome is largely associated with mutations in *SMARCB1* (Brennan, Stiller & Bourdeaut 2013).

A previous study from Poland also identified *SMARCA4* mutations in two cases of SCCOHT (Kupryjańczyk et al. 2013). The authors also implied a resemblance of SCCOHT and rhabdoid tumors, which is in line with the pathological re-examination of our study (Witkowski et al. 2014). In this study, WES analysis of a larger number of SCCOHT cases confirmed that *SMARCA4* is the only recurrently mutated gene in SCCOHT, and that SCCOHTs, like ATRTs, have a remarkably simple genome and harbor fewer somatic mutations and

chromosomal-wide alterations than other cancers, aside from the deficiency in the SWI/SNF chromatin-remodeling complex (Lee et al. 2012).

The presence of *SMARCA4* mutations in all the familial cases studied, the high-incidence of *SMARCA4* alterations in SCCOHT and the lack of other recurrent mutations observed in the tumors showed that SCCOHT is largely a monogenic disorder. Therefore, genetic testing in *SMARCA4* mutations followed by salpingo-oophorectomy may aid the prevention of this disease (Berchuck et al. 2015). In a recent study, a synthetic lethality approach was developed to treat SMARCA4-deficient non-small-cell lung carcinomas (Oike et al. 2013). However, further work is required to investigate the mechanism of tumorigenesis related to *SMARCA4* mutations in SCCOHT, and therefore to ultimately improve the treatment of SCCOHT.

As discussed in Chapter 1, SCCOHTs are difficult for pathologists to diagnose, because they are morphologically overlapped with many other ovarian tumors (Clement 2005). Although missense mutations in *SMARCA4* have been reported in other types of ovarian tumors, such as HGSCs and clear cell ovarian cancers via The Catalogue of Somatic Mutation in Cancer (COSMIC) database, thus far, nonsense mutations have only been seen in SCCOHTs among all types of ovarian cancers. Moreover, similar to the absence of SMARCB1 staining, which was observed in the vast majority of rhabdoid tumors, nearly all the SCCOHTs in this study showed loss of SMARCA4 staining. Therefore, our findings suggested

that SMARCA4 can be a considered as a specific, immunohistochemistry marker for SCCOHT diagnosis – a lack of SMARCA4 staining in an ovarian cancer should distinguish SCCOHT from its morphological mimics.

Chapter 3. WES study of familial breast cancer cases based on the French-Canadian founder population identifies *RECQL* as a new breast cancer susceptibility gene

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

"Germline *RECQL* mutations are associated with breast cancer susceptibility" Cezary Cybulski, Jian Carrot-Zhang, Wojciech Kluźniak, Barbara Rivera, Aniruddh Kashyap, Dominika Wokołorczyk, Sylvie Giroux, Javad Nadaf, Nancy Hamel, Shiyu Zhang, Tomasz Huzarski, Jacek Gronwald, Tomasz Byrski, Marek Szwiec, Anna Jakubowska, Helena Rudnicka, Marcin Lener, Bartłomiej Masojć, Patrica N Tonin, Francois Rousseau, Bohdan Górski, Tadeusz Debniak, Jacek Majewski, Jan Lubinski, William D Foulkes, Steven A Narod & Mohammad R Akbari. *Nature Genetics*, 47, 643-646 (2015)

Permission was granted to reproduce figures and tables in Chapter 3. Author contributions are stated in the Contribution of The Authors section.

3.1 Introduction

Breast cancer is the most common and second leading cause of death in female cancers. Familial breast cancer accounts for approximately 10% of all breast cancer cases (Foulkes 2008). However, *BRCA1* and *BRCA2* together explain only 15% of all familial cases. Other genes with pathogenic mutations conferring moderate to high risk of breast cancer, such as *PALB2*, *ATM*, *CHEK2*, *BRIP1* and *RAD51C*, contribute to another 7% at most (Couch, Nathanson & Offit 2014). Therefore, approximately 60% familial cancer remains unexplained. It has become clear that the remainder of the "missing heritability" for breast cancer cannot be attributed to a small number of undiscovered moderate-to high-penetrance breast cancer susceptibility genes. By targeting the entire coding regions in the genome, WES will likely reveal novel, moderate to high penetrant variants, but those variants will be exceptionally rare. The situation is quite different in founder populations, where single alleles can contribute substantially to the burden of disease attributable to all mutations in a breast cancer susceptibility gene. By focusing on the French-Canadian population, novel genes with breast cancer associated alleles will likely be identified, using a limited number of breast cancer families.

3.2 Materials and Methods

WES data were generated as described in section 2.2.1 and analyzed using the pipeline as described in section 2.2.2. In brief, BWA version 0.5.9 was used for short read alignment (Li, Durbin 2009). SNVs and indels were called by SAMtools mpileup version 0.1.17 (Li 2011). Mutation class, and minor allele frequency in public databases were annotated by ANNOVAR (Wang, Li & Hakonarson 2010). Custom scripts were used to combine and identify shared variants among families. XHMM was used to identify rare CNVs from germ-line samples (Fromer, Purcell 2014). Finally, ExomeAI was used to call somatic mutations using tumor and its matched normal tissue (Nadaf, Majewski & Fahiminiya 2015).

3.3 Results

3.3.1 Filtration and Prioritization of Raw Variants

In order to identify novel breast cancer susceptibility genes, after receiving WES results from 51 French-Canadian, unrelated breast cancer patients with family history, I first confirmed that no case carried mutations in known breast cancer genes, including *BRCA1*, *BRCA2*, *PALB2* and *TP53* (Figure 3.1). Then, variants from all cases were combined and filtered using the following steps: 1) variants with minor allele frequencies greater than 0.001 from the 1000 genome and NHLBI database were filtered; 2) synonymous, non-frameshift, and UTR variants were not considered; 3) variants seen previously in our in-house control dataset (>1000 non-cancer samples) more than five times were filtered. I excluded those variants assuming them to be benign or low-penetrance variants. Finally, 2659 rare variants passed filters.

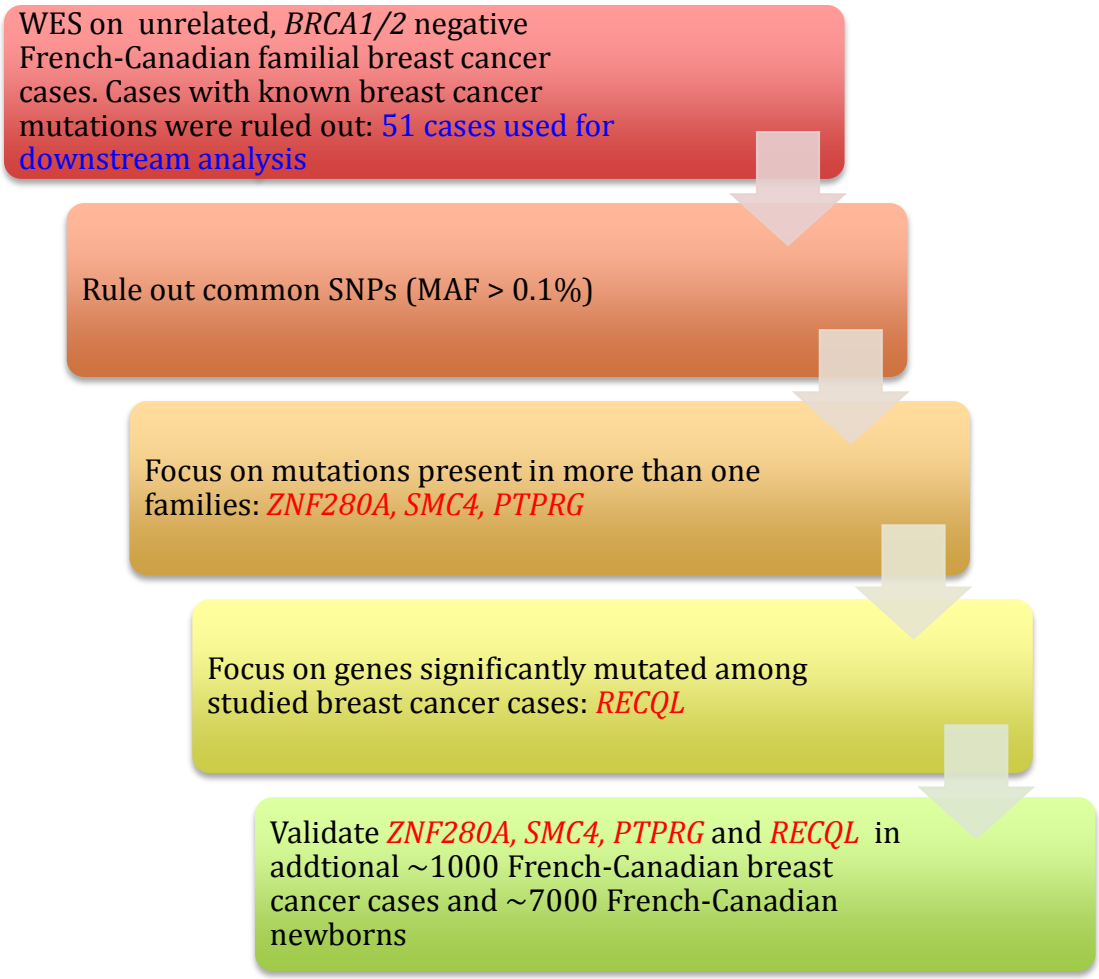


Figure 3.1: Chronology of identifying new breast cancer genes.

The *ZNF280A* variant was included for validation because it was the only variant present in more than two samples from 51 cases, even though this variant had a higher minor allele frequency (MAF >0.1%) in EVS (Exome Variant Server of more than 6000 individuals studied by the NHLBI).

I then prioritized those variants using the following steps: 1) variant recurrently observed in more than one family was considered as potential French-Canadian founder mutations related to breast cancer, and was therefore prioritized. A total

of 330 variants fell into this category. 2) Gene with different variants observed in more than one family, which might be a candidate breast cancer susceptibility gene, but unlikely to represent founder mutations in the French-Canadian population. I found that 953 genes were recurrently mutated in the 51 cases.

3.3.2 Identifying Candidate Breast Cancer Susceptibility Genes

3.3.2.1 Identification of French-Canadian founder mutation in SMC4

I first set out to identify two families with the same protein-truncating variant (nonsense mutation, essential splice-site mutation, frameshift insertion or deletion) in a previously unrecognized gene with biological relevance to cancer (Figure 3.1). No candidates fell into this class. Of note, a frameshift deletion, c.1362delC encoding p.S454fs in *ZNF280A* was found in three cases, whereas 10 carriers were found in 6,259 individuals reported from the NHLBI exome database. Although this database is not specifically from French-Canadian individuals, it was sufficiently large to warrant further investigation of our candidates, and is often used as a reference source of comparison in WES projects containing defined phenotypes (Awadalla et al. 2014, Behlouli et al. 2014, Santen et al. 2013).

The next step was to look for instances where more than one family possessed the same non-protein truncating, but likely deleterious novel variants in the same

gene. As a result of this search, four variants were selected in candidate genes including *PTPRG* (receptor protein tyrosine phosphatase), and *SMC4* (structural maintenance of chromosomes 4), based on their protein functions that were potentially relevant to cancer (Table 3.1). All variants were seen twice in the case cohort, and were predicted to be protein-damaging variants (polyphen2 score larger than 0.95).

Table 3.1: Candidate mutations present in two or more 51 French-Canadian breast cancer families subjected to WES.

Gene name	Frequency	Protein Change	Polyphen2 Score	Protein Function relevant to BC
AATK	2/51	p.P94R	0.992	Tyrosine kinase involved in apoptosis
PTPRG	2/51	p.L1039P	1	Receptor for a Protein Tyrosine Phosphatase
PRKCD	2/51	p.V30M	0.999	Threonine kinase involved in growth, apoptosis & differentiation regulation, cancer metabolism
SMC4	2/51	p.R827C	0.997	Condensin. Implicated in structural maintenance of chromosomes

Moreover, seven copy number variants were recurrently detected in our cases (Table 3.2). A deletion spanning seven exons in *IKL* (integrin-linked kinase) gene (chr11:6629924-6631266) was observed in two samples.

Table 3.2: Copy number variants detected in more than one samples.

Location	CNV	Frequency	Gene	Description
chr1:144916570-144931708	DUP	2/51	*PDE4DIP*	Phosphodiesterase 4D interacting protein
chr1:22328168-22336350	DEL	4/51	*CELA3A*	Chymotrypsin-like elastase family
chr5:814876-825366	DEL	2/51	*AADAC*	Arylacetamide deacetylase
chr6:31616976-31630487	DUP	2/51	*BAG6*	BCL2-associated athanogene 6
chr10:135340900-135379033	DUP	3/51	*CYP2E1*	Cytochrome P450, family 2, subfamily E, polypeptide 1
chr11:6629924-6631266	DEL	2/51	*ILK*	Integrin-linked kinase, over-expression of this gene is implicated in tumor growth and metastasis
chr13:21721325-21746643	DUP	2/51	*SKA3*	Spindle and kinetochore associated complex subunit 3

DUP=duplication; DEL=deletion.

3.3.2.2 Identification of truncating mutations in RECQL

I then looked for genes where two or more mutations were seen in the same gene. Among 953 genes recurrently mutated in the case cohort, 17 had at least two truncating mutations. To prioritize, I tested those genes for excess mutation burden as compared to our 1,095 in-house non-cancer samples as controls, and ranked them by their *p*-values after Fisher's exact test (Table 3.3). 15 genes showed statistical significance.

Table 3.3: Significantly mutated genes among 51 French-Canadian breast cancer cases.

a) Genes with protein-truncating mutations enriched in 51 French-Canadian breast cancer cases. Only rare, truncating mutations are counted here. Fisher's exact test is used to compare the frequency of variants in the case cohort to the control cohort of 1095 non-cancer, WES samples. The *p*-values are corrected for false discovery rate (FDR) using the Benijamini-Hochberg approach (considering the total of genes with truncating mutations in 2 or more cases).

Gene	Rank	Fisher's p-value	Q value	Number of case mutated	Number of case	Number of control mutated	Number of control
RECQL	1	0.00151	0.0110	3	51	3	1095
USHBP1	2	0.00194	0.0110	2	51	0	1095
OR1D2	3	0.00194	0.0110	2	51	0	1095
AGMO	4	0.00566	0.0120	2	51	1	1095
HK3	5	0.00566	0.0120	2	51	1	1095
LDB3	6	0.00566	0.0120	2	51	1	1095
CNTN4	7	0.00566	0.0120	2	51	1	1095
SMPD1	8	0.00566	0.0120	2	51	1	1095
EPX	9	0.0110	0.0156	2	51	2	1095
SLC38A10	10	0.0110	0.0156	2	51	2	1095
EIF2AK4	11	0.0110	0.0156	2	51	2	1095
C1orf56	12	0.0110	0.0156	2	51	2	1095
KIF4B	13	0.0178	0.0216	2	51	3	1095
MAN2C1	14	0.0178	0.0216	2	51	3	1095
STARD9	15	0.0353	0.0400	2	51	5	1095
MYO15A	16	0.0964	0.102	2	51	10	1095
SSPO	17	0.124	0.125	2	51	25	1095

b) Genes with nonsense and missense mutations enriched in 51 French-Canadian breast cancer cases. Only rare mutations are counted here. Only genes that passed Fisher's exact test $p < 0.05$ cutoff are listed.

Gene	Rank	Fisher's p-value	Q value	No. Case mutated	No. Case	No. Control mutated	No. Control
PRIC285	1	8.32E-05	0.0397	3	51	0	1095
RANBP9	2	0.00132	0.123	4	51	8	1095

AP1G1	3	0.00151	0.123	3	51	3	1095
ODZ4	4	0.00194	0.123	2	51	0	1095
ANKRD5	5	0.00194	0.123	2	51	0	1095
CSDA	6	0.00194	0.123	2	51	0	1095
CPQ	7	0.00194	0.123	2	51	0	1095
TROAP	8	0.00331	0.197	4	51	11	1095
GPR126	9	0.00539	0.238	4	51	13	1095
RPP30	10	0.00566	0.238	2	51	1	1095
RGS18	11	0.00566	0.238	2	51	1	1095
SNRPA	12	0.00566	0.238	2	51	1	1095
CEP44	13	0.00566	0.238	2	51	1	1095
LRRC31	14	0.00577	0.238	3	51	6	1095
SYCE1L	15	0.00577	0.238	3	51	6	1095
SLC23A3	16	0.00799	0.238	3	51	7	1095
DCST1	17	0.00822	0.238	4	51	15	1095
C12orf42	18	0.0107	0.238	3	51	8	1095
STAC2	19	0.0107	0.238	3	51	8	1095
OC90	20	0.0107	0.238	3	51	8	1095
VAMP4	21	0.0110	0.238	2	51	2	1095
CALN1	22	0.0110	0.238	2	51	2	1095
BTK	23	0.0110	0.238	2	51	2	1095
UNC119B	24	0.0110	0.238	2	51	2	1095
ZMYND11	25	0.0110	0.238	2	51	2	1095
BFSP2	26	0.0110	0.238	2	51	2	1095
DUSP1	27	0.0110	0.238	2	51	2	1095
INPP1	28	0.0110	0.238	2	51	2	1095
OR6C76	29	0.0110	0.238	2	51	2	1095
FOCAD	30	0.0119	0.238	4	51	17	1095
KCNAB2	31	0.0138	0.238	3	51	9	1095
NCOA2	32	0.0138	0.238	3	51	9	1095
DAB2	33	0.0138	0.238	3	51	9	1095
ARHGEF37	34	0.0173	0.238	3	51	10	1095
CYP4A11	35	0.0173	0.238	3	51	10	1095
GPRIN3	36	0.0173	0.238	3	51	10	1095
SMC5	37	0.0173	0.238	3	51	10	1095
ETV4	38	0.0173	0.238	3	51	10	1095
CADM4	39	0.0178	0.238	2	51	3	1095
PGBD3	40	0.0178	0.238	2	51	3	1095
LIPT2	41	0.0178	0.238	2	51	3	1095
PTGR2	42	0.0178	0.238	2	51	3	1095
EDN3	43	0.0178	0.238	2	51	3	1095
LRRC8D	44	0.0178	0.238	2	51	3	1095
OR6A2	45	0.0178	0.238	2	51	3	1095
ZNF25	46	0.0178	0.238	2	51	3	1095
OR52D1	47	0.0178	0.238	2	51	3	1095
PRKCH	48	0.0178	0.238	2	51	3	1095
OR10A4	49	0.0178	0.238	2	51	3	1095
BEND2	50	0.0178	0.238	2	51	3	1095
HOGA1	51	0.0178	0.238	2	51	3	1095
OR1D2	52	0.0178	0.238	2	51	3	1095
LRRC28	53	0.0178	0.238	2	51	3	1095
DAPL1	54	0.0178	0.238	2	51	3	1095
EPX	55	0.0214	0.238	3	51	11	1095

DNMT3B	56	0.0214	0.238	3	51	11	1095
FBXO4	57	0.0214	0.238	3	51	11	1095
RECQL	58	0.0214	0.238	3	51	11	1095
ULK2	59	0.0214	0.238	3	51	11	1095
GSG2	60	0.0214	0.238	3	51	11	1095
BRDT	61	0.0214	0.238	3	51	11	1095
TBX10	62	0.0214	0.238	3	51	11	1095
SLC2A9	63	0.0214	0.238	3	51	11	1095
COL23A1	64	0.0220	0.238	4	51	21	1095
ARID1A	65	0.0220	0.238	4	51	21	1095
IFT172	66	0.0252	0.238	4	51	22	1095
PCDHB11	67	0.0259	0.238	3	51	12	1095
LDLR	68	0.0259	0.238	3	51	12	1095
MEI1	69	0.0259	0.238	3	51	12	1095
SAMD9	70	0.0259	0.238	3	51	12	1095
ACOXL	71	0.0259	0.238	3	51	12	1095
CLCA2	72	0.0259	0.238	3	51	12	1095
STK24	73	0.0260	0.238	2	51	4	1095
ZNF121	74	0.0260	0.238	2	51	4	1095
CLEC1A	75	0.0260	0.238	2	51	4	1095
APPL1	76	0.0260	0.238	2	51	4	1095
NOP58	77	0.0260	0.238	2	51	4	1095
HCAR1	78	0.0260	0.238	2	51	4	1095
ISCA2	79	0.0260	0.238	2	51	4	1095
HTR3E	80	0.0260	0.238	2	51	4	1095
CCDC121	81	0.0260	0.238	2	51	4	1095
MAGEA6	82	0.0260	0.238	2	51	4	1095
MAGEA10	83	0.0260	0.238	2	51	4	1095
CRISP1	84	0.0260	0.238	2	51	4	1095
GMPR2	85	0.0260	0.238	2	51	4	1095
SCAI	86	0.0260	0.238	2	51	4	1095
RERG	87	0.0260	0.238	2	51	4	1095
MB21D2	88	0.0260	0.238	2	51	4	1095
ANGPTL1	89	0.0260	0.238	2	51	4	1095
HSPH1	90	0.0260	0.238	2	51	4	1095
DNAJC3	91	0.0260	0.238	2	51	4	1095
CADM2	92	0.0260	0.238	2	51	4	1095
LGI2	93	0.0260	0.238	2	51	4	1095
HES4	94	0.0260	0.238	2	51	4	1095
ADTRP	95	0.0260	0.238	2	51	4	1095
CDYL	96	0.0260	0.238	2	51	4	1095
FIBCD1	97	0.0309	0.249	3	51	13	1095
SLC18A3	98	0.0309	0.249	3	51	13	1095
STRA6	99	0.0309	0.249	3	51	13	1095
PLAA	100	0.0309	0.249	3	51	13	1095
CDK5RAP2	101	0.0323	0.249	4	51	24	1095
COL6A3	102	0.0338	0.249	6	51	50	1095
OR8K5	103	0.0353	0.249	2	51	5	1095
PEX19	104	0.0353	0.249	2	51	5	1095
NKX6-2	105	0.0353	0.249	2	51	5	1095
HOXB2	106	0.0353	0.249	2	51	5	1095
HADHB	107	0.0353	0.249	2	51	5	1095
ANKRD65	108	0.0353	0.249	2	51	5	1095

SLC17A3	109	0.0353	0.249	2	51	5	1095
MARS2	110	0.0353	0.249	2	51	5	1095
MANF	111	0.0353	0.249	2	51	5	1095
KRTAP4-11	112	0.0353	0.249	2	51	5	1095
ARL13B	113	0.0353	0.249	2	51	5	1095
ENDOG	114	0.0353	0.249	2	51	5	1095
THOC1	115	0.0353	0.249	2	51	5	1095
CPA4	116	0.0353	0.249	2	51	5	1095
TSEN2	117	0.0353	0.249	2	51	5	1095
PROL1	118	0.0353	0.249	2	51	5	1095
ACOT1	119	0.0353	0.249	2	51	5	1095
SLC26A8	120	0.0353	0.249	2	51	5	1095
HRASLS5	121	0.0353	0.249	2	51	5	1095
ZNF148	122	0.0353	0.249	2	51	5	1095
PDSS2	123	0.0353	0.249	2	51	5	1095
ATG7	124	0.0353	0.249	2	51	5	1095
FAM73A	125	0.0353	0.249	2	51	5	1095
TAS2R41	126	0.0353	0.249	2	51	5	1095
TGS1	127	0.0353	0.249	2	51	5	1095
GIGYF1	128	0.0362	0.249	4	51	25	1095
STAG3	129	0.0363	0.249	3	51	14	1095
UACA	130	0.0363	0.249	3	51	14	1095
CNOT1	131	0.0363	0.249	3	51	14	1095
CCDC168	132	0.0408	0.250	7	51	67	1095
IL17RD	133	0.0423	0.250	3	51	15	1095
SMCR8	134	0.0423	0.250	3	51	15	1095
ZNF672	135	0.0423	0.250	3	51	15	1095
CLSTN1	136	0.0423	0.250	3	51	15	1095
TNFRSF18	137	0.0423	0.250	3	51	15	1095
DDX1	138	0.0458	0.250	2	51	6	1095
USP16	139	0.0458	0.250	2	51	6	1095
SLC5A4	140	0.0458	0.250	2	51	6	1095
DTNB	141	0.0458	0.250	2	51	6	1095
RAB3GAP1	142	0.0458	0.250	2	51	6	1095
IQCF1	143	0.0458	0.250	2	51	6	1095
HCK	144	0.0458	0.250	2	51	6	1095
MLF1	145	0.0458	0.250	2	51	6	1095
ABHD4	146	0.0458	0.250	2	51	6	1095
S1PR3	147	0.0458	0.250	2	51	6	1095
LUZP4	148	0.0458	0.250	2	51	6	1095
GPR64	149	0.0458	0.250	2	51	6	1095
CHRND	150	0.0458	0.250	2	51	6	1095
AHCYL1	151	0.0458	0.250	2	51	6	1095
EXTL2	152	0.0458	0.250	2	51	6	1095
MCOLN2	153	0.0458	0.250	2	51	6	1095
DBF4	154	0.0458	0.250	2	51	6	1095
CHRNB3	155	0.0458	0.250	2	51	6	1095
CHAC2	156	0.0458	0.250	2	51	6	1095
TDRD12	157	0.0458	0.250	2	51	6	1095
BMP3	158	0.0458	0.250	2	51	6	1095
P2RY2	159	0.0458	0.250	2	51	6	1095
RNFT2	160	0.0458	0.250	2	51	6	1095
NKX2-3	161	0.0458	0.250	2	51	6	1095

HAO2	162	0.0458	0.250	2	51	6	1095
KRT23	163	0.0458	0.250	2	51	6	1095
IPO5	164	0.0458	0.250	2	51	6	1095
NPHS1	165	0.0486	0.250	3	51	16	1095
KAT6B	166	0.0486	0.250	3	51	16	1095
EML2	167	0.0486	0.250	3	51	16	1095
CHD9	168	0.0486	0.250	3	51	16	1095
DEPDC5	169	0.0486	0.250	3	51	16	1095
ABCA12	170	0.0498	0.250	4	51	28	1095
TMEM132C	171	0.0498	0.250	4	51	28	1095

When only rare, protein-truncating mutations were considered, *RECQL*, which was mutated in three families with three different mutations (c.132_135delGAAA, p.Lys45fs; c.426delT, p.Ser142fs; c.1138A>T, p.Lys380*), was ranked as the most significantly mutated gene in the list (Figure 3.2, Table 3.3a).

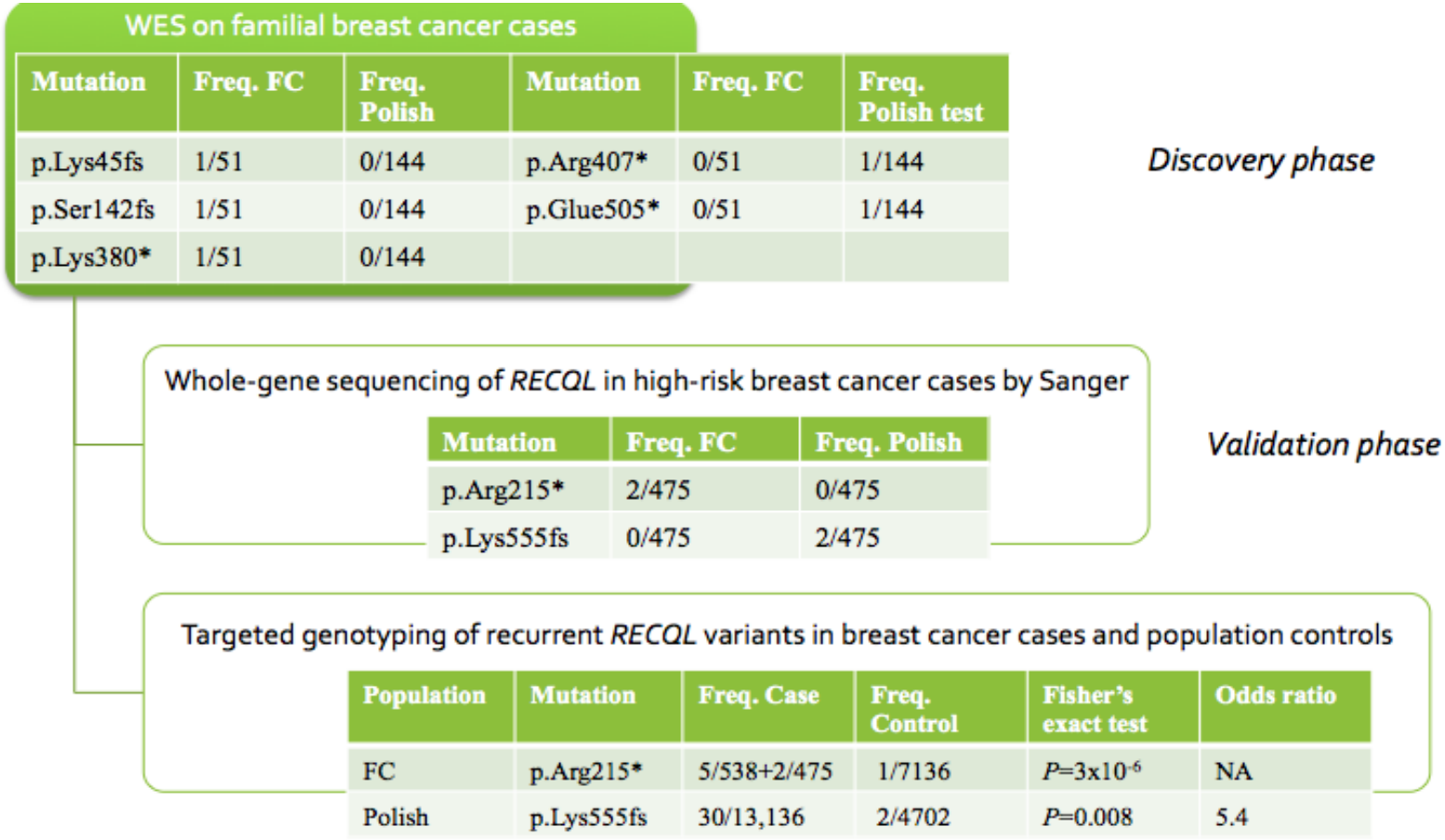


Figure 3.2: *RECQL* truncating mutations identified in French-Canadian and Polish breast cancer cases.

Three mutations were found in 51 French-Canadian familial breast cancer cases using WES. One different mutation was found in two cases from 475 French-Canadian, high-risk breast cancer cases using Sanger sequencing. Recurrent mutation (p.Arg215*) was genotyped in 1,013 high-risk or unselected breast cancer cases. Adapted from Cybulski et al., *Nature Genetics*, 2015.

My colleagues further tested the carrier status in affected relatives in two French-Canadian pedigrees (harboring c.132_135delGAAA and c.426delT mutations) by Sanger sequencing. Both families had a positive family history of ovarian cancer. They confirmed that *RECQL* mutations were segregated with the disease in the families (Figure 3.3). Interestingly, although the presence of *RECQL* mutations was confirmed in tumors from carriers, LOH was not observed as tested by Sanger sequencing (Cybulski et al. 2015).

Meanwhile, a group in Poland who collaborated with us also found two, different truncating mutations in *RECQL* from 144 Polish breast cancer families studied by WES (Cybulski et al. 2015) (Figure 3.2). Therefore, in total, we found five patients (2.6%) who carried different truncating mutations in this gene among a total of 195 French-Canadian and Polish cases studied by WES. In comparison, 0.2% (8/4300) individuals carrying truncating mutations in *RECQL* were reported in the NHLBI exome database ($P = 0.0002$), suggesting a marked difference between the prevalence of the truncating mutation in this gene in our breast cancer cases and in healthy individuals.

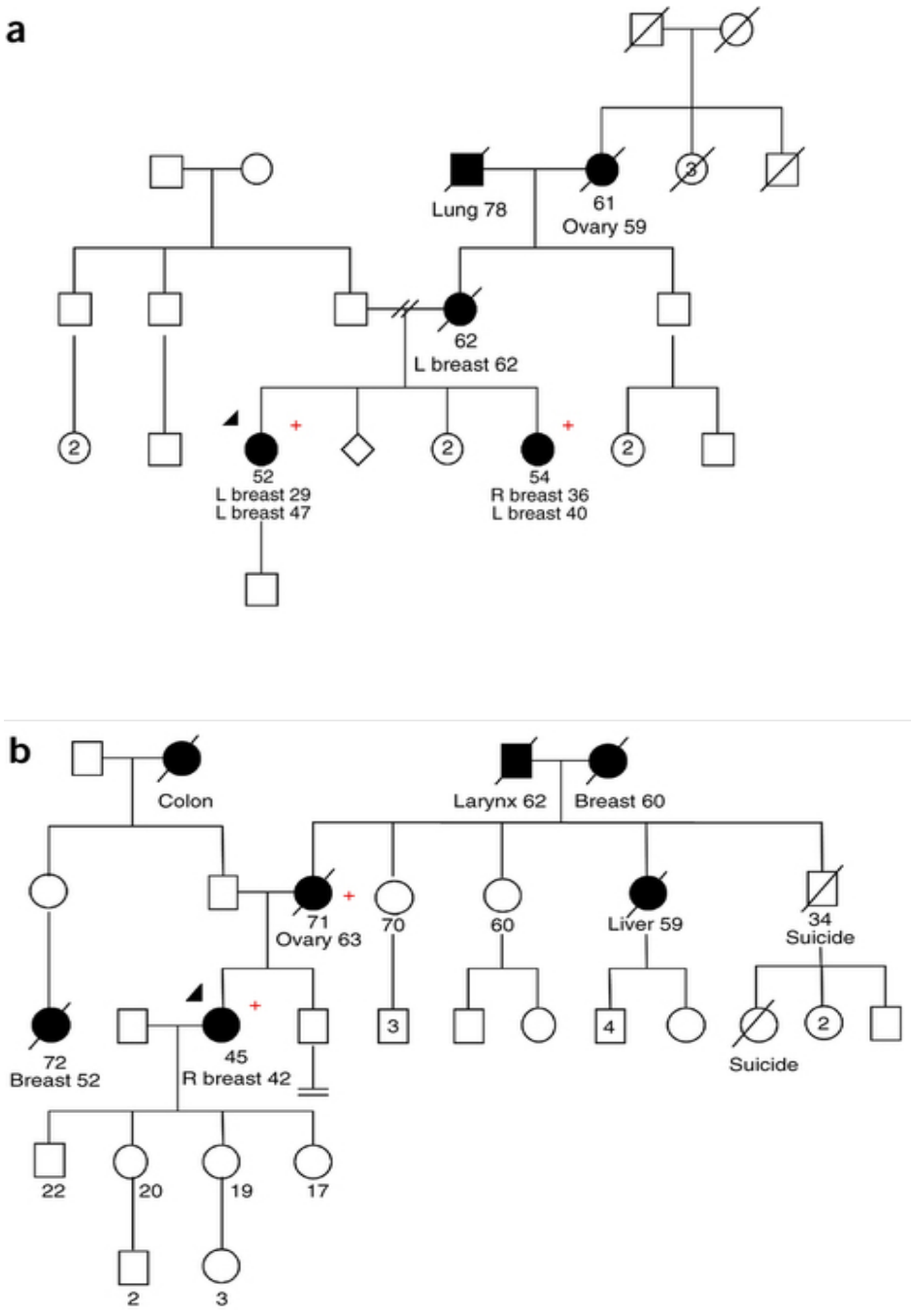


Figure 3.3: Pedigrees corresponding to two French-Canadian *RECQL* mutation carriers.

a) Family carrying c.132_135delGAAA (p.Lys45fs); b) family carrying c.426delT (p.Ser142fs). Adapted from Cybulski et al., *Nature Genetics*, 2015.

3.3.2.3 Identification of French-Canadian founder mutations in ATM and CHEK2

Finally, I searched for potentially interesting mutations in lower-penetrance breast cancer susceptibility genes, including known, moderate-risk genes such as *ATM*, *CHEK2*, and *BRIP1*. Although these genes are known breast cancer susceptibility genes, any observed mutations might potentially represent novel found mutations in the French-Canadian population and were therefore of interest. I found four missense variants in *ATM* and one protein-truncating variant in *CHEK2* (Table 3.4). One variant in *ATM* showed higher minor allele frequency in public databases (>0.001), but was predicted to be protein damaging (Table 3.4).

Table 3.4: Mutations identified in moderate-risk breast cancer susceptibility genes from 51 French-Canadian breast cancer cases.

Gene	Protein change	MAF from 1000 genome	MAF from EVS	Polyphen2 score
ATM	p.V410A	0.0005	0.00177	0.434
	p.L1715P	0	0	1
	p.K1964E	0	0	0.053
	p.G2023R	0.0014	0.002308	1
CHEK2	p.D82fs	0	0	NA

EVS=Exome Variant Server of more than 6000 individuals studied by the NHLBI.

3.3.3 Validating Candidate Genes in Larger Case and Control Cohort

Having a list of candidate variants, the next step was to establish the likelihood that each variant is causal to the disease, rather than a population-specific polymorphism, since it is possible that a French-Canadian specific variant might be identified as a consequence of investigating the French-Canadian population. Thus far, four genes, *ZNF280A*, *PTPRG*, *SMC4 and RECQL* were determined as the most plausible candidates. The *PTPRG* and *SMC4* variants had never been seen in public databases before. The *ZNF280A* variant was found at high-frequency in our cases (6.5%), although 0.16% individuals carried this variant in NHLBI. *RECQL* was the most significantly enriched gene with truncating mutations in our cases, compared to our control samples.

3.3.3.1 Validation of SMC4 mutation in additional French-Canadian breast cancer cases and controls

In the first round of validation, the presence of potential founder variants (variants in *ZNF280A*, *PTPRG*, and *SMC4*) in 1077 unselected French-Canadian breast cancer cases was investigated (Table 3.5a). Those cases were unselected for family history, and all three variants were seen again, with a frequency of 0.65%, 0.84% and 0.46%, respectively. Meanwhile, the frequency of those variants was assessed in the French-Canadian population using a large number of health

controls, based on the established observation that founder populations often harbour population-specific variants not seen in public databases. Because the *ZNF280A* variant was not enriched in additional breast cancer cases (found only in 7/1064 cases), it was not prioritized for further study in the controls. *SMC4* and *PTPRG* variants showed a frequency of 0.10% (2/1932) and 0.45% (8/1745) in French-Canadian newborns, respectively (Table 3.5b). Because of the high frequency of the *PTPRG* variant present in controls, it was categorized as a French-Canadian population-specific variant ($P > 0.05$).

Table 3.5: Validation of potential disease-related founder mutations.

a) Genotyping results of French-Canadian breast cancer cases.

Variant ID	Gene	DNA Change	Protein Change	Frequency
chr22:22868593	*ZNF280A*	c.1362delC	p.S454fs	7/1064
chr3:160143862	*SMC4*	c.C2479T	p.R827C	9/1077
chr3:62257164	*PTPRG*	c.T3116C	p.L1039P	5/1077

b) Genotyping results of French-Canadian newborns. *ZNF280A* variant was excluded because of its higher minor allele frequency from NHLBI exome database and decreased frequency in cases as above.

Variant ID	**Gene**	**DNA Change**	**Protein Change**	**Frequency**
chr3:160143862	*SMC4*	c.C2479T	p.R827C	2/1932
chr3:62257164	*PTPRG*	c.T3116C	p.L1039P	8/1745

Finally, the *SMC4* variant was further pursued in 76 additional high-risk French-Canadian breast cancer cases (defined as aged less than 50 years at diagnosis or had at least two affected first-or second-degree relatives) and 7111 French-Canadian newborns (including the 1932 controls used above) (Table 3.6). Overall, 10 carriers were found in all cases studied (0.87%), whereas 13 carriers were found in controls (0.18%). Our findings suggested that the *SMC4* mutation was significantly associated with French-Canadian breast cancer cases (*P*=0.000545), which showed an enriched frequency among the French-Canadian population.

Table 3.6: Summary of the identification and validation of *SMC4* variant in French-Canadian breast cancer cases and controls.

	***SMC4* (p.R827C)**
FC families	2/51
Frequency	0.0392
FC unselected or high-risk cases	10/1153
Frequency	0.00867
FC newborns	13/7111
Frequency	0.001552
***P*-value**	*P*=0.000545

3.3.3.2 Validation of RECQL mutation in additional breast cancer cases and controls

The validation of *RECQL* was conducted in both French-Canadian and Polish population. In the first phase, all 14 coding exons of *RECQL* were screened by Sanger sequencing among a non-overlapping set of 475 Polish and 475 French-Canadian patients with family history (Figure 3.2). Although none of the five original *RECQL* mutations identified by WES were observed in additional cases, two different truncating mutations in this gene were found (c.643C>T encoding p.Arg215* in French-Canadian cases; c.1667_1667+3delAGTA encoding p.Lys555fs in Polish cases) (Cybulski et al. 2015). Both mutations identified by Sanger sequencing were seen twice in each population (Figure 3.2).

Because the mutations identified from the validation phase were recurrent mutations, they were considered as more likely founder mutations, and therefore, they were further screened in additional breast cancer cases and control cohorts of each population (Figure 3.2). In the French-Canadian cohort, 538 high-risk breast cancer cases (defined same as above) and 7,136 French-Canadian newborns were used. The recurrent mutation (c.634C>T; p.Arg215*) was detected in five cases and one control. Thus, the frequency of this mutation among the two higher-risk sets of French-Canadian cases (475+538) was 0.69% (7/1,013) compared to 0.014% (1/7,136) in controls (P = 0.000003). Likewise, the Polish group also validated that the Polish mutation (c.1667_1667+3delAGTA) in

RECQL was significantly associated with breast cancer ($P = 0.008$). An odds ratio of 5.4 was reported by the Polish group, suggesting *RECQL* as a high-penetrance breast cancer susceptibility gene (Cybulski et al. 2015) (Figure 3.2).

Six missense variants in *RECQL* were identified in the French-Canadian breast cancer cases either by WES or Sanger sequencing (Figure 3.4). All of them were located in the Helicase domain or DNA-binding domain. Three of them were located in conserved residues across species (Figure 3.4). Those three variants had minor allele frequency less than 0.01% reported from The Exome Aggregation Consortium (ExAC), which contains ~60,000 unrelated individuals (http://exac.broadinstitute.org/). C-scores obtained from CADD algorithm indicated that two of them were possible pathogenic (Kircher et al. 2014). However, whether these missense variants in *RECQL* are associated with breast cancer risk requires further functional studies.

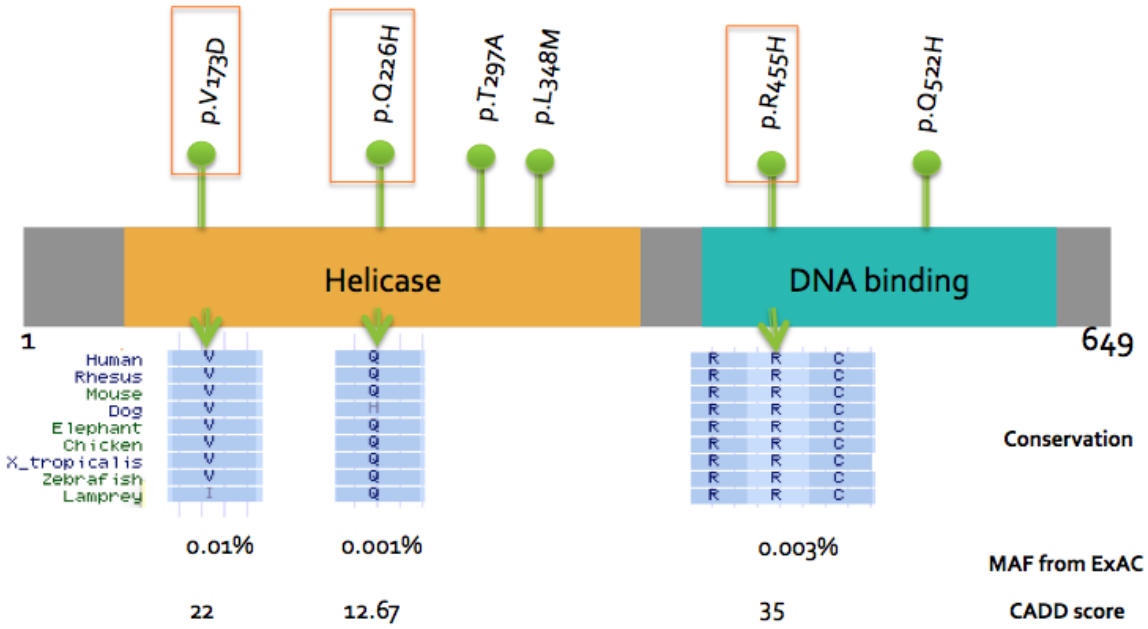


Figure 3.4: Six missense mutations in *RECQL* identified from French-Canadian breast cancer cases.

ExAC = The Exome Aggregation Consortium (http://exac.broadinstitute.org/). CADD score above 15 is usually considered as deleterious.

3.3.4 WES of *RECQL* Mutation Carriers' Tumors

Three tumors (two of them were breast cancer and one was ovarian cancer) and matched normal samples were obtained from two *RECQL*-carrier families for WES. In line with previous results from Sanger sequencing, no obvious LOH of *RECQL* mutations was observed. Moreover, I did not find any somatic mutation in *RECQL*, suggesting the other allele is likely to be functional. However, genome-wide AI analysis identified genomic rearrangements in tumors with

heterozygous *RECQL* mutations, including chromosome 3p, 4p, 8p and 17, which were recurrently observed in three tumors analyzed (Figure 3.5). Additionally, one *TP53* mutation in each tumor (c.G440A, encoding p.Gly147Glu; c.C139T encoding p.His47Tyr; c.374_375insG encoding p.Thr125fs) were identified, suggesting they play a role along with *RECQL* mutation in initiating tumor formation.

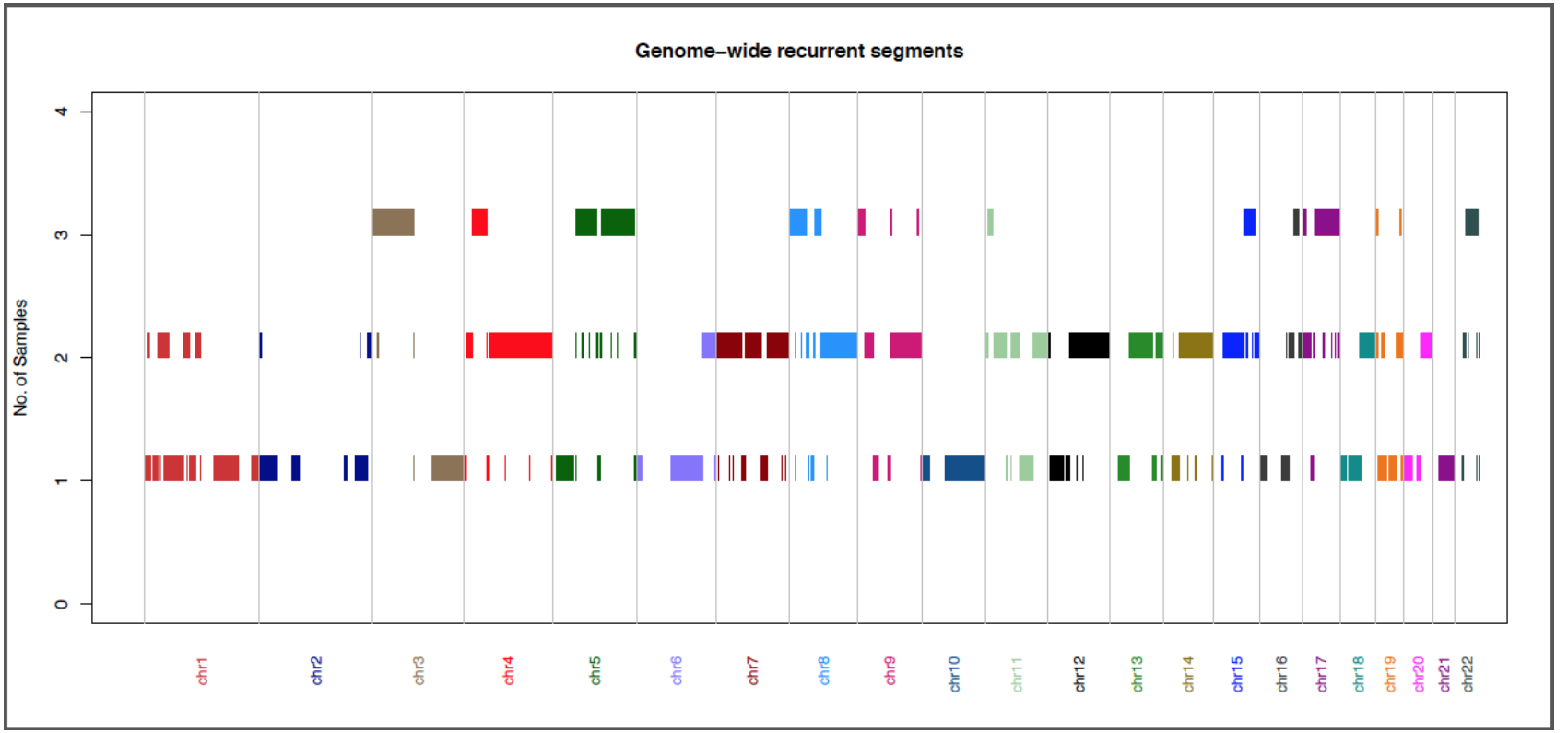


Figure 3.5: Recurrent AI regions identified from *RECQL*+ tumors.

Chromosomes are labled in colors. As shown in the top raw, AI event in chr8p and chr7 (p,q) were seen in the three tumors.

3.4 Discussion

A new breast cancer susceptibility gene *RECQL* was identified via WES analysis of familial breast cancer cases based on the French-Canadian founder population. This gene encodes a member of the RecQ DNA helicase family that is essential for helicase activity. The RecQ helicases prevent the breakdown of DNA replication forks during DNA double-strand repair through the homologous recombination pathway (Chu, Hickson 2009, Wu, Brosh 2010). Other genes in this family, *BLM*, *WRN* and *RECQL4* have been implicated in Bloom, Werner and Rothmund-Thomson syndromes, respectively, which are associated with increased incidence of cancer (Chu, Hickson 2009). Moreover, *BLM* and *RECQL5* have been recently linked to increased susceptibility to breast cancer (Thompson et al. 2012, He et al. 2014). Before this work, no hereditary disorders had been linked with mutations in *RECQL*.

In *RECQL*-knockout mice, aneuploidy, spontaneous chromosomal breakage and translocation events have been observed, suggesting a role for *RECQL* in maintaining genomic stability (Sharma et al. 2006). These observations imply *RECQL* may be a tumor suppressor gene, although complete loss of RECQL was not observed from sequencing the *RECQL* mutated tumors. Previous work suggested that mitotic cell death was induced by suppressing the expression of *RECQL*, leading us to hypothesize that *RECQL*-induced tumorigenesis may occur via *RECQL* haploinsufficiency (Futami et al. 2008). In terms of treatment, it

has previously been suggested that down-regulation of *RECQL* in combination with TOP1 inhibitors could be more effective in treating cancer patients (Berti et al. 2013). Given the fact that *RECQL* mutations are rare among breast cancer patient, further work is required to determine the clinical usefulness of genetic testing of *RECQL* in breast cancer.

Germ-line mutations in *SMC4* (p.R827C) are also found to be associated with breast cancer susceptibility. The missense variant identified through this cohort is likely to represent founder mutation in the French-Canadian population. *SMC4* is believed to play an essential role in chromosome condensation at mitosis (Losada, Hirano 2005). Expression of SMC4 has been suggested as a marker of prognostic prediction in various types of cancers (Wang et al. 2005, Zhou et al. 2012, Jinushi et al. 2014, Zhao et al. 2015). Down-regulation of *SMC4* may affect genomic stability in breast cancer cell lines (Kulawiec et al. 2008). However, the identified *SMC4* variant is not located in a known functional domain and therefore functional study is required to assess the pathogenicity of this variant.

Other interesting variants were identified in genes potentially relevant to cancer. Missense variants in *AATK* and *PRKCD* were found in two French-Canadian breast cancer families. *CNTN4, SMPD1, PRIC285*, *CSDA*, *SMC5* and *SNRPA* were significantly mutated in the breast cancer case cohort. Moreover, novel variants in known breast cancer susceptibility genes *ATM* and *CHEK2* were identified. Further work in genotyping these variants in a larger number of breast

cancer cases and population-specific controls is needed to demonstrate where these variants are associated with breast cancer.

Chapter 4. WES analysis of primary, metastatic and recurrent ovarian carcinomas in a *BRCA1*-positive patient

Methods, figures and tables from this Chapter are published in *BMC Cancer*:

"Exome profiling of primary, metastatic and recurrent ovarian carcinomas in a *BRCA1*-positive patient" Jian Zhang (name before marriage), Yuhao Shi, Emilie Lalonde, Lili Li, Luca Cavallone, Alex Ferenczy, Walter H Gotlieb, William D Foulkes and Jacek Majewski, *BMC cancer*, 13:146 (2013).

4.1 Introduction

Ovarian cancer is the leading cause of death from gynecological cancer in western countries. Approximately 20% of ovarian cancer patients carry an inactivating mutation in either *BRCA1* or *BRCA2* gene, and 15% of them are diagnosed before age 60 (The Cancer Genome Atlas Research Network 2011, Zhang et al. 2011). These observations suggested the pathogenesis of these two

genes in ovarian cancer. DNA-damaging chemotherapeutic drugs for *BRCA1* and *BRCA2* mutation carriers have shown significantly improved outcomes in 5 years, but this survival advantage decreases over time (Bolton et al. 2012, Candido-dos-Reis et al. 2015). These findings emphasized the importance of combining genetic testing and targeted therapy. Moreover, uncovering other genes responsible for ovarian cancer relapse after therapy will be helpful for developing more effective treatments.

To understand the interaction between genetics and treatment response, DNA from four sources (blood, primary tumor, omental metastasis and relapse following standard post-operative therapy with carboplatin and paclitaxel) were obtained from a single patient carrying a deleterious mutation in *BRCA1* (Lee et al. 2010). WES analysis was then performed to allow us identifying subsequent mutations to the *BRCA1* inactivation and gaining insights into the mechanism driving cancer progression.

4.2 Materials and Methods

4.2.1 Clinical History and Tumor samples used for WES

The subject of this study was a 48 year-old patient who had undergone total abdominal hysterectomy for menorrhagia and left salpingectomy for ectopic pregnancy in the past. She had a family history of breast cancer, and was taken to the operating room in September 2003 by general surgery for a suspected diverticular abscess. She was found to have diffuse abdominal carcinomatosis with multiple masses throughout the abdominal cavity. Primary tumor and omental metastatic tumor samples were taken at this time. Final pathology revealed a poorly differentiated serous ovarian cancer. Despite only minimal residual disease being present at the end of the primary surgical resection followed by three courses of neo-adjuvant chemotherapy with carboplatin and paclitaxel, the tumor clinically recurred after only three months of chemotherapy. She then underwent optimal secondary interval cytoreduction with no residual disease. The recurrent sample was taken at this time.

The patient was referred to the medical genetics service and a deleterious missense *BRCA1* mutation (c.5521A>C, p.S1841R), situated in the highly conserved BRCT domain of BRCA1 was identified and found to be segregating with breast and ovarian cancer in her family (Figure 4.1). Despite further

chemotherapy including adjuvant carboplatin-paclitaxel, paclitaxel consolidation, and cisplatin with gemcitabine, liposomal doxorubicin, topotecan, and thalidomide (all of which resulted in short-lived partial responses), the patient died of recurrent disease in August 2007. DNA extracted from the blood used for clinical *BRCA1* testing was subjected to WES. Reproduced from Zhang et al. *BMC cancer*, 2013.

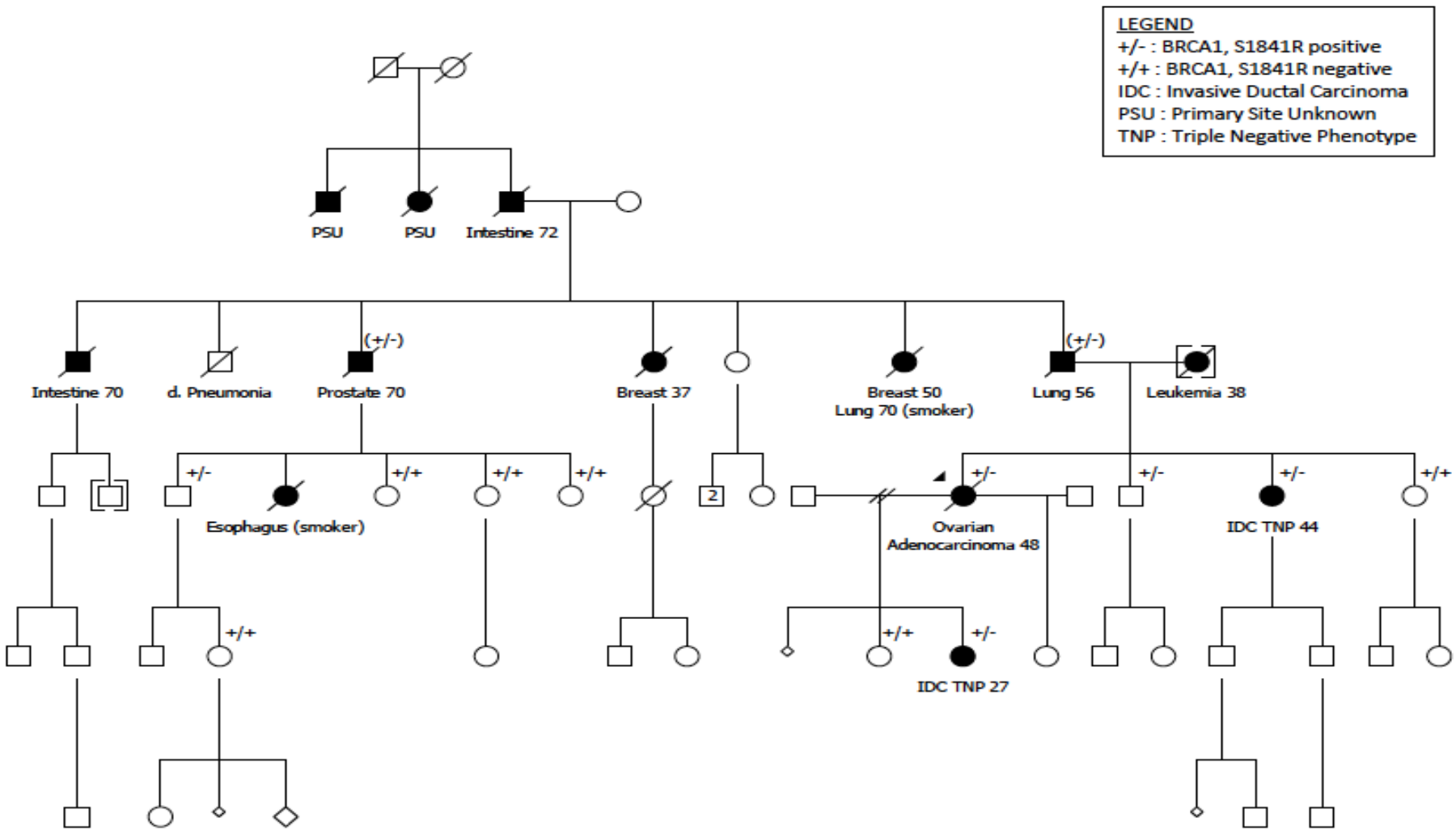

LEGEND
+/- : BRCA1, S1841R positive
+/+ : BRCA1, S1841R negative
IDC : Invasive Ductal Carcinoma
PSU : Primary Site Unknown
TNP : Triple Negative Phenotype
PSU
PSU
Intestine 72
Intestine 70
d. Pneumonia
Prostate 70
(+/-)
Breast 37
Breast 50
Lung 70 (smoker)
Lung 56
(+/-)
Leukemia 38
+/-
Esophagus (smoker)
+/+
+/+
+/+
2
+/-
Ovarian
Adenocarcinoma 48
+/-
+/-
IDC TNP 44
+/+
+/+
+/+
+/-
IDC TNP 27

Figure 4.1: Pedigree of the proband.

The person whose germ-line and tumor DNA was sequenced is indicated with an arrowhead (ovarian adenocarcinoma, age 48). Clear evidence of segregation between the mutation and breast and ovarian cancer is seen by the presence of triple-negative *BRCA1*-related breast cancer in her sister and daughter, who both carry the p.S1841R allele. Other carriers are indicated, with untested obligate carriers indicated as (+/−). Reproduced from Zhang et al. *BMC cancer*, 2013.

4.2.2 Tumor Samples Used For WES

All examined tumor blocks contained poorly differentiated serous adenocarcinoma (Figure 4.2). The histotype was ascertained in routine histological slides obtained from the same tumor, which was fixed in formalin and sections were obtained from paraffin-embedded tissue. This was done because cell morphology was not preserved well enough to provide information on the histotype of the malignant cells. The serous histotype was further demonstrated by immunohistochemistry: the neoplastic cells of all tumor samples stained strongly and diffusely for CA-125, p16, TP53, Ki-67 and WTI. They failed to stain for caldesmon, fascin and only very weakly and focally for B-cadherin. This immunohistochemical profile is consistent with serous differentiation. Reproduced from Zhang et al. *BMC cancer*, 2013.

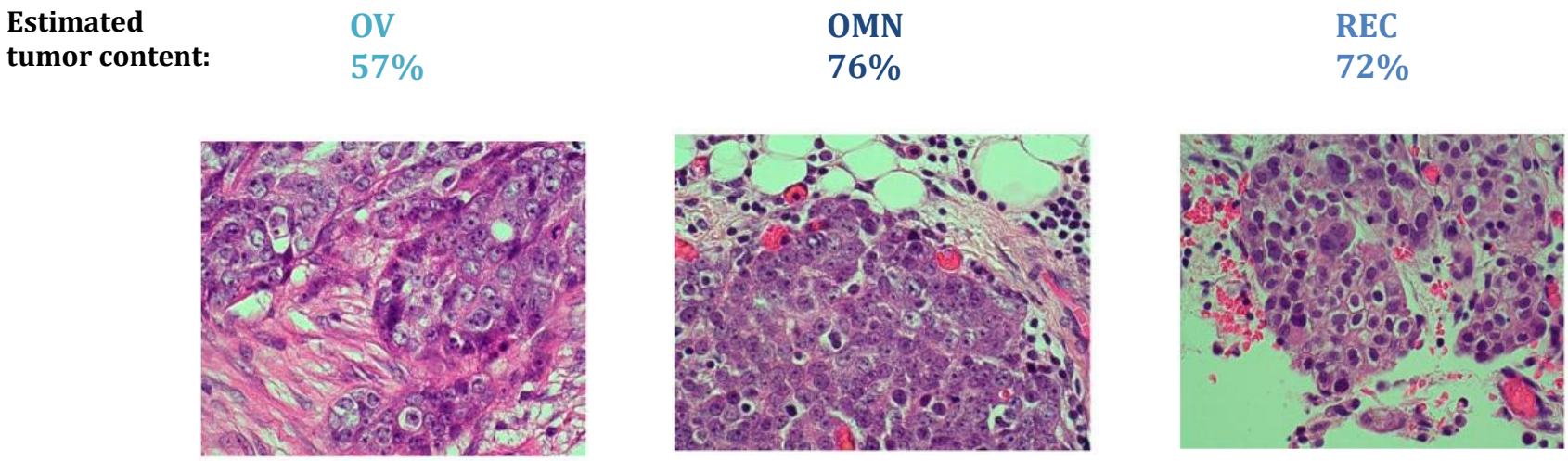


Figure 4.2: Photomicrographs.

Representative frozen tissue was collected at the time of surgery, sections were stained with hematoxylin and eosin and DNA was extracted from the frozen tumors. Because the frozen sections were quite thick, they have not photographed well. Images represented here are the paraffin-embedded tumors that reflect the frozen sections that were used for DNA extraction. A) The poorly differentiated original tumor appeared to be arising from the right ovary (OV). Solid proliferation of highly atypical epithelial cells with enlarged, pleomorphic nuclei and macronucleoli. B) Metastases were widespread, and a biopsy was taken from the omentum (OMN). Solid sheet of malignant cells displaying the same microscopic features as the primary ovarian carcinoma. The tumor cells invade the adjacent fibrofatty tissue of the omentum. C) At laparotomy, the recurrent tumor was found on the surfaces of pelvic and abdominal organs and was biopsied (REC). The

malignant cells are smaller than the primary ovarian and the omentum carcinoma cells. They have clear, cytoplasmic and smudgy nuclear substance, and occasional giant macronuclei and nucleoli. These features may be a reflection of degenerative effects of previous chemotherapy. Estimated tumor content based on the allele frequency of *BRCA1* mutation (discussed above) is indicated. Adapted from Zhang et al. *BMC cancer*, 2013.

4.2.3 Somatic Mutation Detection

Exome of each sample was captured from a total of 3 µg of genomic DNA, using the Illumina TruSeq exome enrichment kit, according to manufacturer's protocols. WES data were generated and processed using the pipeline described above. Next, I applied additional quality control measures to all identified raw variants based on the following criteria: 1) The average Phred-like score is no less than 20 for SNPs and 50 for indels; 2) the total read coverage is no less than five reads; 3) at least three and 10% of covering reads had to support the alternate base.

I further filtered the variants against dbSNP and 1000 genome project dataset (http://www.1000genomes.org), as well as previously identified variants by our lab from more than 100 germ-line, WES samples unrelated to cancer. Only variants that have not been previously observed in any of the control exomes were considered to be potentially functional and selected for downstream analysis. The mutant allelic-fraction of the variants was calculated as:

$$Mutant\ allelic\ fraction = \frac{\text{number of reads supporting alternated base}}{\text{total reads}}$$

In order to validate our WES results, variants with supporting reads increased by at least 10% from the primary tumor to the metastasis or the recurrence were selected for validation using Sanger sequencing. The PeakPicker software was

applied to quantitatively measure the allele proportion of selected SNVs (Ge et al. 2005). The allele proportion from Sanger sequencing data was calculated as:

$$Allele\ proportion = \frac{\text{peak height of alternated base}}{\text{peak height of reference base}}$$

To compare the allelic-fraction from WES and the allele proportion from Sanger sequencing, we converted the allele proportion to mutant allelic-fraction as:

$$Mutant\ allelic\ fraction = \frac{1}{1 + \frac{1}{allele\ proportion}}$$

Adapted from Zhang et al. *BMC cancer*, 2013.

4.2.4 Copy Number Variant Detection

CNV detection was done by comparing normalized read-depth between the blood and each of the primary, metastatic, and recurrent tumors, using a modified algorithm based on ExomeCNV, which was later adapted in FishingCNV (Shi, Majewski 2013, Sathirapongsasuti et al. 2011). In brief, read-depth was first converted to Reads Per Kilobase of exon model per Million mapped reads (RPKM) for each exon, and then the RPKM values were normalized as:

$$Normalized\ RPKM\ value\ = \left(log_2 \frac{RPKM_{tumor}}{RPKM_{blood}}\right)$$

Normalized RPKM values serve as input for DNAcopy, which segments chromosomal regions based on similar log ratios (Venkatraman, Olshen 2007). In this study, because the use of WES data is still not well proven in CNV detection, we refrained from attempting to identify small structural variants and concentrated on larger segments, which we can detect with high confidence. In order to identify large-scale rearrangements, the outputs of DNAcopy were then smoothed by removing small CNV calls and merging adjacent segments. Adapted from Zhang et al. *BMC cancer*, 2013.

4.3 Results

4.3.1 Somatic Mutations Identified by WES in Multiple Tumor Sets

Overall, I identified 39 somatic mutations in the primary and a greater number of somatic mutations in the metastasis and recurrent tumor (Table 4.1). However, I found that all mutations specific to the primary tumor, the metastasis tumor or the recurrent tumor were identified from poor alignments or variant callings (all variants detected in the three tumor samples but also present in the blood sample were excluded). And on visual inspection of the data via IGV, the remaining mutations were also detected in other tumors with small numbers of supporting reads.

Table 4.1: Number of variants called from WES.
Adapted from Zhang et al. *BMC cancer*, 2013.

Sample name	Raw variants	Variants after quality control	Rare variants after filtering	Somatic variants	Validated somatic variants
OV	463944	200059	90	39	24/26
OMN	514227	230935	106	47	24/26
REC	487007	222994	95	52	24/26

OV= primary tumor; OMN=metastatic tumor; REC=recurrent tumor after chemotherapy.

4.3.2 Incomplete LOH of *BRCA1* Mutation Suggesting Tumor Impurity

I then proceeded to examine the change in frequency of the *BRCA1* missense mutation (p.S1841R) and observed that the allelic-faction of this mutation was increased as 0.48 in the blood, 0.57 in the primary tumor, 0.76 in the metastasis, and 0.72 in the recurrence. In Sanger sequencing validation, this mutation showed a consistent increase in ratio: 0.39 in the blood, 0.50 in the primary tumor, 0.68 in the metastasis, and 0.78 in the recurrence.

Of note, the measurements from WES appear more accurate than from Sanger sequencing, because allelic-fraction from WES (0.57) of the inherited *BRCA1* mutation in the blood sample was closer to the expected 0.5, representing heterozygosity. Although we observed an increase in ratio of this mutation during disease progression, we did not observe complete loss of the wild-type allele in the tumors. However, a previous investigation of *BRCA1* mutations in tumors showed frequent LOH events (The Cancer Genome Atlas Research Network 2011). Thus, the allelic-fraction of the *BRCA1* mutation in the three tumor samples should be close to 1, instead of 0.57 to 0.76. Hence, we speculated that the tumor DNA extracted for WES may contain a considerable proportion of non-malignant tissue. Moreover, it appears in the paraffin section image that the primary tumor contains a considerable amount of non-malignant tissue, whereas the percentage of malignant tissue in the omental metastasis and recurrence is

higher (Figure 4.2). Thus, I re-estimated the tumor cell content in the three tumor samples used for WES, based on the allelic-fraction of the *BRCA1* mutation. Therefore, estimated tumor content in the primary, the metastasis and the recurrence is 57%, 76% and 72%, respectively (Figure 4.2).

4.3.3 Validation of Identified Somatic Mutations from Three Tumor Samples

Sanger sequencing validated 24 of 26 somatic mutations as being present in all three tumor samples but not in the blood sample, rendering high confidence in the selected candidate gene list. Observed allelic-fraction was normalized by their estimated tumor content using the allelic-fraction of *BRCA1* mutation, which was approximately 57%, 76% and 72% in the primary, metastasis and recurrence, respectively, thus representing allele frequency of the mutation in the tumors. High concordance of the allelic-fraction estimates from WES and Sanger sequencing was observed (R = 0.78, *P* = 7.865e-15) (Figure 4.3).

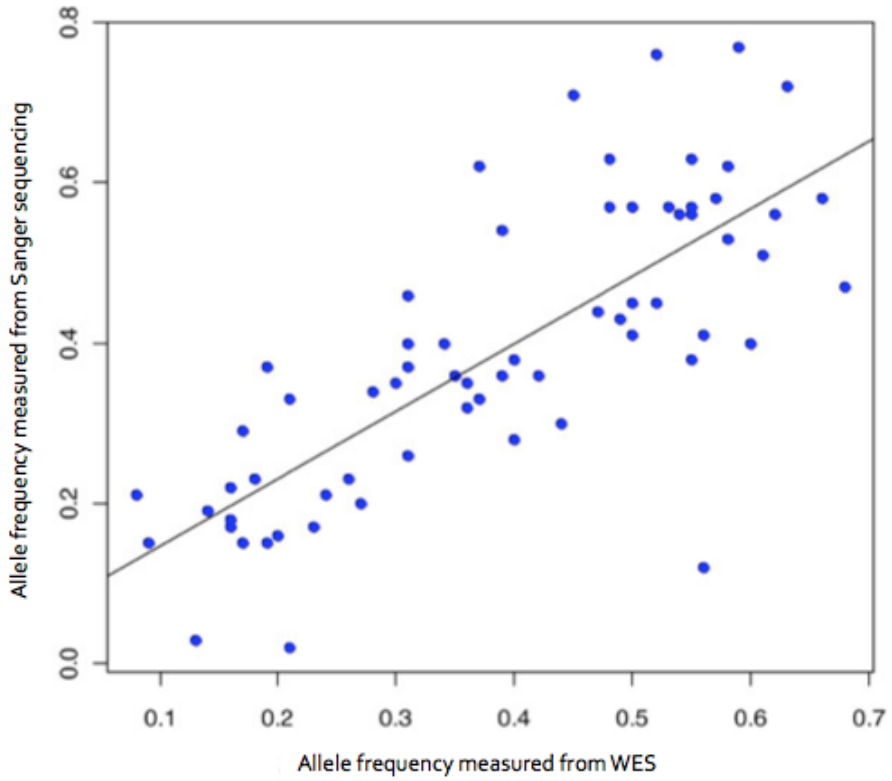


Figure 4.3: Mutation frequencies by two different sequencing methods.
The correlation of mutant allele frequency from WES and Sanger sequencing on validated mutations in the primary tumor, the omental metastasis, and the recurrent tumor after chemotherapy (Spearman's rank correlation= 0.78, *p* = 7.865e-15). Reused from Zhang et al. *BMC cancer*, 2013.

4.3.4 Somatic Mutations Driving Tumor Progression

The presence of important cancer-related mutations was identified, including the above-mentioned *BRCA1* mutation, the missense mutation in *TP53* (c.C329G, p.R110P), and the mutation in *NF1 (*c.G2325+1A*)* damaging the donor site for

splicing (Table 4.2). These mutations showed increased frequency from the primary tumor to the metastatic tumor or the recurrent tumor, emphasizing the importance of their roles in tumor progression. Other genes with novel mutations should also be considered as candidates for intensive investigation, since they were present in all three samples. Our results suggest that the germ-line *BRCA1* mutation might, in combination with somatic mutations in *TP53*, *NF1* and other genes contribute to the tumor initiation and clonal expansion, as well as the relapse of the disease.

Table 4.2: Sanger sequencing confirmed somatic mutations with increased frequencies in tumor samples.

Reused from Zhang et al. *BMC cancer*, 2013.

Position	Gene name	Mutation type	Mutant allele frequency from WES			cDNA change	Protein change	Polyphen score	Mutant allele frequency from sanger sequencing		
			OV	OMN	REC				OV	OMN	REC
chr10:106124579	*CCDC147*	nonsynonymous SNV	0.31	0.45	0.52	c.G529T	p.A177S	0.29	0.40	0.71	0.76
chr17:38173081	*CSF3*	nonsynonymous SNV	0.26	0.49	0.66	c.C493T	p.P162S	0.61	0.23	0.43	0.58
chr15:64496758	*CSNK1G1*	nonsynonymous SNV	0.31	0.50	0.48	c.C881G	p.R294T	1.00	0.46	0.57	0.57
chr17:11696980	*DNAH9*	nonsynonymous SNV	0.24	0.42	0.62	c.A8222C	p.D2741A	0.12	0.21	0.36	0.56
chr4:88533803	*DSPP*	nonsynonymous SNV	0.27	0.61	0.52	c.T465A	p.N155K	0.96	0.20	0.51	0.45
chr20:33874597	*FAM83C*	nonsynonymous SNV	0.16	0.44	0.40	c.G1985A	p.T662M	0.00	0.17	0.30	0.38
chr6:5369392	*FARS2*	nonsynonymous SNV	0.2	0.36	0.35	c.G589A	p.V197M	1.00	0.16	0.35	0.36
chr14:25076412	*GZMH*	nonsynonymous SNV	0.17	0.40	0.37	c.G540T	p.Y180X	NA	0.15	0.28	0.33
chr10:126477647	*METTL10*	nonsynonymous SNV	0.14	0.57	0.60	c.T256C	p.I86V	0.06	0.19	0.58	0.40
chrX:153040228	*PLXNB3*	nonsynonymous SNV	0.17	0.21	0.19	c.G3898C	p.G1323R	0.06	0.29	0.33	0.37
chr12:3692299	*PRMT8*	nonsynonymous SNV	0.30	0.55	0.55	c.G904A	p.D302N	1.00	0.35	0.57	0.58

chr2:65316194	*RAB1A*	nonsynonymous SNV	0.18	0.37	0.39	c.T299C	p.N100S	0.00	0.23	0.62	0.54
chr7:122338859	*RNF133*	nonsynonymous SNV	0.17	0.36	0.34	c.C114T	p.W38X	NA	0.15	0.32	0.40
chrX:30870990	*TAB3*	nonsynonymous SNV	0.09	0.37	0.39	c.C1615T	p.E539K	0.07	0.15	0.33	0.36
chr1:234565362	*TARBP1*	nonsynonymous SNV	0.28	0.50	0.53	c.C2671T	p.D891N	1.00	0.34	0.45	0.57
chr17:7579358	*TP53*	nonsynonymous SNV	0.21	0.47	0.68	c.C329G	p.R110P	0.85	0.02	0.44	0.47
chr7:158824649	*VIPR2*	nonsynonymous SNV	0.13	0.63	0.59	c.G1081T	p.L361M	1.00	0.03	0.72	0.77
chr16:72828578	*ZFHX3*	nonsynonymous SNV	0.23	0.54	0.58	c.C8003T	p.R1754Q	0.45	0.17	0.56	0.53
chr19:58420819	*ZNF417*	nonsynonymous SNV	0.19	0.56	0.5	c.G827C	p.S276C	0.89	0.15	0.41	0.42
chr17:29554310	*NF1*	splice site SNV	0.16	0.56	0.48	c.G2325+1A	NA	NA	0.18	0.12	0.63
chr19:46192605	*SNRPD2*	splice site SNV	0.31	0.58	0.55	c.G378-1A	NA	NA	0.26	0.62	0.63
chr3:195022735-195022753	*ACAP2*	frameshift deletion	0. 15	0.41	0.55	c.1267_1285del	p.R423fs*26	NA	NA	NA	NA
chr1:201983017-201983030	*ELF3*	frameshift deletion	0.17	0.15	0.34	c.866_879del	p.N289fs*7	NA	NA	NA	NA
chr13:108922263-108922263	*TNFSF13B*	frameshift deletion	0.17	0.36	0.31	c.20delG	p.E8fs*15	NA	0.21	0.22	0.37

OV= primary tumor; OMN=metastatic tumor; REC=recurrent tumor after chemotherapy.

4.3.5 Landscape of Three Tumors Revealed by WES

Numerous CNVs were identified, which is consistent with aneuploidy in high-grade serous ovarian cancers. CNV detection suggested that chromosome 17q containing *BRCA1, TP53 and NF1* genes were deleted in all three tumors, which should be considered as evidence of LOH of the *BRCA* mutation in all tumors (Table 4.3). Meanwhile, CNVs associated with known ovarian cancer mutations, such as the amplification of 8q harboring the *MYC* oncogene were found in all tumors including the primary tumor (Table 4.3).

Table 4.3: CNVs that were detected in primary, metastatic and recurrent tumors.

Reused from Zhang et al. *BMC cancer*, 2013.

Region	Type	CNV segements indicating deletion/amplification					
		OV		OMN		REC	
		Coordinates	Mean log ratio	Coordinates	Mean log ratio	Coordinates	Mean log ratio
1p35-1p36	Del	861393-12980233 13910301-22895846	-0.3979 -0.391	861322-27589726 28059114-29652173	-0.5557 -0.5209	861322-27589726 28059114-29650008	-0.5263 -0.5321
Chr4	Del	264888-42088143 42145445-88235112 88258428-190874280	-0.1326 -0.1643 -0.1689	264888-1389640 20255439-145040934 148785997-189026086	-0.5261 -0.5052 -0.5084	264888-1389640 18023221-141832508 147227078-190873442	-0.5903 -0.5217 -0.5302
6q16-6q25	Del	153313992-170176161	-0.2665	96971022-170893669	-0.5104	96969750-170893669	-0.5462
8p21-8p23	Del	117024-28385681	-0.287	190896-28385681	-0.5488	190896-28385681	-0.5817
8q21-8q24	Amp	90775210-122641580 123963751-142226069 142227189-145278133 145515440-146279543	0.5658 0.98 0.5909 0.5688	90926305-95709154 97605708-122641580 123963751-145725582	0.5043 0.927 1.3829	91836945-97172920 97243283-121357802 121379410-145622144	0.5658 0.9853 1.4429
11q12-11q14	Amp	64676463p-134251918	0.1758	63581159-94354158	0.7324	63766427-94354158	0.7829
12p12-12p13	Amp	250451-6637339 6638679-9262631 9264755-13140266 13208485-31107009	0.1653 0.188 0.3317 0.2592	247439-22089608	0.4673	247439-22089608	0.4963
12q21-12q24	Del	31116761-121883221 121970711-131616361 132195775-133781116	-0.1361 -0.3135 -0.3871	65078567-113909303 114282473-133781116	-0.5148 -0.55	64668681-133781116	-0.5465

16q21-16q24	Del	3725325-90142318*	-0.2189	50102691-90030718	-0.5425	50069328-69988476 70428885-90142318	-0.563 -0.5792
17p + 17q11-17q21	Del	171206-7755654 7758393-18286499 18539775-42328956	-0.3947 -0.3397 -0.3036	63643-36881851 36894606-41234592	-0.5335 -0.5191	63643-36709091 36865426-41256973	-0.5552 -0.546
19p13.3	Del	374421-8429523	-0.448	474621-8194249	-0.5189	110679-8402712	-0.5409
19p13.2	Amp	8555110-11531615 11559037-16639066	0.1418 0.1043	8429206-18541740	0.4018	8429206-10625687 10677734-11031424 11031510-18548570	0.4414 0.8088 0.4299
19q13.2- 19q13.4	Del	17317922-59082756	-0.2849	41626252-59082756	-0.5468	41306478-59082756	-0.5686
22q	Del	17073440-18909917 19029320-42999166 43023310-51065480	-0.362 -0.3716 -0.4172	16448824-51133476	-0.517	17071767-51065188	-0.5632

Mean log ratio = mean of the RPKM log ratio of the segment; Del = deletion (mean log ratio < 0); Amp = amplification (mean log ratio >1); OV= primary tumor; OMN=metastatic tumor; REC=recurrent tumor after chemotherapy; Del=Deletion; Amp=amplification.

WES results showed that the degree of all the identified CNVs was increased from the primary tumor to the metastatic and the recurrent tumors (Figure 4.4, Tables 4.3). Moreover, no *de novo* CNVs were found specific to the primary tumor or in the subsequent tumors (Table 4.3). The landscape of the three tumor sets appeared to be identical (Figure 4.4). This again, supported our hypothesis that the primary tumor sample we obtained for WES contained a relatively larger proportion of normal tissue than the metastatic and recurrent samples. The increased degree of structural variants was likely to reflect tumor purity, as opposed to a selection process.

.

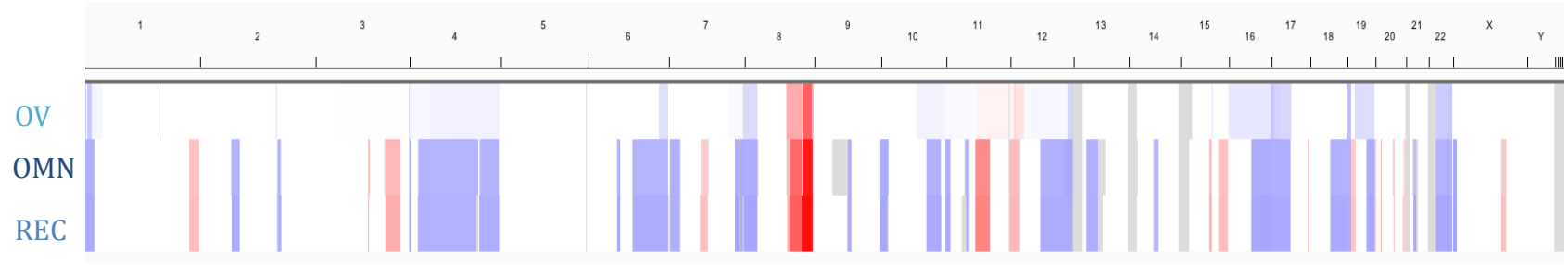


Figure 4.4: Copy number variants in the ovarian tumors.

Filtered CNVs in the OV, OMN, and REC tumors across the genome, with chromosomal labels at the top. Because we were only interested in large-scale deletions and amplifications, smaller CNV calls were removed and adjacent segments were merged. In the heat map, red indicates amplifications, blue indicates deletions, grey indicates missing data. The magnified CNV patterns from OV, to OMN, to REC are likely due to differences in tumor purity. Notable amplifications are seen in 8q and 11q. Deletions are seen in chr4, 6q, 7q, 12q, 16q, chr17, chr19, chr22. Reused from Zhang et al. *BMC cancer*, 2013.

4.4 Discussion

In this study, an analysis on the whole exome was performed to identify potential driver mutations, as well as large chromosomal rearrangements. We observed the LOH in the *BRCA1* mutation in the primary and subsequent tumors, and somatic mutations in the *TP53* and *NF1* genes were identified, suggesting their role along with *BRCA1* driving the tumor development. Notably, the patient responded very poorly to platinum-based therapy and relapsed quickly. This early platinum failure is somewhat less common in *BRCA1*-related cancer than in non-hereditary ovarian cancer (Bolton et al. 2012). Deleterious somatic mutations present in the primary tumor likely contributed to the rapid progression of the disease.

Genetic evolution of tumors from diagnosis to relapse following highly active chemotherapy was not observed. Instead, all the cancer-driving events (deleterious mutations in *TP53* and *NF1*, amplification in *MYC*) were already present in the primary tumor. Although increased mutant allele frequencies were found in the three tumors, little selection might exist thereafter. During the primary surgery, it was not possible to identify the tumor of origin. It is possible that the primary tumor was, in fact, a secondary tumor of the tumor origin with a full capacity of metastasis. It is believed that the origin of HGSC carrying *BRCA1* mutation is in the fallopian tube (Piek et al. 2001). Thus, it is likely that the

“primary tumor” used in this study was evolved and metastasized from the fallopian tube.

The *NF1*-associated RAS pathway is often activated in ovarian cancers (The Cancer Genome Atlas Research Network 2011, Sangha N. et al. 2008). In additional, somatic mutations in *NF1* may frequently co-occur with *TP53* mutations in HGSC (Sangha N. et al. 2008). With a 2.5 fold increase in frequency from the primary tumor to the metastatic tumor, it is possible that the *NF1* mutation appeared in the primary tumor later than the *TP53* mutation. However, further research in a larger number of patients is required to fully understand whether combined therapy targeting both the RAS signaling pathway and the DNA repair pathway will benefit the patient’s outcome (Downward 2003, Lord, Ashworth 2012).

Chapter 5. LoLoPicker — Detecting Low Allelic-Fraction Variants in Low-Quality Cancer Samples from Whole-exome Sequencing Data

A version of Chapter 5 is under preparation for publication. Author contributions are stated in the Contribution of The Authors section.

5.1 Introduction

Cancer arises from cells that have acquired somatic mutations conferring selective advantages to allow the cells proliferating autonomously (Stratton, Campbell & Futreal 2009). For this reason, when sequencing the tumor cells, identified driving mutations should not present in the normal cells, either from the same cancer patient or from other cancer-free individuals. However, identifying those somatic events remains challenging. One of the major complexities is that variants with low allelic-fraction are commonly observed in tumor samples owing to normal tissue contamination, local copy number change and cancer heterogeneity. The difficulty of identifying those low allelic-fraction variants is magnified by the fact that sequencing technologies are imperfect and produce errors (Flickinger et al. 2015).

Previously identified sources suggest that artifacts can occur both randomly and systematically in a manner of sequence-dependent and site-specific (Wilm et al. 2012, Gerstung, Papaemmanuil & Campbell 2013, Wang et al. 2013). However, other sources of artifacts likely exist and may not be characterized yet. Failing to remove unknown artifacts can significantly affect the specificity of variant calling, especially in calling low allelic-fraction variants, since their allelic-fraction in the tumor approaches the error rate of sequencing technologies. For instance, technical artifacts may arise from the formalin fixation process, and therefore, decrease the accuracy of calling variants from FFPE samples (Williams et al. 1999, Van Allen et al. 2014).

WES has emerged as a promising tool to discover disease-causing genes. For many basic research or clinical laboratories, the number of samples being sequenced has increased dramatically. Some laboratories build their in-house database of WES data to enable them to filter out false-positive calls that are specific to library preparation, protocols, instruments, environmental factors or analytical pipeline. Such database also provides an opportunity to 1) rule out polymorphisms not reported by the public databases, and to 2) precisely estimate the site-specific error rates using control samples. This precise, site-specific error rate gives the advantage to increase the sensitivity of calling low allelic-fraction SNVs on sites with lower error rates, and reduce false positives on sites with high error rates. This idea has been successfully implemented in Shearwater for targeted re-sequencing experiments, which is not designed for analyzing WES

data (Gerstung, Papaemmanuil & Campbell 2013). MuTect recommends using a panel of normal samples to filter missed germ-line variants and error-prone sites (Cibulskis et al. 2013). However, this strategy is limited in identifying error-prone sites or retaining sites with low-level artifacts, even using MuTect's artifact detection mode. A comparison of the performance of current somatic SNV callers has suggested that they have significant room for improvement, especially for detecting low allelic-fraction variants with high accuracy (Wang et al. 2013).

Here, I present LoLoPicker, a tool dedicated to call somatic SNVs from WES data using tumor and its matched normal tissue, plus a user-defined control cohort of germ-line, non-cancer samples. The goal is to reach higher specificity than current somatic SNV callers, including MuTect, VarScan2 and LoFreq (Wilm et al. 2012, Cibulskis et al. 2013, Koboldt et al. 2012). We observed a superior performance of LoLoPicker compared to other programs. Our approach is particularly suited for FFPE samples, since FFPE-specific errors can be identified from a panel of FFPE controls.

The LoLoPicker algorithm is implemented in Python language and the package is released at https://github.com/jcarrotzhang/LoLoPicker.

5.2 Materials and Methods

All samples underwent the same protocol for sequencing and pipeline for alignments as described in Chapter 2. FFPE tissue-derived DNA was captured by using the Nextera Rapid-Capture Exome kit and processed as described in section 2.2.2. To ensure the best performance of MuTect and LoFreq, GATK BaseRecalibrator was used to increase the quality score accuracy (McKenna et al. 2010). Selected SNVs were validated by performing targeted re-sequencing, using a MiSeq sequencing platform with an average coverage of 5000X.

Table 5.1: Primers used for targeted re-sequencing validation

Position	Gene	Forward	Reverse
chr1:89521863	*GBP1*	ACACTGACGACATGGTTCTACAGCTTGGTCACCTTGGTGTTT	TACGGTAGCAGAGACTTGGTCTCATTAAAGGCCCAGCTAGAAAA
chr13:24895566	*C1QTNF9*	ACACTGACGACATGGTTCTACACAGGGTGAGCCAGGAGTC	TACGGTAGCAGAGACTTGGTCTAGGCCTGGTCCTCAGAGC
chr3:178952085	*PIK3CA*	ACACTGACGACATGGTTCTACAATGATGCTTGGCTCTGGAAT	TACGGTAGCAGAGACTTGGTCTCAATTCCTATGCAATCGGTCT
chr9:5231708	*INSL4*	ACACTGACGACATGGTTCTACACCCATGCCTGAGAAGACATT	TACGGTAGCAGAGACTTGGTCTCCCATGAGATTTCTGGTGAGA
chr11:71907000	*FOLR1*	ACACTGACGACATGGTTCTACAGGCTGGCAGACCTCAAGATA	TACGGTAGCAGAGACTTGGTCTTCATGGCTGCAGCATAGAAC
chr13:25378544	*RNF17*	ACACTGACGACATGGTTCTACATCATCCACCTATTTTGCCTAAAG	TACGGTAGCAGAGACTTGGTCTAAATCATATAAACTTGTTTGAAGTTGC
chr19:40580859	*ZNF780A*	ACACTGACGACATGGTTCTACAGAGTTTTCTGATGTTGGGAAAG	TACGGTAGCAGAGACTTGGTCTTCCAATGAGAAACCTTTTGTATG
chrX:3240813	*MXRA5*	ACACTGACGACATGGTTCTACAAAGGTGTGCAAAGGTGTCTTC	TACGGTAGCAGAGACTTGGTCTTGAACCATCTCCTACTCTGCAC
chr1:65301884	*JAK1*	ACACTGACGACATGGTTCTACACAGCCATGGGACTAGAATCTG	TACGGTAGCAGAGACTTGGTCTACCATAGCAGCGTATACATGG
chr2:62099221	*CCT4*	ACACTGACGACATGGTTCTACAATGCAGCAGGCCTCATATTT	TACGGTAGCAGAGACTTGGTCTTGCTTTTGCAGATGCTATGG
chr15:22742690	*GOLGA6L1*	ACACTGACGACATGGTTCTACAGAACCAGCAACAGGAGGAGA	TACGGTAGCAGAGACTTGGTCTTGCATCTTCTCTTCCAGCTCC
chr3:4715013	*ITPR1*	ACACTGACGACATGGTTCTACAGATATCAGCTGAACCTCTTTGC	TACGGTAGCAGAGACTTGGTCTGGCTATCTACTGCCGCACA
chr18:77805926	*RBFA*	ACACTGACGACATGGTTCTACAGTGCTTGGTGTGAAGCCTCT	TACGGTAGCAGAGACTTGGTCTTCTGCCTCCAACTCCTCTGT
chr6:116912080	*RWDD1*	ACACTGACGACATGGTTCTACAGCAGATACATTTCATATGCCACTT	TACGGTAGCAGAGACTTGGTCTTGGAATTCTACTATTATCTTACCATCC

5.3 Results

5.3.1 Description of LoLoPicker Method

LoLoPicker allows users to provide a control cohort, which contains normal samples that underwent similar procedures as the test sample (tumor), and uses the control cohort to estimate site-specific error across the exomes. For variants observed in the tumor sample, a binomial test is performed to determine whether the ratio of altered reads of the tumor variant exceeds the background error rate obtained from the control samples, which enables high sensitivity and specificity (Figure 5.1). A detailed description of this algorithm is followed.

Step one: raw variant calling

LoLoPicker first walks through the tumor exome using pysamstats (https://github.com/alimanfoo/pysamstats), and identifies sites with reads containing non-reference bases. Then, for each of those potential SNV, LoLoPicker filters reads based on their base quality scores (default setting is 30) and mapping quality scores (default setting is 30) (Table 5.2). Particularly, overlapping read-pair covering the same variant, meaning that they sequence a variant from the same DNA fragment, are counted once. Additionally, a variant site is filtered out if 1) altered base is supported by less than two reads, and if 2) in either strand, the altered alleles are clustered at the first five or the last five bases of the reads. The same filters are applied to the matched normal sample.

Finally, variants with allelic-fraction greater than 10% or with read count larger than two are flagged as possible germ-line variants and are excluded from downstream analysis (Figure 5.1, Table 5.2).

Figure 5.1: Overview of the workflow of LoLoPicker.

Step 1: LoLoPicker first performs raw variant calling using tumor and matched normal sample. Step 2&3: LoLoPicker then performs its core statistical framework using a user-provided control cohort.

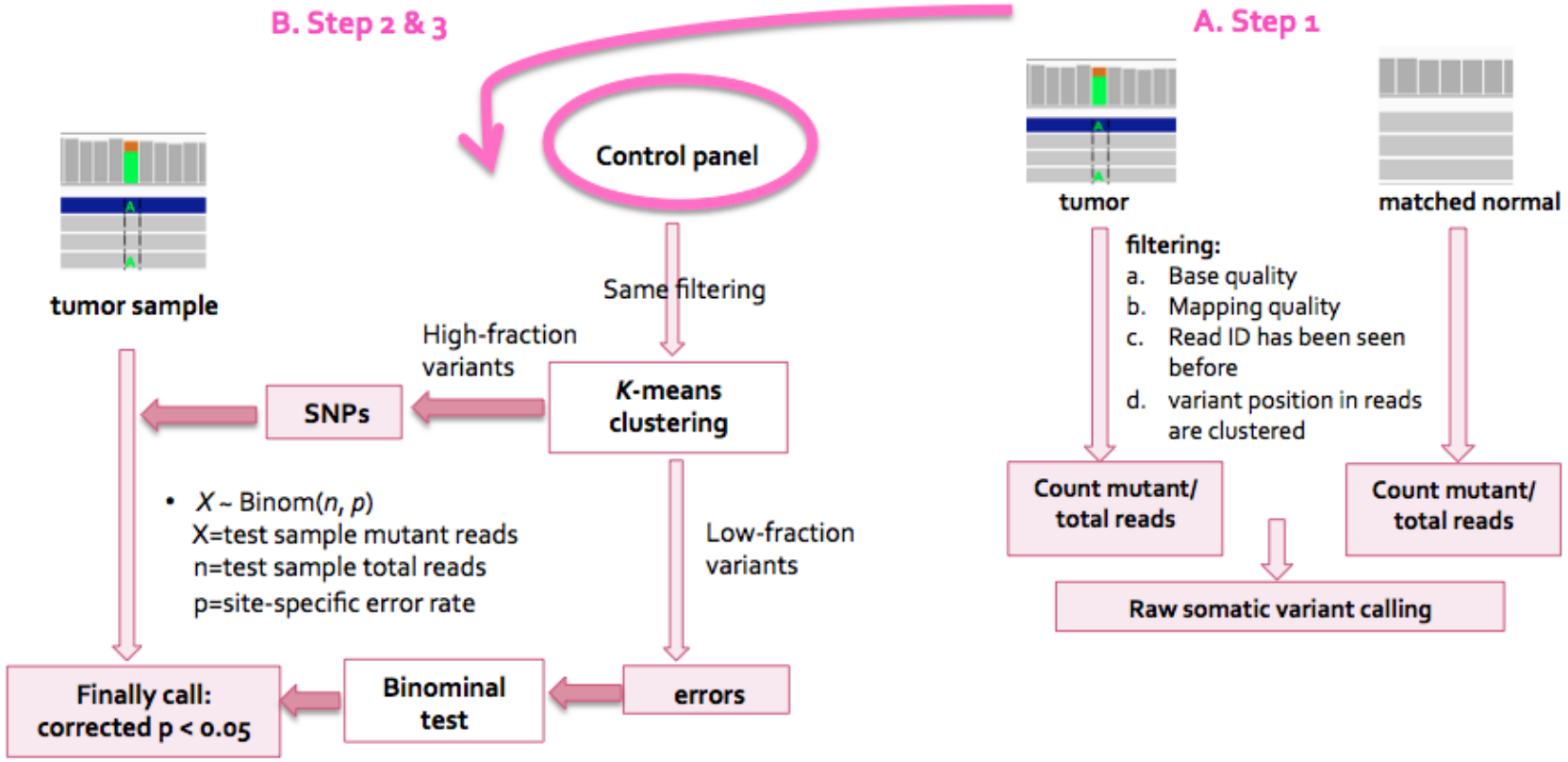


Table 5.2: Default thresholds of LoLoPicker filters.

	Class	Default thresholds
Base quality	Read filter	Base quality score <30
Mapping quality	Read filter	Mapping quality score < 30
Overlapped read pairs	Read filter	Read mate is previously counted
Position in read	Site filter	In either forward or reverse strand, all altered bases are located in the first or last five read positions
Variant in matched normal	Site filter	Number of altered reads < 2 in the matched normal sample
Variant in control cohort	Site filter	Number of altered reads >= 2 in the matched normal sample

Step two: statistical framework

At a given site, LoLoPicker counts reads of each sample in the user-provided control cohort, with the same filters as described above. An SNP is called, if the variant allelic-fraction is larger than 50%, and the total coverage is larger than 10. Then, a *K*-means clustering is performed one the remaining variants based on their allelic-fraction in order to identify two clusters representing true variants and errors, respectively. Variants in the cluster of the larger mean are considered as SNPs; whereas variants in the cluster of the smaller mean are considered as errors. If a variant is observed in the control cohort for more than three times, it is flagged as a possible SNP and is excluded from further consideration.

Then, using site errors identified from the control dataset, a site-specific error rate is calculated as the ratio of read count supporting altered base and the total

number of reads covering the site. Finally, a binomial test is performed, followed by Bonferroni correction with a cut-off of 0.05 for significance. The statistical power is calculated as:

$$\text{power} = 1 - \sum_{Ct=0}^{n} \text{binom}(Ct, n, f)$$

Ct = critical value obtained from the first binomial test; n = total coverage; f = 0.3 for tumor and f = 0.5 for normal sample. A cut-off of 0.95 is applied to ensure using sites covered with 95% power to be able to detect a variant with 30% allelic-fraction in the tumor, and 50% allelic-fraction in matched normal sample.

5.3.2 Benchmarking Analysis

To access the performance of LoLoPicker in comparison to other variant callers, I benchmarked LoLoPicker, MuTect (Version 1.1.6), VarScan2 (Version 2.3.6) and LoFreq (Version 2.1.2) against two datasets (Wilm et al. 2012, Wang et al. 2013, Koboldt et al. 2012). For the assessment of sensitivity, I used data of the primary ovarian tumor and matched blood sample described in Chapter 4. Variants were called in a new tumor sample, which was a mixture of the primary tumor and the blood, to ensure that variants were present in low allelic-fraction (between 1% to 12%). The new tumor sample was also randomly down-sampled to 10% of total reads. Only those Sanger validated variants were considered as true positives

(Table 5.3). For the assessment of specificity, a sample that underwent WES twice in two different batches was used, and all variants called between the two batches were considered as false positives.

Table 5.3: Status of true positives used for benchmarking analysis.

An ovarian tumor with validated somatic mutations was merged with its matched blood sample and down-sampled to 10%.

Position	Gene	Mutation	Reference coverage	Altered_base coverage	Allelic-fraction	Reference coverage *	Altered_base coverage*	Allelic-fraction*
chr10:106124579	*CCDC147*	p.A177S	185	25	0.12	15	6	0.29
chr17:38173081	*CSF3*	p.P162S	89	8	0.08	NA	NA	NA
chr15:64496758	*CSNK1G1*	p.R294T	301	46	0.13	29	5	0.15
chr17:11696980	*DNAH9*	p.D2741A	277	31	0.10	26	6	0.10
chr4:88533803	*DSPP*	p.N155K	202	21	0.09	18	3	0.14
chr20:33874597	*FAM83C*	p.T662M	139	8	0.05	4	2	0.33
chr6:5369392	*FARS2*	p.V197M	53	4	0.07	NA	NA	NA
chr14:25076412	*GZMH*	p.Y180X	428	34	0.07	35	4	0.10
chr10:126477647	*METTL10*	p.I86V	446	28	0.06262	58	5	0.08
chrX:153040228	*PLXNB3*	p.G1323R	39	3	0.07	NA	NA	.NA
chr12:3692299	*PRMT8*	p.D302N	124	19	0.13	16	3	0.16
chr2:65316194	*RAB1A*	p.N100S	715	52	0.07	80	6	0.07
chr7:122338859	*RNA133*	p.W38X	422	34	0.07	37	5	0.12
chrX:30870990	*TAB3*	p.E539K	262	11	0.04	22	2	0.08
chr1:234565362	*TARBP1*	p.D891N	385	61	0.14	NA	NA	NA
chr17:7579358	*TP53*	p.R110P	161	10	0.06	32	5	0.14
chr7:158824649	*VIPR2*	p.L361M	116	6	0.05	18	2	0.10
chr16:72828578	*ZFHX3*	p.R1754Q	428	34	0.07	66	6	0.08
chr19:58420819	*ZNF417*	p.S276C	113	11	0.09	49	5	0.09
chr17:29554310	*NF1*	NA	172	12	0.07	29	6	0.17
chr19:46192605	*SNRPD2*	NA	128	21	0.14	NA	NA	NA

*****Variant status after down-sampling.

Default parameters were applied in all programs for variant calling, except in VarScan2, --somatic-p-value 1, and --min-var-freq 0 were used to allow calling low allelic-fraction SNVs. In MuTect, the --normal_panel option is used to allow filtering missed germ-line variants from a panel of normal samples (same panel as used in LoLoPicker) as suggested by the program. The performances of tested algorithms were analyzed and visualized by receiver operating characteristic (ROC) analysis, based on the binomial *p*-value for LoLoPicker, the tumor Fstar LOD score for MuTect, the somatic *p*-value for VarScan2, and the VCF quality score for LoFreq. The ROC analysis was conducted using ROCR package (Sing et al. 2005). To allow fair comparisons, variants present in more than one read in the matched normal sample were removed from the ROC analysis, whereas variants flagged as “possible contamination” from MuTect, and dbSNP variants flagged by LoFreq, were retained.

As the results, LoLoPicker showed much better specificity, while maintained the highest sensitivity, particularly for calling variants at very low allelic-fraction (Figure 5.2). When reducing the coverage of variants between two to 10, the sensitivity of all callers was dropped; whereas LoLoPicker and MuTect showed the highest sensitivity (Table 5.4).

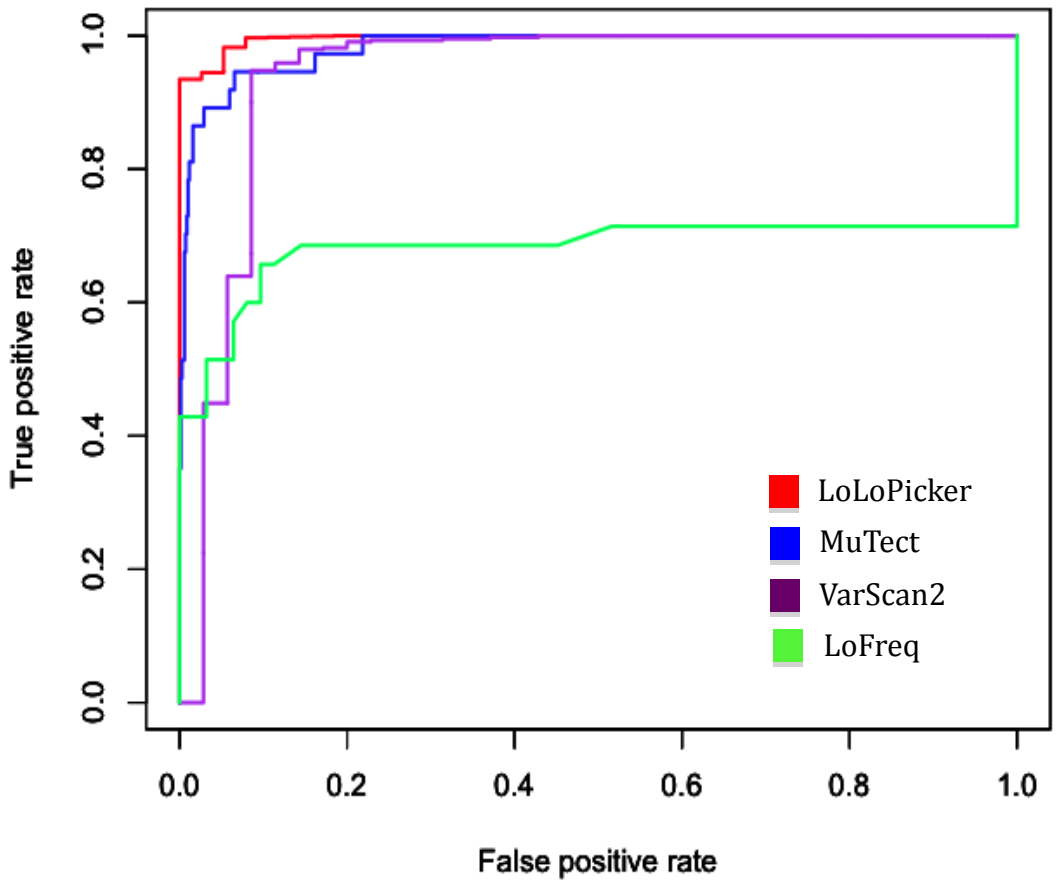


Figure 5.2: ROC analysis.

The performances of LoLoPicker, MuTect, VarScan2, and LoFreq in calling low-fraction SNVs are compared using benchmarked samples.

Table 5.4: Number of true positives and false positives called by LoLoPicker, MuTect, VarScan and LoFreq from benchmarked samples.

Tools	True Positives		False Positives
	High Coverage	Low Coverage	
LoLoPicker	18/18	9/13	3
MuTect	18/18	9/13	15
VarScan2	18/18	8/13	21
LoFreq	18/18	7/13	53

5.3.3 Applying LoLoPicker to Read Data

5.3.3.1 High-quality tumor samples

Because LoLoPicker, MuTect and VarScan2 showed better performances in calling low-fraction SNVs, we then applied them on a real cancer sample with matched blood sample from a glioblastoma (GBM) patient (GBM_9). About 500 germ-line samples were used as controls. In GBM_9, LoLoPicker called 60 somatic variants, while MuTect and VarScan2 called 182 variants and 503 variants, respectively (Figure 5.3).

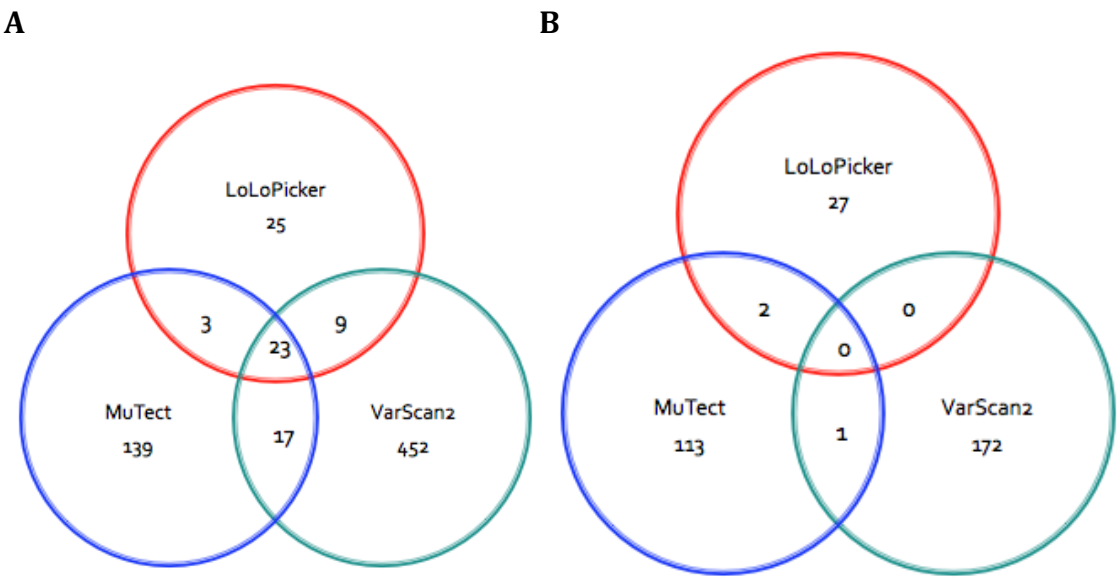


Figure 5.3: Venn diagram of called variants from different tools.
A) Number of SNVs called by LoLoPicker, MuTect and VarScan2. B) Number of SNVs with less than 10% allelic-fraction called by three callers.

Known GBM driving mutations were identified, including *TP53* (p.R174X), *H3F3A* (p.K28M), *ATRX* (p.R1480X), and *PIK3CA* (p.H1047R). *TP53* and *PIK3CA* mutations were present in both Catalogue of Somatic Mutations in Cancer (COSMIC) and dbSNP. LoLoPicker successfully identified all of them; whereas MuTect filtered out the *TP53* mutation because it found reads supporting the variant in the normal sample. However, those reads were overlapping pairs, thus resulting from a single DNA fragment (Figure 5.4). VarScan2 did not call the *PIK3CA* mutation as a high-confidence variant. In particular, the *PIK3CA* mutation showed low allelic-fraction at 6%. Again, this demonstrates that LoLoPicker has a high sensitivity of calling low-fraction SNVs.

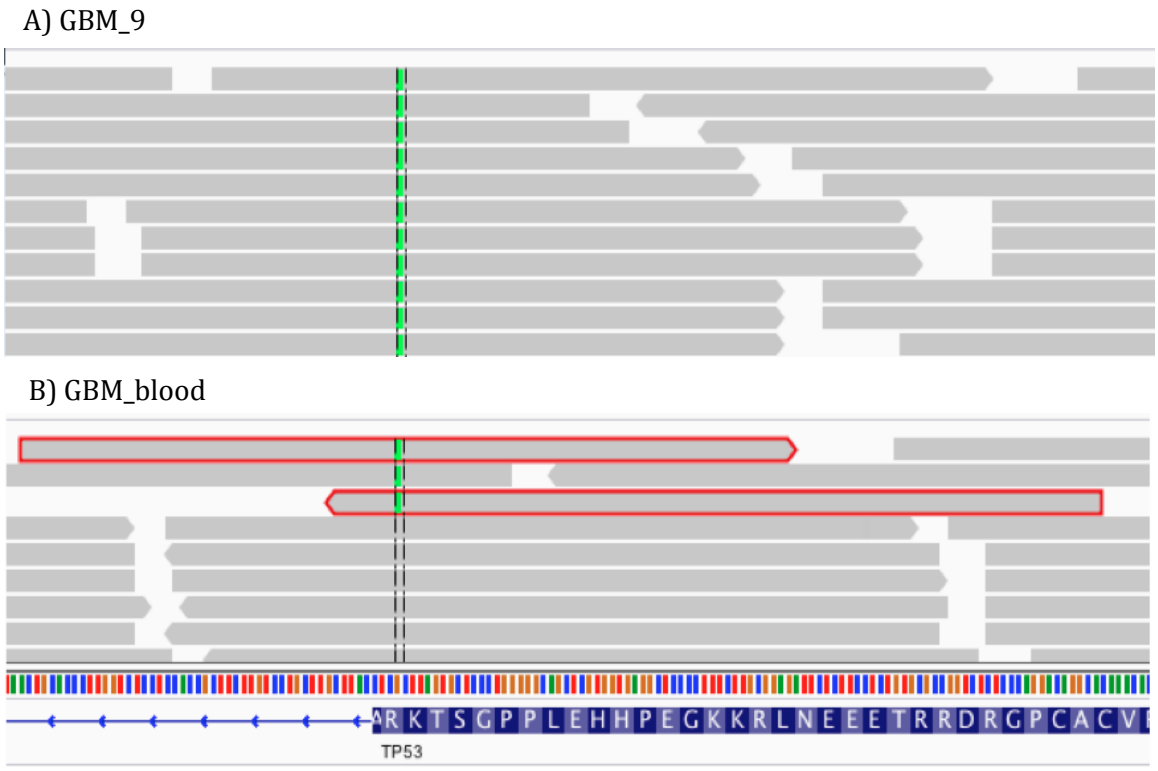


Figure 5.4: Snapshot of IGV showing the *TP53* mutation in GBM_9 and its matched blood sample.

MuTect filtered the *TP53* mutation (p.R174X) because three reads supporting the variant were observed in the matched normal. However, two of them were overlapping read-pairs (highlighted in red), which were counted as one in LoLoPicker. The third read had low mapping quality. Therefore, this mutation was retained in LoLoPicker.

Other somatic variants may exist in GBM_9. Variants reported in COSMIC were enriched among variants called by LoLoPicker (10%), compared with 2% among rejected calls (Figure 5.5). In variants called from MuTect and VarScan2, 7% and 2% were reported in COSMIC, respectively. MuTect also rejected 5% COSMIC reported variants. This suggested that LoLoPicker called more cancer-related SNVs.

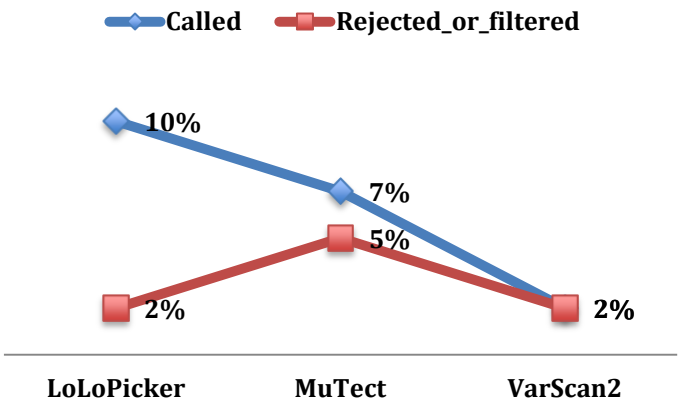


Figure 5.5: Proportion of variants reported in COSMIC among variants called and rejected or filtered by LoLoPicker, MuTect and VarScan2 in GBM_9.

LoLoPicker also identified 131 SNPs in GBM_9 from the control cohort, without comparing them to any public databases (Figure 5.6). All of the variants classified by LoLoPicker as “possible SNP” were reported in dbSNP, suggesting possibly missed germ-line variants in the normal tissue sequencing experiments or foreign DNA contamination.

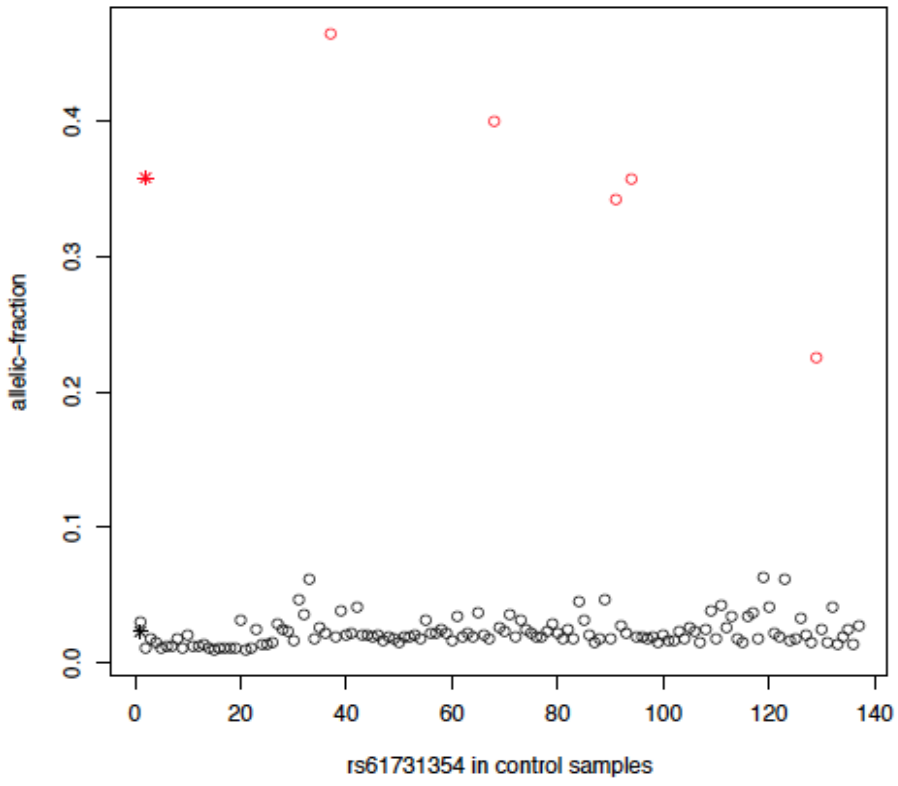


Figure 5.6: *K*-means clustering distinguishes true SNPs from background noise.

Identified true variants (red circles) and background noise (black circles) from 500 germ-line, non-cancer samples. At this specific site, rs61731354 was found

in five samples with mean allelic-fraction at 0.35 (red star). 130 samples showed background noise with mean allelic-fraction at 0.02 (black star).

Finally, 14 low-fraction SNVs in GBM_9 were selected for targeted re-sequencing validation. These variants were selected on the basis of being called by both LoLoPicker and MuTect, or by MuTect only. As the results, all the variants called by both tools were validated as true positives, whereas the ones that LoLoPicker rejected were not validated. These included four variants with higher coverage (>=5X) supporting the altered bases (Table 5.5). These results suggested that the specificity of LoLoPicker was improved without rejecting true positives as a trade-off.

Table 5.5: Low allelic-fraction SNVs either called by both MuTect and LoLoPicker, or called by MuTect but rejected by LoLoPicker were selected for MiSeq validation.

Position	Ref	Alt	Altered Reads	Total Reads	Allelic-fraction	Protein Change	Gene	LoLoPicker *p*_value	MuTect Judgement	MiSeq Validation
chr1:89521863	C	A	16	92	0.17	p.A402S	*GBP1*	5.30e-31	KEEP	Yes
chr13:24895566	T	C	5	47	0.11	p.I221T	*C1QTNF9*	9.66e-06	KEEP	Yes
chr3:178952085	A	G	5	94	0.06	p.H1047R	*PIK3CA*	2.62e-05	KEEP	Yes
chr9:5231708	G	A	4	22	0.17	p.G62E	*INSL4*	2.04e-04	KEEP	Yes
chr11:71907000	C	A	4	65	0.06	p.P185T	*FOLR1*	0.32	KEEP	No
chr13:25378544	C	A	5	75	0.07	p.P690T	*RNF17*	0.6	KEEP	No
chr19:40580859	C	A	4	68	0.06	p.G497V	*ZNF780A*	1	KEEP	No
chrX:3240813	C	A	3	31	0.09	p.E971D	*MXRA5*	1	KEEP	No
chr1:65301884	C	A	3	44	0.06	p.C1052F	*JAK1*	1	KEEP	No
chr2:62099221	C	A	3	29	0.1	p.R466L	*CCT4*	1	KEEP	No
chr15:22742690	T	C	8	121	0.07	p.W359R	*GOLGA6L1*	1	KEEP	No
chr3:4715013	A	C	7	23	0.3	p.T785P	*ITPR1*	1	KEEP	No
chr18:77805926	T	G	5	13	0.38	p.W240G	*RBFA*	1	KEEP	No
chr6:116912080	C	A	3	58	0.05	p.L193I	*RWDD1*	1	KEEP	No

5.3.3.2 FFPE samples

Error rates across different sites vary. Site-specific error rates in low-quality samples, such as FFPE samples are much higher than high-quality samples (Figure 5.7).

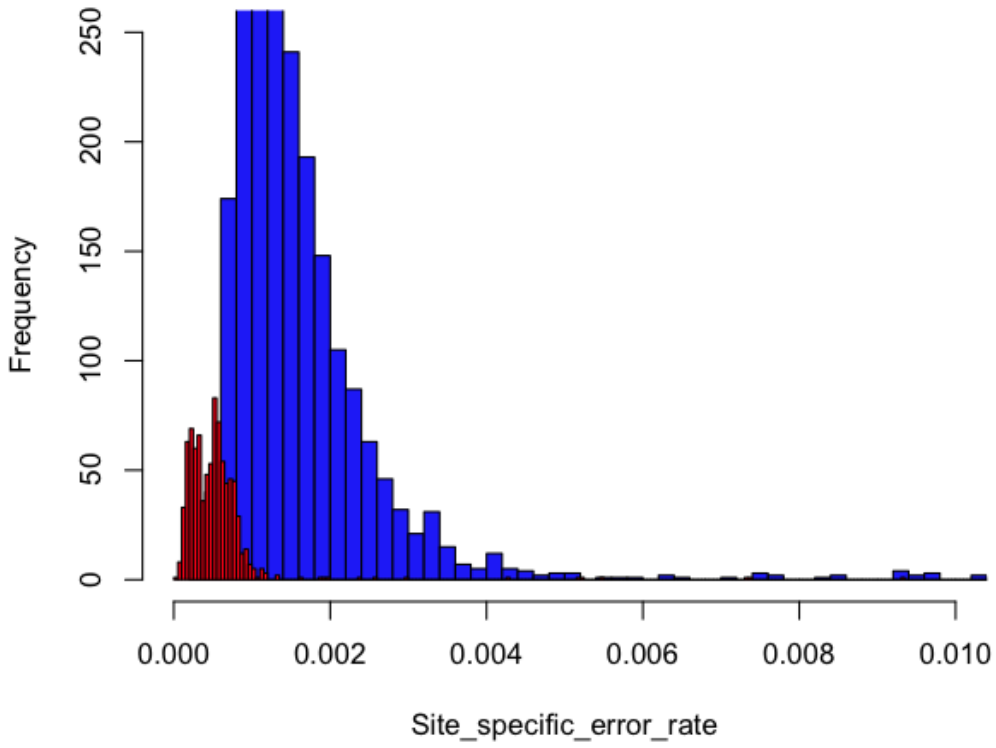


Figure 5.7: Frequency distribution of site-specific error rate.
Blue=FFPE sample; Red=germ-line sample.

In Chapter 3, I showed that no recurrent mutations, other than *SMARCA4* mutations were observed in SCCOHT tumors. Therefore, I tested LoLoPicker,

MuTect and VarScan2 using a FFPE sample of SCCOHT (UN5) carrying a somatic mutation in *SMARCA4* (c.2275-1G>T). Although I expected to see very few somatic mutations from this sample, both MuTect and VarScan2 called a large number of variants (483 and 143, respectively). Although a panel of normal FFPE samples was provided to MuTect, only 19 variants were filtered by the –normal_panel option. When using 500 germ-line samples as my controls, LoLoPicker called 92 variants. However, when 35 FFPE samples from normal tissues were used, only 18 variants were called, and most of the LoLoPicker rejected calls were C to T or G to A transitions known to be induced by the FFPE protocol, suggesting the necessity of providing a control cohort to further reduce false positive calls related to batch effects, especially FFPE-specific artifacts (Figure 5.8).

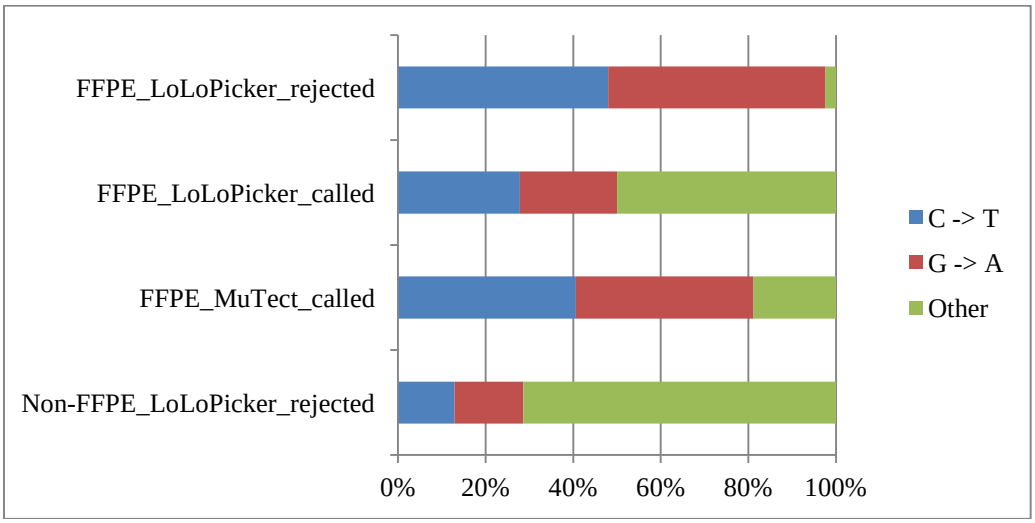


Figure 5.8: Percentage of base changes.

In a FFPE sample, C to T and G to A transitions, which are mostly likely FFPE-induced artifacts, are frequently observed among LoLoPicker rejected variants, and MuTect called variants. By contrast, these transitions are less frequent among LoLoPicker called variants. In a fresh-frozen sample, C to T and G to A transitions are less frequent among LoLoPicker rejected variants.

5.4 Discussion

LoLoPicker is a new algorithm designed to detect somatic SNVs, particularly tailored for low-fraction SNVs. I observed a superior performance of LoLoPicker in comparison with MuTect, VarScan2 and LoFreq. While LoLoPicker maintains highest sensitivity among other programs, the specificity of LoLoPicker is significantly improved, highlighting the importance of precisely measuring site-specific error rate from a larger number of control samples, rather than from a matched normal sample solely.

Samples provided as additional controls are essential in estimating the background error rate. Although I expect that LoLoPicker will handle WES data from any sequencing platforms and alignment methods, I suggest that samples processed in similar experimental protocols should be used. For example, having a panel of FFPE samples helped in filtering FFPE-specific artifacts. Compared to MuTect, which simply filtering out recurrent calls from the panel of normal, LoLoPicker's statistical framework identifies sites with high error-rate and retains sites with low-level artifacts, allowing high-accuracy.

Poor coverage uniformity has been a hurdle in precisely detecting variants in WES due to bias in exome-capture efficiency. The sensitivity of all callers was decreased when total coverage was reduced to less than 10X. Site-specific artifacts missed from the matched normal sample can be revealed using

additional control samples, giving the advantage for LoLoPicker to filter out most of the false positives. However, as the number of control samples available for research projects is usually limited, the final calls of LoLoPicker should be also filtered against other public databases to further identify possible SNPs.

Finally, the LoLoPicker algorithm can be easily parallelized to allow the analysis of WES data of one pair of tumor and normal sample against a larger number of control samples in a reasonable time. Our method will provide unprecedented information for analyzing FFPE samples and pave the way to apply WES into clinical testing.

Chapter 6. General Discussion

While sequencing an individual's whole genome remains expensive, WES that targets the protein-coding region of the genome has emerged as a cost-effective method to identify highly penetrant variants causing Mendelian or monogenic diseases. This thesis work first focused on finding the genetic causes in unexplained cancer families. Using WES as the primary investigative tool, I successfully identified the gene predisposing to SSCOHT gene – *SMARCA4* – as well as a new breast cancer gene – *RECQL* – both of which were subsequently found to play a role in non-inherited cases as well. These findings suggest that WES is a powerful tool to study the genetic basis underlying both rare and common cancers. Then, by performing WES on multiple tumor sets collected at different stages from a single patient carrying a germ-line *BRCA1* mutation, an *NF1* alteration was identified as a major contributor to cancer progression, suggesting that combined therapy targeting both the RAS signaling and the DNA repair pathway will benefit the patient's outcome.

Studying inherited variants is relatively easy, because heterozygous variants usually display allelic-fraction from 20% to 80% in germ-line samples. On the other hand, detecting somatic alterations from tumor samples is more challenging, because low-fraction variants are commonly observed when analyzing tumor samples, but these low-fraction variants are hard to distinguish

from artifacts. Therefore, an efficient tool to accurately identify low-fraction variants from WES data is warranted. By estimating the site-specific error rate from a project-specific control cohort, LoLoPicker significantly reduces false-positive rate in calling low-fraction SNVs. The performance of LoLoPicker is superior to other existing tools, especially in calling variants from low quality data, such as FFPE samples. Next, I will improve this algorithm for indel calling, using a similar strategy. For instance, observed indel artifacts from the controls will be characterized and used as a training dataset for a supervised learning algorithm to allow the identification of indels from the test sample.

6.1 *SMARCA4* and SCCOHT

By performing WES analysis of six individuals from three families with SCCOHT, I discovered segregating deleterious germ-line mutations in *SMARCA4* in all three families. Somatic mutation or loss of the wild-type allele was found in all the familial tumors harboring the germ-line *SMARCA4* mutation. By combining WES and Sanger sequencing of cases with available DNA, at least one deleterious *SMARCA4* mutation was identified in 30 of 32 cases. Genomic analysis of SCCOHT tumors and their matched normal tissues confirmed that *SMARCA4* is the only recurrently mutated gene. Moreover, recurrent allelic imbalance events were observed exclusively on chromosome 19p, where *SMARCA4* is located. Our findings suggest that SCCOHT is universally characterized by *SMARCA4* mutations.

Two different studies showed similar findings. In the first study, germ-line or somatic mutations in *SMARCA4* were found in 75% of the 12 cases they analyzed, and 82% of those cases showed loss of SMARCA4 protein (Ramos et al. 2014). In the second study, inactivating mutations in *SMARCA4* were found in all the 12 cases they analyzed (Jelinic et al. 2014). Taken together, clear evidence of the genetic cause of SCCOHT is provided. While the other two studies discovered *SMARCA4* mutations through WES of fresh-frozen tissues and targeted re-sequencing of a panel of 300 cancer-related genes, only our

work successfully applied WES to FFPE samples, which unlocked archived tissue for genome-wide analysis.

6.1.1 Is SCCOHT Rhabdoid Tumor of the Ovary?

Although named as small cell carcinoma, pathological examination observed large cells in approximately 50% of SCCOHT cases we studied. These large cells display a ‘rhabdoid’ features, with eccentric nuclei, prominent nucleoli and abundant glassy eosinophilic cytoplasm (Young, Oliva & Scully 1994). Defects in the SWI/SNF complexes have been observed in rhabdoid tumors and lung cancers (Matsubara et al. 2013). Notably, hypercalcemia has been clinically linked to SCCOHT, rhabdoid tumors and lung tumors (Brennan, Stiller & Bourdeaut 2013, Hsu et al. 2011). In essence, germ-line mutations in *SMARCA4* are also known to rarely predispose ATRTs (Schneppenheim et al. 2010, Hasselblatt et al. 2011, Witkowski et al. 2013). Thus, it is reasonable to speculate that SCCOHT falls within the category of rhabdoid tumors.

6.1.2 Molecular Analysis Revealed Close Similarities between SCCOHT and ATRT

Clinical findings are mirrored by genome-wide sequencing analysis, which revealed that germ-line or somatic mutations in *SMARCA4* and recurrent allelic imbalance on chromosome 19p, where *SMARCA4* resides, exclusively

characterized SCCOHT. By comparing SCCOHT, ATRT and HGSC, remarkably simple genomes in SCCOHT and ATRT were shown. Interestingly, both SCCOHT and ATRT harbor much fewer somatic mutations and chromosomal alterations than HGSC. Furthermore, our recent study comparing the global DNA methylation profiles of SCCOHT, ATRT, and HGSC demonstrates a strong epigenetic correlation between SCCOHT and ATRT, suggesting that similar mechanisms linked to the defects in the SWI/SNF complexes might contribute to similar methylation alterations in SCCOHT and ATRT (Fahiminiya et al. 2015).

6.1.3 WES Opened New Avenues for SCCOHT Treatment

In this study, genetic findings from WES pinpointed to *SMARCA4* mutations as the potential therapeutic target. Although an approach based on the inhibition of the *SMARCA4* counterpart *SMARCA2* has been developed to treat *SMARCA4*-deficient cancers, recent work suggested that SMARCA2 protein is absent in SCCOHT (Oike et al. 2013, Jelinic et al. 2016). Although no inactivating mutations in *SMARCA*2 found in any of our SCCOHT cases, a epigenetic mechanism may underlie the *SMARCA2* silencing.

Meanwhile, new treatments targeting SMARCB1 loss have been developed with active responses in ATRT patients (Smith et al. 2011, Wetmore, Bendel & Gajjar 2014). Given the molecular and clinical similarities between SCCOHT and ATRT, similar method used for rhabdoid tumor treatment might help to improve the

outcome of SCCOHT. Finally, this study extended the application of WES from gene discovery to assistance in pathological analysis, inferring the role of WES in guiding decisions about treatment in future.

6.2 New Breast Cancer Genes

6.2.1 *RECQL* and Breast Cancer

WES study of 51 French-Canadian breast cancer families revealed *RECQL* as a new breast cancer susceptibility gene, with protein-damaging mutations associated with a remarkably increased risk for breast cancer. The identification of different truncating mutations in *RECQL* from WES and recurrent mutations in additional cases from Sanger sequencing prompted the in-depth investigation of the *RECQL* mutations in large samples of breast cancer cases and controls in two specific populations, the French-Canadian and the Polish population.

In addition, a recent study identified *RECQL* mutations in association with breast cancer in the Chinese population (Sun et al. 2015). Functional analysis suggested that missense mutations identified from breast cancer cases disrupted the helicase activity, further supporting the pathogenic role of *RECQL* mutations in breast cancer (Sun et al. 2015). Given that the frequency of *RECQL* mutations in breast cancer patients is relatively low, founder population-based WES study may serve as an effective approach in the search for other rarely mutated cancer susceptibility genes. Although no obvious founder mutation was observed in *RECQL* using WES, other missense mutations identified through this cohort are more likely to represent founder mutations in the French-Canadian population, including the *SMC4* variant.

6.2.2 *SMC4* and Breast Cancer

The *SMC4* mutation (p.R827C) was originally identified in two cases through WES analysis of 51 French-Canadian breast cancer cases. Genotyping of this variant in a large number of breast cancer cases and population-specific controls demonstrated that the *SMC4* mutation was strongly associated with breast cancer. Furthermore, the frequency of this variant in healthy controls is increased from 0.00005 in unselected population to 0.001552 in the French-Canadian population, suggesting the underlying driving force due to the founder effect. *SMC*4 is the core of the condensin complex, which is essential for chromosome assembly and segregation (Losada, Hirano 2005). To establish the pathogenicity of the *SMC4* mutation (p.R827C) in breast cancer, further work is required to assess its functional significance. As an example, cells with the p.R827C mutation may be characterized in cell cycle and chromosome segregation. WES study of the *SMC4* carrier's tumor will be probably informative as well – a "second hit" in the tumor will provide further evidence for the role of this gene in the etiology of the breast cancer.

6.2.3 Other Candidates

Previous work on *PALB2* suggested that breast cancer associated founder mutations can present in low frequency in the French-Canadian population

(Foulkes et al. 2007). While WES analysis of 51 French-Canadian breast cancer families has provided valuable leads, other breast cancer alleles could have been missed. One example is the frameshift mutation in *CHEK2* (p.D82fs), which was observed only once in the 51 cases. This mutation would not be selected based on the current filtering criteria, if no previous knowledge about *CHEK2* in breast cancer were known. Thus, sequencing additional families will add confidence that our prioritization method has not missed a rare but important French-Canadian founder mutation.

6.2.4 Limitations of This Study

Although many other genes identified in this study are potentially interesting and probably worth pursuing – examples including *SMPD1* and *CNTN4*, both of which were significantly mutated in the case cohort (two truncating mutations in different families) and have implications in cancer – some families will remain unexplained. It is likely that some causal CNVs were missed from the discovery phase, because of the difficulties in distinguishing true CNVs from background noise created by WES. Furthermore, more and more evidence suggest that germ-line variants in transcript regulation are associated with breast cancer risk (Li et al. 2013, Glubb et al. 2015, Castro et al. 2016). Regulatory variants for breast cancer will not be revealed by WES studies. Therefore, other methods will be applied to discover new breast cancer susceptibility genes. For example, WGS of the germ-line sample in combination with RNA-seq of the tumor and

matched normal tissue will be informative to identify inherited, causal variants missed by WES. Finally, mutation discovery using sequencing technologies is only the first step. Proving that a candidate variant truly is disease-causing can be a large challenge, especially for missense mutations in functionally unknown genes when choosing a relevant functional assay is not obvious. Therefore, combining bioinformatics analysis and functional screening will be essential for future work.

6.3 Current Issues in Studying Somatic Alterations

Unraveling somatic alterations is crucial for cancer studies. Presently, the need in cancer genomics is shifting from data generation to data analysis, which accelerated the development of numerous computational tools to precisely identify tumor-only events. Point mutations are most frequently observed in cancer genomes. Algorithms of calling somatic SNVs evolved from simply subtracting germ-line variants, to jointly calling tumor and matched normal samples, to integrating clonal information into the models (Ding et al. 2014). Although several tools have been developed to improve the detection of somatic SNVs (e.g., MuTect, VarScan2, LoLoPicker), indel detection remains challenging. One of the major barriers is that current alignment methods are not optimized for mapping short reads containing indels to the reference genome. To tackle this issue, local re-alignment algorithms (e.g., GATK IndelRealigner) have been applied to improve the mapping surrounding potential indels, allowing improved indel identification using standard variant callers. However, novel algorithms are needed to identify somatic indels, especially to distinguish low allelic-fraction indels from artifacts.

Genetic factors contributing to cancer progression have been identified, such as *NT5C2* mutations in leukemia relapse, SERPINE2 overexpression in breast cancer metastasis, and activation of the RAS-PI3K pathway in GBM progression (Ding et al. 2012, Ma et al. 2015, Bai et al. 2016). In HGSC, inactivating

mutations in *RB1*, *NF1*, *RAD51B* and *PTEN* might contribute to the disease relapse after treatment (Zhang et al. 2013, Schwarz et al. 2015, Patch et al. 2015). Nowadays, sequencing tumor samples from multiple sites, multiple stages, or multiple tumor sections has become common in cancer research. More and more subclonal variants related to tumor progression will be identified using different sequencing strategies. On the other hand, somatic mutations in hematopoietic stem cells might be misunderstood as germ-line variants from deep sequencing of whole blood-derived DNA (Kurek et al. 2012). Advance in single cell sequencing holds great potential for studying tumor evolution, where subclonal information may be missed from sequencing mixed cell populations (Navin et al. 2011). The bottleneck is the development of computational tools to reconstruct the genotypes of each tumor subpopulations. Current models in subclonal reconstruction are based on the measurement of allelic-fraction of the somatic mutations, and therefore, accurate detection and quantification of somatic events, including point mutations and CNVs, is the cornerstone (Carter et al. 2012, Larson, Fridley 2013, Deshwar et al. 2015).

6.4 Future Directions

6.4.1 WES vs. WGS and RNA-seq

While WES is rapidly entering research and clinical laboratories as a routinely used investigative and diagnostic tool, the cost of sequencing the whole genome is decreasing, and therefore, WGS is becoming increasingly attractive. Now mapping the epigenome is completed, and regulatory elements have been better characterized than ever (The ENCODE Project Consortium 2012). Sequencing the whole genome enables us to identify causal variants in regulatory regions. Another advantage of using WGS instead of WES is that WGS is more reliable for calling of structural variants. Current methods of CNV detection from NGS data are based on four sources of information: abnormally mapped read pairs, split reads that span breakpoints, read-depth and unmapped reads (Alkan, Coe & Eichler 2011). Because the majority of structural breakpoints lie outside the exonic regions, the broader coverage of WGS will provide more evidence for CNV calling.

Perhaps RNA-seq, which has been fruitfully applied in cancer studies, will become more and more popular. Compared to WGS, RNA-seq costs less and requires less computational time. Although variant calling from RNA-seq data remains challenging, (e.g., variants nearby the splicing junctions), recent improvements in mapping RNA-seq reads and variant filtering strategies have

enabled accurate identification of disease-associated variants from RNA-seq (Piskol, Ramaswami & Li 2013). Additionally, RNA-seq is capable of addressing a broader range of important questions in cancer, such as measurement of gene expression, detection of novel transcript, gene fusion, alternative splicing, allelic-specific expression, or RNA-editing (Costa et al. 2013). Therefore, RNA-seq should be considered as a complementary technology to DNA sequencing in cancer research.

6.4.2 More Samples, More Opportunities

Current studies in cancer genomics are limited by the availability of fresh tumor samples, however, FFPE samples are used routinely for diagnosis. Our work successfully produced reliable results from FFPE-WES data. Recent work also showed the feasibility of performing RNA-seq analysis on FFPE samples (Adiconis et al. 2013, Majewski et al. 2013, Graw et al. 2015). These breakthroughs will enable the set-up of large-scale cancer research and significantly benefit hereditary cancer research. As an example, combining WES analysis of germ-line DNA with WES analysis of FFPE tissue from the same patient will increase our ability to distinguish potentially disease-causing variants from variants unrelated to the disease. In our study of identifying novel breast cancer variants, WES data from tumor samples will be particularly helpful in prioritizing genes for further study: a non-recurrent mutation with a potentially

interesting function related to cancer will be prioritized, if persuasive evidence is found in the tumor (e.g., second hit).

6.4.3 Cancer Research: Entering the Era of Epigenetics

In this thesis, inactivating *SMARCA4* mutations are identified as the major driver in SCCOHT. SCCOHT is a devastating disease, yet is characterized by very little genomic alterations, suggesting the key role of epigenetic disruption in the tumorigenesis of this disease. In fact, genetic mutations resulting in epigenetic regulation alterations have been linked to different types of cancers (Baylin and Jones 2011). Besides the SIW/SNF chromatin remodeling complex, somatic mutations have been found in DNA methyltransferase gene *DNMT3A* in acute myeloid leukemia, which altered DNA methylation activity and expression of genes including *IDH1* (Yan et al. 2011). Inactivating mutations in *SETD2*, which is responsible for trimethylation of the histone mark H3K36, have been found in renal cell carcinoma (Duns et al. 2010). Mutations in histone gene *H3F3A* have been recurrently found in pediatric GBM (Schwartzentruber et al. 2012). Global DNA methylation profiles have shown distinct epigenetic phenotypes driven by these mutations (Fahiminiya et al. 2015, Fontebasso et al. 2014). However, how genetic alterations interact with epigenetic regulations and lead to cancer development remains unclear.

Next generation sequencing technologies facilitate the understanding of the epigenome. For instance, bisulfite-treated DNA coupled with high-throughput sequencing enables genome-wide profiling of the methylation state on single-nucleotide level (Cokus et al. 2008). Chromatin immunoprecipitation-sequencing (ChIP-seq) enables the examination of the DNA-protein interactions throughout the genomes (Robertson et al. 2007). Concurrently, the demand of identifying DNA-binding sites from ChIP-seq data motivates the development of computational tools. In cancer research, methods that identify differential signals between tumor and matched normal tissue will be particularly useful. Finally, high-throughput technologies in the proteomic analysis will one day, unveil all the protein-protein interactions, and ultimately translate our knowledge in cancer genomics into diagnosis, prognosis and personalized medicine.

Reference

Adiconis, X., Borges-Rivera, D., Satija, R., DeLuca, D.S., Busby, M.A., Berlin, A.M., Sivachenko, A., Thompson, D.A., Wysoker, A. & Fennell, T. 2013, "Comparative analysis of RNA sequencing methods for degraded or low-input samples", *Nature methods,* vol. 10, no. 7, pp. 623-629.

Adzhubei, I., Jordan, D.M. & Sunyaev, S.R. 2013, "Predicting functional effect of human missense mutations using PolyPhen‐2", *Current protocols in human genetics,* pp. 7-20.

Alkan, C., Coe, B.P. & Eichler, E.E. 2011, "Genome structural variation discovery and genotyping", *Nature reviews Genetics,* vol. 12, no. 5, pp. 363-376.

Amberger J, Bocchini C, Hamosh A 2011, “A new face and new challenges for Online Mendelian Inheritance in Man (OMIM(R))”, *Human Mutation* 2011, vol. 32, pp. 564–567.

Arcand, S., Maugard, C., Ghadirian, P., Robidoux, A., Perret, C., Zhang, P., Fafard, E., Mes-Masson, A., Foulkes, W., Provencher, D., Narod, S. & Tonin, P. 2008, "Germline TP53 mutations in BRCA1 and BRCA2 mutation-negative French Canadian breast cancer families", *Breast cancer research and treatment,* vol. 108, no. 3, pp. 399-408.

Armitage, P. & Doll, R. 1957, "A two-stage theory of carcinogenesis in relation to the age distribution of human cancer", *British journal of cancer,* vol. 11, no. 2, pp. 161-169.

Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy- Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J. and Banks, E., 2013, “From FastQ data to high- confidence variant calls: the genome analysis toolkit best practices pipeline”, *Current Protocols in Bioinformatics*, pp.11-10.

Awadalla, M.S., Burdon, K.P., Souzeau, E. & al, e. 2014, "Mutation in TMEM98 in a large white kindred with autosomal dominant nanophthalmos linked to 17p12-q12", *JAMA Ophthalmology,* vol. 132, no. 8, pp. 970-977.

Badve, S., Dabbs, D.J., Schnitt, S.J., Baehner, F.L., Decker, T., Eusebi, V., Fox, S.B., Ichihara, S., Jacquemier, J., Lakhani, S.R., Palacios, J., Rakha, E.A., Richardson, A.L., Schmitt, F.C., Tan, P., Tse, G.M., Weigelt, B., Ellis, I.O. & Reis-Filho, J. 2011, "Basal-like and triple-negative breast cancers: a critical review with an emphasis on the implications for pathologists and oncologists", *Modern pathology,* vol. 24, no. 2, pp. 157-167.

Bai, H., Harmanci, A.S., Erson-Omay, E., Li, J., Coskun, S., Simon, M., Krischek, B., Ozduman, K., Omay, S.B., Sorensen, E.A., Turcan, S., Bakirciglu, M., Carrion-Grant, G., Murray, P.B., Clark, V.E., Ercan-Sencicek, A., Knight, J., Sencar, L., Altinok, S., Kaulen, L.D., Gulez, B., Timmer, M., Schramm, J., Mishra-Gorur, K., Henegariu, O., Moliterno, J., Louvi, A., Chan, T.A., Tannheimer, S.L., Pamir, M.N., Vortmeyer, A.O., Bilguvar, K., Yasuno, K. &

Gunel, M. 2016, "Integrated genomic characterization of IDH1-mutant glioma malignant progression", *Nature genetics,* vol. 48, no. 1, pp. 59-66.

Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A. & Shendure, J. 2011, "Exome sequencing as a tool for Mendelian disease gene discovery", *Nature reviews Genetics,* vol. 12, no. 11, pp. 745-755.

Baylin, S.B. and Jones, P.A., 2011, “A decade of exploring the cancer epigenome – biological and translational implications”, *Nature Reviews Cancer*, vol. 12, no.11, pp.726-734.

Behlouli, A., Bonnet, C., Abdi, S., Bouaita, A., Lelli, A., Hardelin, J., Schietroma, C., Rous, Y., Louha, M. & Cheknane, A. 2014, "EPS8, encoding an actin-binding protein of cochlear hair cell stereocilia, is a new causal gene for autosomal recessive profound deafness", *Orphanet Journal of Rare Diseases,* vol. 9, no. 1, pp. 55.

Berchuck, A., Witkowski, L., Hasselblatt, M. & Foulkes, W.D. 2015, "Prophylactic oophorectomy for hereditary small cell carcinoma of the ovary, hypercalcemic type", *Gynecologic Oncology Reports,* vol. 12, pp. 20-22.

Berti, M., Chaudhuri, A.R., Thangavel, S., Gomathinayagam, S., Kenig, S., Vujanovic, M., Odreman, F., Glatter, T., Graziano, S. & Mendoza-Maldonado, R. 2013, "Human RECQ1 promotes restart of replication forks reversed by DNA topoisomerase I inhibition", *Nature structural & molecular biology,* vol. 20, no. 3, pp. 347-354.

Bolton, K.L., Chenevix- Trench, G., Goh, C., Sadetzki, S., Ramus, S.J., Karlan, B.Y., Lambrechts, D., Despierre, E., Barrowdale, D., McGuffog, L., Healey, S., Easton, D.F., Sinilnikova, O., BenAtez, J., GarcAa, M.J., Neuhausen, S., Gail, M.H., Hartge, P., Peock, S., Frost, D., Evans, D.G., Eeles, R., Godwin, A.K., Daly, M.B., Kwong, A., Ma, E., LAzaro, C., Blanco, I. & Montagna, M. 2012, "Association between BRCA1 and BRCA2 mutations and survival in women with invasive epithelial ovarian cancer", *JAMA,* vol. 307, no. 4, pp. 382-389.

Bray, F., Ren, J., Masuyer, E. & Ferlay, J. 2013, "Global estimates of cancer prevalence for 27 sites in the adult population in 2008", *International Journal of Cancer,* vol. 132, no. 5, pp. 1133-1145.

Brennan, B., Stiller, C. & Bourdeaut, F. 2013, "Extracranial rhabdoid tumours: what we have learned so far and future directions", *The Lancet Oncology,* vol. 14, no. 8, pp. e329-e336.

Caburet, S., Anttonen, M., Todeschini, A., Unkila-Kallio, L., Mestivier, D., Butzow, R. & Veitia, R.A. 2015, "Combined comparative genomic hybridization and transcriptomic analyses of ovarian granulosa cell tumors point to novel candidate driver genes", *BMC Cancer,* vol. 15, pp. 251.

Candido-dos-Reis, F.J., Song, H., Goode, E.L., Cunningham, J.M., Fridley, B.L., Larson, M.C., Alsop, K., Dicks, E., Harrington, P., Ramus, S.J. and de Fazio, A., 2015, “Germline mutation in BRCA1 or BRCA2 and ten-year survival for women diagnosed with epithelial ovarian cancer”, *Clinical Cancer Research*, vol. 21, no. 3, pp.652-657.

Carter, S.L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P.W., Onofrio, R.C., Winckler, W. & Weir, B.A. 2012, "Absolute quantification of somatic DNA alterations in human cancer", *Nature biotechnology,* vol. 30, no. 5, pp. 413-421.

Castro, M.A.A., de Santiago, I., Campbell, T.M., Vaughn, C., Hickey, T.E., Ross, E., Tilley, W.D., Markowetz, F., Ponder, B.A.J. & Meyer, K.B. 2016, "Regulators of genetic risk of breast cancer identified by integrative network analysis", *Nature genetics,* vol. 48, no. 1, pp. 12-21.

Cavallone, L., Arcand, S., Maugard, C., Nolet, S., Gaboury, L., Mes-Masson, A., Ghadirian, P., Provencher, D. & Tonin, P. 2010, "Comprehensive BRCA1 and BRCA2 mutation analyses and review of French Canadian families with at least three cases of breast cancer", *Familial Cancer,* vol. 9, no. 4, pp. 507-517.

Cavenee, W.K., Dryja, T.P., Phillips, R.A., Benedict, W.F., Godbout, R., Gallie, B.L., Murphree, A.L., Strong, L.C. & White, R.L. 1983, "Expression of recessive alleles by chromosomal mechanisms in retinoblastoma", *Nature,* vol. 305, no. 5937, pp. 779-784.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A.P., Sander, C. & Schultz, N. 2012, "The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data", *Cancer Discovery,* vol. 2, no. 5, pp. 401-404.

Cheng, L., Roth, L.M., Zhang, S., Wang, M., Morton, M.J., Zheng, W., Abdul Karim, F.W., Montironi, R. & Lopez-Beltran, A. 2011, "KIT gene mutation and amplification in dysgerminoma of the ovary", *Cancer,* vol. 117, no. 10, pp. 2096-2103.

Choi, M., Scholl, U.I., Ji, W., Liu, T., Tikhonova, I.R., Zumbo, P., Nayir, A., Bakkaloğlu, A., Özen, S., Sanjad, S., Nelson-Williams, C., Farhi, A., Mane, S. & Lifton, R.P. 2009, "Genetic diagnosis by whole exome capture and massively parallel DNA sequencing", *Proceedings of the National Academy of Sciences,* vol. 106, no. 45, pp. 19096-19101.

Chu, W.K. & Hickson, I.D. 2009, "RecQ helicases: multifunctional genome caretakers", *Nature reviews Cancer,* vol. 9, no. 9, pp. 644-654.

Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S. & Getz, G. 2013, "Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples", *Nature Biotechnology,* vol. 31, no. 3, pp. 213-219.

Clement, P.B. 2005, "Selected miscellaneous ovarian lesions: small cell carcinomas, mesothelial lesions, mesenchymal and mixed neoplasms, and non-neoplastic lesions", *Modern pathology,* vol. 18, pp. S113-S129.

Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M. & Jacobsen, S.E. 2008, "Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning", *Nature,* vol. 452, no. 7184, pp. 215-219.

Collins, F.S. 1996, "BRCA1 — lots of mutations, lots of dilemmas", *New England Journal of Medicine,* vol. 334, no. 3, pp. 186-188.

Costa, V., Aprile, M., Esposito, R. & Ciccodicola, A. 2013, "RNA-Seq and human complex diseases: recent accomplishments and future perspectives", *European Journal of Human Genetics,* vol. 21, no. 2, pp. 134-142.

Costello, M., Pugh, T.J., Fennell, T.J., Stewart, C., Lichtenstein, L., Meldrim, J.C., Fostel, J.L., Friedrich, D.C., Perrin, D., Dionne, D., Kim, S., Gabriel, S.B., Lander, E.S., Fisher, S. & Getz, G. 2013, "Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation", *Nucleic acids research,* vol. 41, no. 6, pp. e67-e67.

Couch, F.J., Nathanson, K.L. & Offit, K. 2014, "Two decades after BRCA: setting paradigms in personalized cancer care and prevention", *Science,* vol. 343, no. 6178, pp. 1466-1470.

Cybulski, C., Carrot-Zhang, J., Kluzniak, W., Rivera, B., Kashyap, A., Wokolorczyk, D., Giroux, S., Nadaf, J., Hamel, N., Zhang, S., Huzarski, T., Gronwald, J., Byrski, T., Szwiec, M., Jakubowska, A., Rudnicka, H., Lener, M., Masojc, B., Tonin, P.N., Rousseau, F., Gorski, B., Debniak, T., Majewski, J., Lubinski, J., Foulkes, W.D., Narod, S.A. & Akbari, M.R. 2015, "Germline RECQL mutations are associated with breast cancer susceptibility", *Nature genetics,* vol. 47, no. 6, pp. 643-646.

DePristo, M.A., Banks, E., Poplin, R.E., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D. & Daly, M.J. 2011, "A framework for variation discovery and genotyping using next-generation DNA sequencing data", *Nature genetics,* vol. 43, no. 5, pp. 491-498.

Deshwar, A.G., Vembu, S., Yung, C.K., Jang, G.H., Stein, L. & Morris, Q. 2015, "PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors", *Genome biology,* vol. 16, pp. 35.

Ding, L., Ellis, M.J., Li, S., Larson, D.E., Chen, K., Wallis, J.W., Harris, C.C., McLellan, M.D., Fulton, R.S., Fulton, L.L., Abbott, R.M., Hoog, J., Dooling, D.J., Koboldt, D.C., Schmidt, H., Kalicki, J., Zhang, Q., Chen, L., Lin, L., Wendl, M.C., McMichael, J.F., Magrini, V.J., Cook, L., McGrath, S.D., Vickery, T.L., Appelbaum, E., DeSchryver, K., Davies, S., Guintoli, T., Lin, L., Crowder, R., Tao, Y., Snider, J.E., Smith, S.M., Dukes, A.F., Sanderson, G.E., Pohl, C.S., Delehaunty, K.D., Fronick, C.C., Pape, K.A., Reed, J.S., Robinson, J.S., Hodges, J.S., Schierding, W., Dees, N.D., Shen, D., Locke, D.P., Wiechert, M.E., Eldred, J.M., Peck, J.B., Oberkfell, B.J., Lolofie, J.T., Du, F., Hawkins, A.E., O'Laughlin, M.,D., Bernard, K.E., Cunningham, M., Elliott, G., Mason, M.D., Thompson, D.M., Ivanovich, J.L., Goodfellow, P.J., Perou, C.M., Weinstock, G.M., Aft, R., Watson, M., Ley, T.J., Wilson, R.K. & Mardis, E.R. 2010, "Genome remodeling in a basal-like breast cancer metastasis and xenograft", *Nature,* vol. 464, no. 7291, pp. 999-1005.

Ding, L., Ley, T.J., Larson, D.E., Miller, C.A., Koboldt, D.C., Welch, J.S., Ritchey, J.K., Young, M.A., Lamprecht, T., McLellan, M.D., McMichael, J.F., Wallis, J.W., Lu, C., Shen, D., Harris, C.C., Dooling, D.J., Fulton, R.S., Fulton, L.L., Chen, K., Schmidt, H., Kalicki-Veizer, J., Magrini, V.J., Cook, L., McGrath,

S.D., Vickery, T.L., Wendl, M.C., Heath, S., Watson, M.A., Link, D.C., Tomasson, M.H., Shannon, W.D., Payton, J.E., Kulkarni, S., Westervelt, P., Walter, M.J., Graubert, T.A., Mardis, E.R., Wilson, R.K. & DiPersio, J.F. 2012, "Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing", *Nature,* vol. 481, no. 7382, pp. 506-510.

Ding, L., Wendl, M.C., McMichael, J.F. & Raphael, B.J. 2014, "Expanding the computational toolbox for mining cancer genomes", *Nature reviews Genetics,* vol. 15, no. 8, pp. 556-570.

Domchek, S.M., Friebel, T.M., Singer, C.F., Evans, D.G., Lynch, H.T., Isaacs, C., Garber, J.E., Neuhausen, S.L., Matloff, E., Eeles, R., Pichert, G., Van t'veer, L., Tung, N., Weitzel, J.N., Couch, F.J., Rubinstein, W.S., Ganz, P.A., Daly, M.B., Olopade, O.I., Tomlinson, G., Schildkraut, J., Blum, J.L. & Rebbeck, T.R. 2010, "Association of risk-reducing surgery in BRCA1 or BRCA2 mutation carriers with cancer risk and mortality", *JAMA,* vol. 304, no. 9, pp. 967-975.

Downward, J. 2003, "Targeting RAS signalling pathways in cancer therapy", *Nature reviews Cancer,* vol. 3, no. 1, pp. 11-22.

Duns, G., van den Berg, E., van Duivenbode, I., Osinga, J., Hollema, H., Hofstra, R.M.W. & Kok, K. 2010, "Histone Methyltransferase Gene SETD2 Is a Novel Tumor Suppressor Gene in Clear Cell Renal Cell Carcinoma", *Cancer research,* vol. 70, no. 11, pp. 4287-4291.

Estel, R., Hackethal, A., Kalder, M. & Münstedt, K. 2011, "Small cell carcinoma of the ovary of the hypercalcaemic type: an analysis of clinical and prognostic aspects of a rare disease on the basis of cases published in the literature", *Archives of Gynecology and Obstetrics,* vol. 284, no. 5, pp. 1277-1282.

Fahiminiya, S., Witkowski, L., Nadaf, J., Carrot-Zhang, J., Goudie, C., Hasselblatt, M., Johann, P., Kool, M., Lee, R.S., Gayden, T., Roberts, C.W., Biegel, J.A., Jabado, N., Majewski, J. & Foulkes, W.D. 2015, "Molecular analyses reveal close similarities between small cell carcinoma of the ovary, hypercalcemic type and atypical teratoid/rhabdoid tumor", *Oncotarget*. Doi:10.18632/oncotarget.6459

Flicek, P. & Birney, E. 2009, "Sense from sequence reads: methods for alignment and assembly", *Nature methods,* vol. 6, no. 11, pp. S6-S12.

Flickinger, M., Jun, G., Abecasis, G., Boehnke, M. & Kang, H. 2015, "Correcting for sample contamination in genotype calling of DNA sequence data", *The American Journal of Human Genetics,* vol. 97, no. 2, pp. 284-290.

Fontebasso, A.M., Papillon-Cavanagh, S., Schwartzentruber, J., Nikbakht, H., Gerges, N., Fiset, P., Bechet, D., Faury, D., De Jay, N., Ramkissoon, L.A., Corcoran, A., Jones, D.T.W., Sturm, D., Johann, P., Tomita, T., Goldman, S., Nagib, M., Bendel, A., Goumnerova, L., Bowers, D.C., Leonard, J.R., Rubin, J.B., Alden, T., Browd, S., Geyer, J.R., Leary, S., Jallo, G., Cohen, K., Gupta, N., Prados, M.D., Carret, A., Ellezam, B., Crevier, L., Klekner, A., Bognar, L., Hauser, P., Garami, M., Myseros, J., Dong, Z., Siegel, P.M., Malkin, H., Ligon, A.H., Albrecht, S., Pfister, S.M., Ligon, K.L., Majewski, J., Jabado, N. & Kieran, M.W. 2014, "Recurrent somatic mutations in ACVR1 in pediatric

midline high-grade astrocytoma", *Nature genetics,* vol. 46, no. 5, pp. 462-466.

Ford, D., Easton, D.F., Stratton, M., Narod, S., Goldgar, D., Devilee, P., Bishop, D.T., Weber, B., Lenoir, G., Chang-Claude, J., Sobol, H., Teare, M.D., Struewing, J., Arason, A., Scherneck, S., Peto, J., Rebbeck, T.R., Tonin, P., Neuhausen, S., Barkardottir, R., Eyfjord, J., Lynch, H., Ponder, B.A.J., Gayther, S.A., Birch, J.M., Lindblom, A., Stoppa-Lyonnet, D., Bignon, Y., Borg, A., Hamann, U., Haites, N., Scott, R.J., Maugard, C.M., Vasen, H., Seitz, S., Cannon-Albright, L., Schofield, A. & Zelada-Hedman, M. 1998, "Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families", *The American Journal of Human Genetics,* vol. 62, no. 3, pp. 676-689.

Foulkes, W.D., Ghadirian, P., Akbari, M.R., Hamel, N., Giroux, S., Sabbaghian, N., Darnel, A., Royer, R., Poll, A. & Fafard, E. 2007, "Identification of a novel truncating PALB2 mutation and analysis of its contribution to early-onset breast cancer in French-Canadian women", *Breast Cancer Research,* vol. 9, no. 6, pp. R83.

Foulkes, W.D. 2008, "Inherited Susceptibility to Common Cancers", *New England Journal of Medicine,* vol. 359, no. 20, pp. 2143-2153.

Foulkes, W.D., Ghadirian, P., Akbari, M., Hamel, N., Giroux, S., Sabbaghian, N., Darnel, A., Royer, R., Poll, A., Fafard, E., Robidoux, A., Martin, G., Bismar, T., Tischkowitz, M., Rousseau, F. & Narod, S. 2007, "Identification of a novel truncating PALB2 mutation and analysis of its contribution to early-onset breast cancer in French-Canadian women", *Breast Cancer Research,* vol. 9, no. 6, pp. R83.

Foulkes, W.D., Smith, I.E. and Reis-Filho, J.S., 2010, “Triple-negative breast cancer”, *New England journal of medicine*, vol. 363, no. 20, pp.1938-1948.

Friend, S.H., Bernards, R., Rogelj, S., Weinberg, R.A., Rapaport, J.M., Albert, D.M. & Dryja, T.P. 1986, "A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma", *Nature,* vol. 323, no. 6089, pp. 643-646.

Frio, T.R., Bahubeshi, A., Kanellopoulou, C., Hamel, N., Niedziela, M., Sabbaghian, N., Pouchet, C., Gilbert, L., O’Brien, P.K., Serfas, K. and Broderick, P., 2011, “DICER1 mutations in familial multinodular goiter with and without ovarian Sertoli-Leydig cell tumors”, *JAMA*, vol. 305, no. 1, pp.68-77.

Fromer, M., Moran, J., Chambert, K., Banks, E., Bergen, S., Ruderfer, D., Handsaker, R., McCarroll, S., O’Donovan, M., Owen, M., Kirov, G., Sullivan, P., Hultman, C., Sklar, P. & Purcell, S. 2012, "Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth", *American Journal of Human Genetics,* vol. 91, no. 4, pp. 597-607.

Fromer, M. & Purcell, S.M. 2014, "Using XHMM software to detect copy number variation in whole-exome sequencing data", *Current protocols in human genetics,* vol. 81, pp. 7.23.1-7.23.21.

Fuller, C.W., Middendorf, L.R., Benner, S.A., Church, G.M., Harris, T., Huang, X., Jovanovich, S.B., Nelson, J.R., Schloss, J.A., Schwartz, D.C. and Vezenov,

D.V. 2009, “The challenges of sequencing by synthesis”, *Nature biotechnology*, Vol. 27, pp.1013-1023.

Futami, K., Kumagai, E., Makino, H., Goto, H., Takagi, M., Shimamoto, A. & Furuichi, Y. 2008, "Induction of mitotic cell death in cancer cells by small interference RNA suppressing the expression of RecQL1 helicase", *Cancer Science,* vol. 99, no. 1, pp. 71-80.

Ge, B., Gurd, S., Gaudin, T., Dore, C., Lepage, P., Harmsen, E., Hudson, T.J. & Pastinen, T. 2005, "Survey of allelic expression using EST mining", *Genome research,* vol. 15, no. 11, pp. 1584-1591.

Gerlinger, M., Rowan, A.J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., Varela, I., Phillimore, B., Begum, S., McDonald, N.Q., Butler, A., Jones, D., Raine, K., Latimer, C., Santos, C.R., Nohadani, M., Eklund, A.C., Spencer-Dene, B., Clark, G., Pickering, L., Stamp, G., Gore, M., Szallasi, Z., Downward, J., Futreal, P.A. & Swanton, C. 2012, "Intratumor heterogeneity and branched evolution revealed by multiregion sequencing", *New England Journal of Medicine,* vol. 366, no. 10, pp. 883-892.

Gerstung, M., Beisel, C., Rechsteiner, M., Wild, P., Schraml, P., Moch, H. & Beerenwinkel, N. 2012, "Reliable detection of subclonal single-nucleotide variants in tumour cell populations", *Nature communications,* vol. 3, pp. 811.

Gerstung, M., Papaemmanuil, E. & Campbell, P.J. 2013, "Subclonal variant calling with multiple samples and prior knowledge", *Bioinformatics,* vol. 30, no. 9, pp. 1198-1204.

Ghadirian, P., Robidoux, A., Zhang, P., Royer, R., Akbari, M., Zhang, S., Fafard, E., Costa, M., Martin, G., Potvin, C., Patocskai, E., Larouche, N., Younan, R., Nassif, E., Giroux, S., Narod, S.A., Rousseau, F. & Foulkes, W.D. 2009, "The contribution of founder mutations to early-onset breast cancer in French-Canadian women", *Clinical genetics,* vol. 76, no. 5, pp. 421-426.

Gilbert, M.T., Haselkorn, T., Bunce, M., Sanchez, J.J., Lucas, S.B., Jewell, L.D., Marck, E.V. & Worobey, M. 2007, "The isolation of nucleic acids from fixed, paraffin-embedded tissues–which methods are useful when?", *PLoS ONE,* vol. 2, no. 6, pp. e537.

Glubb, D., Maranian, M., Michailidou, K., Pooley, K., Meyer, K., Kar, S., Carlebur, S., O’Reilly, M., Betts, J., Hillman, K., Kaufmann, S., Beesley, J., Canisius, S., Hopper, J., Southey, M., Tsimiklis, H., Apicella, C., Schmidt, M., Broeks, A., Hogervorst, F., van der Schoot, C. ., Muir, K., Lophatananon, A., Stewart-Brown, S., Siriwanarangsan, P., Fasching, P., Ruebner, M., Ekici, A., Beckmann, M., Peto, J., dos-Santos-Silva, I., Fletcher, O., Johnson, N., Pharoah, P.P., Bolla, M., Wang, Q., Dennis, J., Sawyer, E., Tomlinson, I., Kerin, M., Miller, N., Burwinkel, B., Marme, F., Yang, R., Surowy, H., Guénel, P., Truong, T., Menegaux, F., Sanchez, M., Bojesen, S., Nordestgaard, B., Nielsen, S., Flyger, H., González-Neira, A., Benitez, J., Zamora, M. ., Arias Perez, J., Anton-Culver, H., Neuhausen, S., Brenner, H., Dieffenbach, A., Arndt, V., Stegmaier, C., Meindl, A., Schmutzler, R., Brauch, H., Ko, Y., Brüning, T., Nevanlinna, H., Muranen, T., Aittomäki, K., Blomqvist, C., Matsuo, K., Ito, H., Iwata, H., Tanaka, H., Dörk, T., Bogdanova, N., Helbig,

S., Lindblom, A., Margolin, S., Mannermaa, A., Kataja, V., Kosma, V., Hartikainen, J., Wu, A., Tseng, C., Van Den Berg, D., Stram, D., Lambrechts, D., Zhao, H., Weltens, C., van Limbergen, E., Chang-Claude, J., Flesch-Janys, D., Rudolph, A., Seibold, P., Radice, P., Peterlongo, P., Barile, M., Capra, F., Couch, F., Olson, J., Hallberg, E., Vachon, C., Giles, G., Milne, R., McLean, C., Haiman, C., Henderson, B., Schumacher, F., Le Marchand, L., Simard, J., Goldberg, M., Labrèche, F., Dumont, M., Teo, S., Yip, C., See, M., Cornes, B., Cheng, C., Ikram, M. ., Kristensen, V., Zheng, W., Halverson, S., Shrubsole, M., Long, J., Winqvist, R., Pylkäs, K., Jukkola-Vuorinen, A., Kauppila, S., Andrulis, I., Knight, J., Glendon, G., Tchatchou, S., Devilee, P., Tollenaar, R.E.M., Seynaeve, C., Van Asperen, C., García-Closas, M., Figueroa, J., Chanock, S., Lissowska, J., Czene, K., Klevebring, D., Darabi, H., Eriksson, M., Hooning, M., Hollestelle, A., Martens, J.M., Collée, J. ., Hall, P., Li, J., Humphreys, K., Shu, X., Lu, W., Gao, Y., Cai, H., Cox, A., Cross, S., Reed, M.R., Blot, W., Signorello, L., Cai, Q., Shah, M., Ghoussaini, M., Kang, D., Choi, J., Park, S., Noh, D., Hartman, M., Miao, H., Lim, W., Tang, A., Hamann, U., Torres, D., Jakubowska, A., Lubinski, J., Jaworska, K., Durda, K., Sangrajrang, S., Gaborieau, V., Brennan, P., McKay, J., Olswold, C., Slager, S., Toland, A., Yannoukakos, D., Shen, C., Wu, P., Yu, J., Hou, M., Swerdlow, A., Ashworth, A., Orr, N., Jones, M., Pita, G., Alonso, M. ., Álvarez, N., Herrero, D., Tessier, D., Vincent, D., Bacot, F., Luccarini, C., Baynes, C., Ahmed, S., Healey, C., Brown, M., Ponder, B.J., Chenevix-Trench, G., Thompson, D., Edwards, S., Easton, D., Dunning, A. & French, J. 2015, "Fine-Scale mapping of the 5q11.2 breast cancer locus reveals at least three independent risk variants regulating MAP3K1", *The American Journal of Human Genetics,* vol. 96, no. 1, pp. 5-20.

Graw, S., Meier, R., Minn, K., Bloomer, C., Godwin, A.K., Fridley, B., Vlad, A., Beyerlein, P. & Chien, J. 2015, "Robust gene expression and mutation analyses of RNA-sequencing of formalin-fixed diagnostic tumor samples", *Scientific Reports,* vol. 5, pp. 12335.

Gundry, M. & Vijg, J. 2012, "Direct mutation analysis by high-throughput sequencing: From germline to low-abundant, somatic variants", *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis,* vol. 729, no. 1–2, pp. 1-15.

Haas, J., Katus, H.A. & Meder, B. 2011, "Next-generation sequencing entering the clinical arena", *Molecular and cellular probes,* vol. 25, no. 5–6, pp. 206-211.

Hanahan, D. & Weinberg, R. 2011, "Hallmarks of Cancer: The Next Generation", *Cell,* vol. 144, no. 5, pp. 646-674.

Hasselblatt, M., Gesk, S., Oyen, F., Rossi, S., Viscardi, E., Giangaspero, F., Giannini, C., Judkins, A.R., Frühwald, M.,C., Obser, T., Schneppenheim, R., Siebert, R. & Paulus, W. 2011, "Nonsense mutation and inactivation of SMARCA4 (BRG1) in an atypical teratoid/rhabdoid tumor showing retained SMARCB1 (INI1) expression", *The American Journal of Surgical Pathology,* vol. 35, no. 6, pp. 933-935.

He, Y., Qiao, Z., Gao, B., Zhang, X. & Wen, Y. 2014, "Association between RECQL5 genetic polymorphisms and susceptibility to breast cancer", *Tumor Biology,* vol. 35, no. 12, pp. 12201-12204.

Hemminki, K., Sundquist, J. & Bermejo, J.L. 2008, "How common is familial cancer?", *Annals of Oncology,* vol. 19, no. 1, pp. 163-167.

Heravi-Moussavi, A., Anglesio, M.S., Cheng, S.-.G., Senz, J., Yang, W., Prentice, L., Fejes, A.P., Chow, C., Tone, A., Kalloger, S.E., Hamel, N., Roth, A., Ha, G., Wan, A.N.C., Maines-Bandiera, S., Salamanca, C., Pasini, B., Clarke, B.A., Lee, A.F., Lee, C., Zhao, C., Young, R.H., Aparicio, S.A., Sorensen, P.H.B., Woo, M.M.M., Boyd, N., Jones, S.J.M., Hirst, M., Marra, M.A., Gilks, B., Shah, S.P., Foulkes, W.D., Morin, G.B. & Huntsman, D.G. 2012, "Recurrent Somatic DICER1 Mutations in Nonepithelial Ovarian Cancers", *New England Journal of Medicine,* vol. 366, no. 3, pp. 234-242.

Hilbers, F.S., Meijers, C.M., Laros, J.F.J., van Galen, M., Hoogerbrugge, N., Vasen, H.F.A., Nederlof, P.M., Wijnen, J.T., van Asperen, C.,J. & Devilee, P. 2012, "Exome Sequencing of Germline DNA from Non-BRCA1/2 Familial Breast Cancer Cases Selected on the Basis of aCGH Tumor Profiling", *PLoS ONE,* vol. 8, no. 1, pp. e55734.

Hilbers, F., Vreeswijk, M., van Asperen, C. & Devilee, P. 2013, "The impact of next generation sequencing on the analysis of breast cancer susceptibility: a role for extremely rare genetic variation?", *Clinical genetics,* vol. 84, no. 5, pp. 407-414.

Ho, C., Kurman, R.J., Dehari, R., Wang, T. & Shih, I. 2004, "Mutations of BRAF and KRAS rrecede the development of ovarian serous borderline tumors", *Cancer research,* vol. 64, no. 19, pp. 6915-6918.

Houlston, R.S. & Peto, J. 2004, "The search for low-penetrance cancer susceptibility alleles", *Oncogene,* vol. 23, no. 38, pp. 6471-6476.

Hsu, Y.L., Huang, M.S., Yang, C.J., Hung, J.Y., Wu, L.Y. & Kuo, P.L. 2011, "Lung tumor-associated osteoblast-derived bone morphogenetic protein-2 increased epithelial-to-mesenchymal transition of cancer by Runx2/Snail signaling pathway", *The Journal of biological chemistry,* vol. 286, no. 43, pp. 37335-37346.

Jamieson, S., Butzow, R., Andersson, N., Alexiadis, M., Unkila-Kallio, L., Heikinheimo, M., Fuller, P.J. & Anttonen, M. 2010, "The FOXL2 C134W mutation is characteristic of adult granulosa cell tumors of the ovary", *Modern pathology,* vol. 23, no. 11, pp. 1477-1485.

Jasmine, F., Rahaman, R., Roy, S., Raza, M., Paul, R., Rakibuz-Zaman, M., Paul-Brutus, R., Dodsworth, C., Kamal, M., Ahsan, H. & Kibriya, M.G. 2012, "Interpretation of genome-wide infinium methylation data from ligated DNA in formalin-fixed, paraffin-embedded paired tumor and normal tissue", *BMC Research Notes,* vol. 5, no. 1, pp. 1-11.

Jayson, G.C., Kohn, E.C., Kitchener, H.C. & Ledermann, J.A. 2014, "Ovarian cancer", *The Lancet,* vol. 384, no. 9951, pp. 1376-1388.

Jelinic, P., Mueller, J.J., Olvera, N., Dao, F., Scott, S.N., Shah, R., Gao, J., Schultz, N., Gonen, M. & Soslow, R.A. 2014, "Recurrent SMARCA4

mutations in small cell carcinoma of the ovary", *Nature genetics,* vol. 46, no. 5, pp. 424-426.

Jelinic, P., Schlappe, B.,A., Conlon, N., Tseng, J., Olvera, N., Dao, F., Mueller, J.,J., Hussein, Y., Soslow, R.,A. & Levine, D.,A. 2016, "Concomitant loss of SMARCA2 and SMARCA4 expression in small cell carcinoma of the ovary, hypercalcemic type", *Modern pathology,* vol. 29, no. 1, pp. 60-66.

Jenne, D.E., Reomann, H., Nezu, J., Friedel, W., Loff., S., Jeschke, R., Muller, O., Back, W. & Zimmer, M. 1998, "Peutz-Jeghers syndrome is caused by mutations in a novel serine threoninekinase", *Nature genetics,* vol. 18, no. 1, pp. 38-43.

Jinushi, T., Shibayama, Y., Kinoshita, I., Oizumi, S., Jinushi, M., Aota, T., Takahashi, T., Horita, S., Dosaka‐Akita, H. & Iseki, K. 2014, "Low expression levels of microRNA‐124‐5p correlated with poor prognosis in colorectal cancer via targeting of SMC4", *Cancer medicine,* vol. 3, no. 6, pp. 1544-1552.

Johns, L.E. & Houlston, R.S. 2001, "A systematic review and meta-analysis of familial colorectal cancer risk", *The American Journal of Gastroenterology,* vol. 96, no. 10, pp. 2992-3003.

Jones, S., Wang, T., Shih, I., Mao, T., Nakayama, K., Roden, R., Glas, R., Slamon, D., Diaz, L.A., Vogelstein, B., Kinzler, K.W., Velculescu, V.E. & Papadopoulos, N. 2010, "Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma", *Science,* vol. 330, no. 6001, pp. 228-231.

Kadoch, C., Hargreaves, D.C., Hodges, C., Elias, L., Ho, L., Ranish, J. & Crabtree, G.R. 2013, "Proteomic and bioinformatic analysis of mSWI/SNF (BAF) complexes reveals extensive roles in human malignancy", *Nature genetics,* vol. 45, no. 6, pp. 592-601.

Kaku, T., Ogawa, S., Kawano, Y., Ohishi, Y., Kobayashi, H., Hirakawa, T. & Nakano, H. 2003, "Histological classification of ovarian cancer", *Medical Electron Microscopy,* vol. 36, no. 1, pp. 9-17.

Kemmer, K., Corless, C.L., Fletcher, J.A., McGreevey, L., Haley, A., Griffith, D., Cummings, O.W., Wait, C., Town, A. & Heinrich, M.C. 2004, "KIT mutations are common in testicular seminomas", *The American Journal of Pathology,* vol. 164, no. 1, pp. 305-313.

Kiiski, J.I., Pelttari, L.M., Khan, S., Freysteinsdottir, E.S., Reynisdottir, I., Hart, S.N., Shimelis, H., Vilske, S., Kallioniemi, A., Schleutker, J., Leminen, A., Bützow, R., Blomqvist, C., Barkardottir, R.B., Couch, F.J., Aittomäki, K. & Nevanlinna, H. 2014, "Exome sequencing identifies FANCM as a susceptibility gene for triple-negative breast cancer", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 111, no. 42, pp. 15172-15177.

Kim, M.Y., Oskarsson, T., Acharyya, S., Nguyen, D.X., Zhang, X.H.F., Norton, L. and Massagué, J., 2009, “Tumor self-seeding by circulating cancer cells”, *Cell*, vol. 139, no. 7, pp.1315-1326.

Kinzler, K.W. & Vogelstein, B. 1997, "Gatekeepers and caretakers", *Nature,* vol. 386, no. 6627, pp. 761-763.

Kinzler, K.W. & Vogelstein, B. 1996, "Lessons from hereditary colorectal cancer", *Cell,* vol. 87, no. 2, pp. 159-170.

Kircher, M., Stenzel, U. & Kelso, J. 2009, "Improved base calling for the Illumina Genome Analyzer using machine learning strategies", *Genome biology,* vol. 10, no. 8, pp. R83.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.,J., Cooper, G.M. & Shendure, J. 2014, "A general framework for estimating the relative pathogenicity of human genetic variants", *Nature genetics,* vol. 46, no. 3, pp. 310-315.

Klein, C.A. 2009, "Parallel progression of primary tumours and metastases", *Nature reviews Cancer,* vol. 9, no. 4, pp. 302-312.

Knudson, A.G. 1971, "Mutation and cancer: statistical study of retinoblastoma", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 68, no. 4, pp. 820-823.

Knudson, A.G., Meadows, A.T., Nichols, W.W. & Hill, R. 1976, "Chromosomal deletion and retinoblastoma", *New England Journal of Medicine,* vol. 295, no. 20, pp. 1120-1123.

Köbel, M., Kalloger, S.E., Boyd, N., McKinney, S., Mehl, E., Palmer, C., Leung, S., Bowen, N.J., Ionescu, D.N., Rajput, A., Prentice, L.M., Miller, D., Santos, J., Swenerton, K., Gilks, C.B. & Huntsman, D. 2008, "Ovarian carcinoma subtypes are different diseases: Implications for biomarker studies", *PLoS Medicine,* vol. 5, no. 12, pp. e232.

Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L. & Wilson, R.K. 2012, "VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing", *Genome research,* vol. 22, no. 3, pp. 568-576.

Kuilman, T., Velds, A., Kemper, K., Ranzani, M., Bombardelli, L., Hoogstraat, M., Nevedomskaya, E., Xu, G., de Ruiter, J., Lolkema, M.P. and Ylstra, B., 2015, "CopywriteR: DNA copy number detection from off-target sequence data", *Genome Biology*, vol. 1, pp. 49.

Kulawiec, M., Safina, A.F., Desouki, M.M., Still, I., Matsui, S., Bakin, A. & Singh, K.K. 2008, "Tumorigenic transformation of human breast epithelial cells induced by mitochondrial DNA depletion", *Cancer biology & therapy,* vol. 7, no. 11, pp. 1732-1743.

Kuo, K., Mao, T., Chen, X., Feng, Y., Nakayama, K., Wang, Y., Glas, R., Ma, M.J., Kurman, R.J., Shih, I. & Wang, T. 2010, "DNA copy number profiles in affinity-purified ovarian clear cell carcinoma", *Clinical cancer research,* vol. 16, no. 7, pp. 1997-2008.

Kupryjańczyk, J., Dansonka-Mieszkowska, A., Moes-Sosnowska, J., Plisiecka-Hałasa, J., Szafron, L., Podgórska, A., Rzepecka, I.K., Konopka, B., Budziłowska, A. & Rembiszewska, A. 2013, "Ovarian small cell carcinoma of hypercalcemic type-evidence of germline origin and smarca4 gene inactivation. a pilot study", *Polish Journal of Pathology,* vol. 64, no. 4, pp. 238-246.

Kurek, K.C., Luks, V.L., Ayturk, U.M., Alomari, A.I., Fishman, S.J., Spencer, S.A., Mulliken, J.B., Bowen, M.E., Yamamoto, G.L., Kozakewich, H.P. and Warman, M.L. 2012, "Somatic mosaic activating mutations in PIK3CA cause

CLOVES syndrome", *The American Journal of Human Genetics*, vol. 90, pp.1108-1115.

Laberge, A., Michaud, J., Richter, A., Lemyre, E., Lambert, M., Brais, B. & Mitchell, G.A. 2005, "Population history and its impact on medical genetics in Quebec", *Clinical genetics,* vol. 68, no. 4, pp. 287-301.

Lamovec, J., Bracko, M. & Cerar, O. 1995, "Familial occurrence of small-cell carcinoma of the ovary", *Archives of Pathology & Laboratory Medicine,* vol. 119, no. 6, pp. 523-527.

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. 2009, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome", *Genome biology,* vol. 10, no. 3, pp. R25-R25.

Larson, D.E., Harris, C.C., Chen, K., Koboldt, D.C., Abbott, T.E., Dooling, D.J., Ley, T.J., Mardis, E.R., Wilson, R.K. & Ding, L. 2012, "SomaticSniper: identification of somatic point mutations in whole genome sequencing data", *Bioinformatics,* vol. 28, no. 3, pp. 311-317.

Larson, N.B. & Fridley, B.L. 2013, "PurBayes: estimating tumor cellularity and subclonality in next-generation sequencing data", *Bioinformatics,* vol. 29, no. 15, pp. 1888-1889.

Ledergerber, C. & Dessimoz, C. 2010, "Base-calling for next-generation sequencing platforms", *Briefings in Bioinformatics,* vol. 12, no. 5, pp. 489-497.

Lee, M.S., Green, R., Marsillac, S.M., Coquelle, N., Williams, R.S., Yeung, T., Foo, D., Hau, D.D., Hui, B., Monteiro, A. & Glover, J. 2010, "Comprehensive analysis of missense variations in the BRCT domain of BRCA1 by structural and functional assays", *Cancer Res,* vol. 70, no. 12, pp. 4880-4890.

Lee, R.S., Stewart, C., Carter, S.L., Ambrogio, L., Cibulskis, K., Sougnez, C., Lawrence, M.S., Auclair, D., Mora, J., Golub, T.R., Biegel, J.A., Getz, G. & Roberts, C.W.M. 2012, "A remarkably simple genome underlies highly malignant pediatric rhabdoid cancers", *The Journal of clinical investigation,* vol. 122, no. 8, pp. 2983-2988.

Li, H. 2011, "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data", *Bioinformatics,* vol. 27, no. 21, pp. 2987-2993.

Li, H. & Durbin, R. 2009, "Fast and accurate short read alignment with Burrows-Wheeler transform", *Bioinformatics,* vol. 25, no. 14, pp. 1754-1760.

Li, H., Ruan, J. & Durbin, R. 2008, "Mapping short DNA sequencing reads and calling variants using mapping quality scores", *Genome research,* vol. 18, no. 11, pp. 1851-1858.

Li, Q., Seo, J., Stranger, B., McKenna, A., Pe'er, I., LaFramboise, T., Brown, M., Tyekucheva, S. & Freedman, M. 2013, "Integrative eQTL-based analyses reveal the biology of breast cancer risk loci", *Cell,* vol. 152, no. 3, pp. 633-641.

Liaw, D., Marsh, D.J., Li, J., Dahia, P.L.M., Wang, S.I., Zheng, Z., Bose, S., Call, K.M., Tsou, H.C., Peacoke, M., Eng, C. & Parsons, R. 1997, "Germline mutations of the PTEN gene in Cowden disease, an inherited breast and thyroid cancer syndrome", *Nature genetics,* vol. 16, no. 1, pp. 64-67.

Longy, M., Toulouse, C., Mage, P., Chauvergne, J. & Trojani, M. 1996, "Familial cluster of ovarian small cell carcinoma: a new mendelian entity?", *Journal of medical genetics,* vol. 33, no. 4, pp. 333-335.

Lord, C.J. & Ashworth, A. 2012, "The DNA damage response and cancer therapy", *Nature,* vol. 481, no. 7381, pp. 287-294.

Losada, A. & Hirano, T., 2005, “Dynamic molecular linkers of the genome: the first decade of SMC proteins”, *Genes & development*, vol.19, no. 11, pp.1269-1287.

Loveday, C., Turnbull, C., Ramsay, E., Hughes, D., Ruark, E., Frankum, J.R., Bowden, G., Kalmyrzaev, B., Warren-Perry, M., Snape, K., Adlard, J.W., Barwell, J., Berg, J., Brady, A.F., Brewer, C., Brice, G., Chapman, C., Cook, J., Davidson, R., Donaldson, A., Douglas, F., Greenhalgh, L., Henderson, A., Izatt, L., Kumar, A., Lalloo, F., Miedzybrodzka, Z., Morrison, P.J., Paterson, J., Porteous, M., Rogers, M.T., Shanley, S., Walker, L., Eccles, D., Evans, D.G., Renwick, A., Seal, S., Lord, C.J., Ashworth, A., Reis-Filho, J., Antoniou, A.C. & Rahman, N. 2011, "Germline mutations in RAD51D confer susceptibility to ovarian cancer", *Nature genetics,* vol. 43, no. 9, pp. 879-882.

Luvero, D., Milani, A. & Ledermann, J.A. 2014, "Treatment options in recurrent ovarian cancer: latest evidence and clinical potential", *Therapeutic Advances in Medical Oncology,* vol. 6, no. 5, pp. 229-239.

Ma, X., Edmonson, M., Yergeau, D., Muzny, D.M., Hampton, O.A., Rusch, M., Song, G., Easton, J., Harvey, R.C., Wheeler, D.A., Ma, J., Doddapaneni, H., Vadodaria, B., Wu, G., Nagahawatte, P., Carroll, W.L., Chen, I., Gastier-Foster, J., Relling, M.V., Smith, M.A., Devidas, M., Auvil, J.M.G., Downing, J.R., Loh, M.L., Willman, C.L., Gerhard, D.S., Mullighan, C.G., Hunger, S.P. & Zhang, J. 2015, "Rise and fall of subclones from diagnosis to relapse in pediatric B-acute lymphoblastic leukaemia", *Nature Communications,* vol. 6, pp. 6604.

Majewski, I.J., Mittempergher, L., Davidson, N.M., Bosma, A., Willems, S.M., Horlings, H.M., de Rink, I., Greger, L., Hooijer, G.K. & Peters, D. 2013, "Identification of recurrent FGFR3 fusion genes in lung cancer through kinome‐centred RNA sequencing", *The Journal of pathology,* vol. 230, no. 3, pp. 270-276.

Majewski, J., Schwartzentruber, J., Lalonde, E., Montpetit, A. & Jabado, N. 2011, "What can exome sequencing do for you?", *Journal of medical genetics,* vol. 48, no. 9, pp. 580-589.

Malkin, D., Li, F., Strong, L., Fraumeni, J., Nelson, C., Kim, D., Kassel, J., Gryka, M., Bischoff, F., Tainsky, M. & et, a. 1990, "Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms", *Science,* vol. 250, no. 4985, pp. 1233-1238.

Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E.H., Kumar, A., Howard, E., Shendure, J. & Turner, D.J. 2010, "Target-enrichment strategies for next-generation sequencing", *Nature Method,* vol. 7, no. 2, pp. 111-118.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk,

G.P., Jando, S.C., Alenquer, M.L.I., Jarvie, T.P., Jirage, K.B., Kim, J., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F. & Rothberg, J.M. 2005, "Genome sequencing in microfabricated high-density picolitre reactors", *Nature,* vol. 437, no. 7057, pp. 376-380.

Martinez-Borges, A., Petty, J.K., Hurt, G., Stribling, J.T., Press, J.Z. & Castellino, S.M. 2009, "Familial small cell carcinoma of the ovary", *Pediatric Blood & Cancer,* vol. 53, no. 7, pp. 1334-1336.

Marusyk, A., Almendro, V. & Polyak, K. 2012, "Intra-tumour heterogeneity: a looking glass for cancer?", *Nat Rev Cancer,* vol. 12, no. 5, pp. 323-334.

Matsubara, D., Kishaba, Y., Ishikawa, S., Sakatani, T., Oguni, S., Tamura, T., Hoshino, H., Sugiyama, Y., Endo, S., Murakami, Y., Aburatani, H., Fukayama, M. & Niki, T. 2013, "Lung cancer with loss of BRG1/BRM, shows epithelial mesenchymal transition phenotype and distinct histologic and genetic features", *Cancer Science,* vol. 104, no. 2, pp. 266-273.

Maxwell, K.N. & Nathanson, K.L. 2013, "Common breast cancer risk variants in the post-COGS era: a comprehensive review", *Breast Cancer Research,* vol. 15, no. 6, pp. 212-212.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & DePristo, M.A. 2010, "The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data", *Genome research,* vol. 20, no. 9, pp. 1297-1303.

Meindl, A., Hellebrand, H., Wiek, C., Erven, V., Wappenschmidt, B., Niederacher, D., Freund, M., Lichtner, P., Hartmann, L., Schaal, H., Ramser, J., Honisch, E., Kubisch, C., Wichmann, H.E., Kast, K., Deiszler, H., Engel, C., Muller-Myhsok, B., Neveling, K., Kiechle, M., Mathew, C.G., Schindler, D., Schmutzler, R.K. & Hanenberg, H. 2010, "Germline mutations in breast and ovarian cancer pedigrees establish RAD51C as a human cancer susceptibility gene", *Nature genetics,* vol. 42, no. 5, pp. 410-414.

Meyer, K., O'Reilly, M., Michailidou, K., Carlebur, S., Edwards, S., French, J., Prathalingham, R., Dennis, J., Bolla, M.K., Wang, Q., de Santiago, I., Hopper, J., Tsimiklis, H., Apicella, C., Southey, M., Schmidt, M., Broeks, A., Van 't Veer, L., Hogervorst, F., Muir, K., Lophatananon, A., Stewart-Brown, S., Siriwanarangsan, P., Fasching, P., Lux, M., Ekici, A., Beckmann, M., Peto, J., dos Santos Silva, I., Fletcher, O., Johnson, N., Sawyer, E., Tomlinson, I., Kerin, M., Miller, N., Marme, F., Schneeweiss, A., Sohn, C., Burwinkel, B., Guénel, P., Truong, T., Laurent-Puig, P., Menegaux, F., Bojesen, S., Nordestgaard, B., Nielsen, S., Flyger, H., Milne, R., Zamora, M. ., Arias, J., Benitez, J., Neuhausen, S., Anton-Culver, H., Ziogas, A., Dur, C., Brenner, H., Müller, H., Arndt, V., Stegmaier, C., Meindl, A., Schmutzler, R., Engel, C., Ditsch, N., Brauch, H., Brüning, T., Ko, Y., Nevanlinna, H.,

Muranen, T., Aittomäki, K., Blomqvist, C., Matsuo, K., Ito, H., Iwata, H., Yatabe, Y., Dörk, T., Helbig, S., Bogdanova, N., Lindblom, A., Margolin, S., Mannermaa, A., Kataja, V., Kosma, V., Hartikainen, J., Chenevix-Trench, G., Wu, A., Tseng, C., Van Den Berg, D., Stram, D., Lambrechts, D., Thienpont, B., Christiaens, M., Smeets, A., Chang-Claude, J., Rudolph, A., Seibold, P., Flesch-Janys, D., Radice, P., Peterlongo, P., Bonanni, B., Bernard, L., Couch, F., Olson, J., Wang, X., Purrington, K., Giles, G., Severi, G., Baglietto, L., McLean, C., Haiman, C., Henderson, B., Schumacher, F., Le Marchand, L., Simard, J., Goldberg, M., Labrèche, F., Dumont, M., Teo, S., Yip, C., Phuah, S., Kristensen, V., Grenaker Alnæs, G., Børresen-Dale, A., Zheng, W., Deming-Halverson, S., Shrubsole, M., Long, J., Winqvist, R., Pylkäs, K., Jukkola-Vuorinen, A., Kauppila, S., Andrulis, I., Knight, J., Glendon, G., Tchatchou, S., Devilee, P., Tollenaar, R.E.M., Seynaeve, C., García-Closas, M., Figueroa, J., Chanock, S.J., Lissowska, J., Czene, K., Darabi, H., Eriksson, K., Hooning, M., Martens, J.M., van den Ouweland, A.W., van Deurzen, C.M., Hall, P., Li, J., Liu, J., Humphreys, K., Shu, X., Lu, W., Gao, Y., Cai, H., Cox, A., Reed, M.R., Blot, W., Signorello, L.B., Cai, Q., Pharoah, P.P., Ghoussaini, M., Harrington, P., Tyrer, J., Kang, D., Choi, J., Park, S., Noh, D., Hartman, M., Hui, M., Lim, W., Buhari, S., Hamann, U., Försti, A., Rüdiger, T., Ulmer, H., Jakubowska, A., Lubinski, J., Jaworska, K., Durda, K., Sangrajrang, S., Gaborieau, V., Brennan, P., McKay, J., Vachon, C., Slager, S., Fostira, F., Pilarski, R., Shen, C., Hsiung, C., Wu, P., Hou, M., Swerdlow, A., Ashworth, A., Orr, N., Schoemaker, M., Ponder, B.J., Dunning, A. & Easton, D. 2013, "Fine-scale mapping of the FGFR2 breast cancer risk locus: putative functional variants differentially bind FOXA1 and E2F1", *The American Journal of Human Genetics,* vol. 93, no. 6, pp. 1046-1060.

Meyerson, M., Gabriel, S. & Getz, G. 2010, "Advances in understanding cancer genomes through second-generation sequencing", *Nature reviews Genetics,* vol. 11, no. 10, pp. 685-696.

Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R.L., Schmidt, M.K., Chang-Claude, J., Bojesen, S.E., Bolla, M.K., Wang, Q., Dicks, E., Lee, A., Turnbull, C., Rahman, N., Fletcher, O., Peto, J., Gibson, L., dos, S.S., Nevanlinna, H., Muranen, T.A., Aittomaki, K., Blomqvist, C., Czene, K., Irwanto, A., Liu, J., Waisfisz, Q., Meijers-Heijboer, H., Adank, M., van der Luijt, R.,B., Hein, R., Dahmen, N., Beckman, L., Meindl, A., Schmutzler, R.K., Muller-Myhsok, B., Lichtner, P., Hopper, J.L., Southey, M.C., Makalic, E., Schmidt, D.F., Uitterlinden, A.G., Hofman, A., Hunter, D.J., Chanock, S.J., Vincent, D., Bacot, F., Tessier, D.C., Canisius, S., Wessels, L.F.A., Haiman, C.A., Shah, M., Luben, R., Brown, J., Luccarini, C., Schoof, N., Humphreys, K., Li, J., Nordestgaard, B.G., Nielsen, S.F., Flyger, H., Couch, F.J., Wang, X., Vachon, C., Stevens, K.N., Lambrechts, D., Moisse, M., Paridaens, R., Christiaens, M., Rudolph, A., Nickels, S., Flesch-Janys, D., Johnson, N., Aitken, Z., Aaltonen, K., Heikkinen, T., Broeks, A., Veer, L.J.V., van, d.S., Guenel, P., Truong, T., Laurent-Puig, P., Menegaux, F., Marme, F., Schneeweiss, A., Sohn, C., Burwinkel, B., Zamora, M.P., Perez, J.I.A., Pita, G., Alonso, M.R., Cox, A., Brock, I.W., Cross, S.S., Reed,

M.W.R., Sawyer, E.J., Tomlinson, I., Kerin, M.J., Miller, N., Henderson, B.E., Schumacher, F., Le Marchand, L., Andrulis, I.L., Knight, J.A., Glendon, G., Mulligan, A.M., Lindblom, A., Margolin, S., Hooning, M.J., Hollestelle, A., van den Ouweland, A.,M.W., Jager, A., Bui, Q.M., Stone, J., Dite, G.S., Apicella, C., Tsimiklis, H., Giles, G.G., Severi, G., Baglietto, L., Fasching, P.A., Haeberle, L., Ekici, A.B., Beckmann, M.W., Brenner, H., Muller, H., Arndt, V., Stegmaier, C., Swerdlow, A., Ashworth, A., Orr, N., Jones, M., Figueroa, J., Lissowska, J., Brinton, L., Goldberg, M.S., Labreche, F., Dumont, M., Winqvist, R., Pylkas, K., Jukkola-Vuorinen, A., Grip, M., Brauch, H., Hamann, U., Bruning, T., Radice, P., Peterlongo, P., Manoukian, S., Bonanni, B., Devilee, P., Tollenaar, R.A.E.M., Seynaeve, C., van Asperen, C.,J., Jakubowska, A., Lubinski, J., Jaworska, K., Durda, K., Mannermaa, A., Kataja, V., Kosma, V., Hartikainen, J.M., Bogdanova, N.V., Antonenkova, N.N., Dork, T., Kristensen, V.N., Anton-Culver, H., Slager, S., Toland, A.E., Edge, S., Fostira, F., Kang, D., Yoo, K., Noh, D., Matsuo, K., Ito, H., Iwata, H., Sueta, A., Wu, A.H., Tseng, C., Van, D.B., Stram, D.O., Shu, X., Lu, W., Gao, Y., Cai, H., Teo, S.H., Yip, C.H., Phuah, S.Y., Cornes, B.K., Hartman, M., Miao, H., Lim, W.Y., Sng, J., Muir, K., Lophatananon, A., Stewart-Brown, S., Siriwanarangsan, P., Shen, C., Hsiung, C., Wu, P., Ding, S., Sangrajrang, S., Gaborieau, V., Brennan, P., McKay, J., Blot, W.J., Signorello, L.B., Cai, Q., Zheng, W., Deming-Halverson, S., Shrubsole, M., Long, J., Simard, J., Garcia-Closas, M., Pharoah, P.D.P., Chenevix-Trench, G., Dunning, A.M., Benitez, J. & Easton, D.F. 2013, "Large-scale genotyping identifies 41 new loci associated with breast cancer risk", *Nature genetics,* vol. 45, no. 4, pp. 353-361.

Michailidou, K., Beesley, J., Lindstrom, S., Canisius, S., Dennis, J., Lush, M.J., Maranian, M.J., Bolla, M.K., Wang, Q., Shah, M. and Perkins, B.J., 2015, “Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer”, *Nature genetics*, vol. 47, no. 4, pp.373-380.

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L., Ding, W. & et, a. 1994, "A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1", *Science,* vol. 266, no. 5182, pp. 66-71.

Mullighan, C.G., Phillips, L.A., Su, X., Ma, J., Miller, C.B., Shurtleff, S.A. & Downing, J.R. 2008, "Genomic Analysis of the Clonal Origins of Relapsed Acute Lymphoblastic Leukemia", *Science,* vol. 322, no. 5906, pp. 1377-1380.

Nadaf, J., Majewski, J. & Fahiminiya, S. 2015, "ExomeAI: detection of recurrent allelic imbalance in tumors using whole-exome sequencing data", *Bioinformatics,* vol. 31, no. 3, pp. 429-431.

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M.C., Hirai, A., Takahashi, H., Altaf-Ul-Amin, M., Ogasawara, N. & Kanaya, S. 2011, "Sequence-specific error profile of Illumina sequencers", *Nucleic acids research,* vol. 39, no. 13, pp. e90-e90.

Narod, S.A. & Foulkes, W.D. 2004, "BRCA1 and BRCA2: 1994 and beyond", *Nature reviews Cancer,* vol. 4, no. 9, pp. 665-676.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., Muthuswamy, L., Krasnitz, A., McCombie, W.R., Hicks, J. & Wigler, M. 2011, "Tumour evolution inferred by single-cell sequencing", *Nature,* vol. 472, no. 7341, pp. 90-94.

Ng, P.C. & Henikoff, S. 2003, "SIFT: predicting amino acid changes that affect protein function", *Nucleic acids research,* vol. 31, no. 13, pp. 3812-3814.

Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E. and Bamshad, M. 2009, “Targeted capture and massively parallel sequencing of 12 human exomes”, *Nature*, vol. 461, pp.272-276.

Nielsen, R., Paul, J.S., Albrechtsen, A. & Song, Y.S. 2011, "Genotype and SNP calling from next-generation sequencing data", *Nature reviews Genetics,* vol. 12, no. 6, pp. 443-451.

Nordling, C.O. 1953, "A New Theory on the Cancer-inducing Mechanism", *British journal of cancer,* vol. 7, no. 1, pp. 68-72.

Novak, D.J., Chen, L.Q., Ghadirian, P., Hamel, N., Zhang, P., Rossiny, V., Cardinal, G., Robidoux, A., Tonin, P.N., Rousseau, F., Narod, S.A. & Foulkes, W.D. 2008, "Identification of a novel CHEK2 variant and assessment of its contribution to the risk of breast cancer in French Canadian women", *BMC Cancer,* vol. 8, no. 1, pp. 1-8.

Nowell, P. 1976, "The clonal evolution of tumor cell populations", *Science,* vol. 194, no. 4260, pp. 23-28.

Obata, K., Morland, S.J., Watson, R.H., Hitchcock, A., Chenevix-Trench, G., Thomas, E.J. & Campbell, I.G. 1998, "Frequent PTEN/MMAC mutations in endometrioid but not serous or mucinous epithelial ovarian tumors", *Cancer research,* vol. 58, no. 10, pp. 2095-2097.

Oike, T., Ogiwara, H., Tominaga, Y., Ito, K., Ando, O., Tsuta, K., Mizukami, T., Shimada, Y., Isomura, H., Komachi, M., Furuta, K., Watanabe, S., Nakano, T., Yokota, J. & Kohno, T. 2013, "A synthetic lethality–based strategy to treat cancers harboring a genetic deficiency in the chromatin remodeling factor BRG1", *Cancer research,* vol. 73, no. 17, pp. 5508-5518.

Oros, K.K., Leblanc, G., Arcand, S.L., Shen, Z., Perret, C., Mes-Masson, A., Foulkes, W.D., Ghadirian, P., Provencher, D. & Tonin, P.N. 2006, "Haplotype analysis suggest common founders in carriers of the recurrent BRCA2 mutation, 3398delAAAAG, in French Canadian hereditary breast and/ovarian cancer families", *BMC Medical Genetics,* vol. 7, no. 1, pp. 1-7.

Paszkiewicz, K. & Studholme, D.J. 2010, "De novo assembly of short sequence reads", *Briefings in Bioinformatics,* vol. 11, no. 5, pp. 457-472.

Patch, A., Christie, E.L., Etemadmoghadam, D., Garsed, D.W., George, J., Fereday, S., Nones, K., Cowin, P., Alsop, K., Bailey, P.J., Kassahn, K.S., Newell, F., Quinn, M.C.J., Kazakoff, S., Quek, K., Wilhelm-Benartzi, C., Curry, E., Leong, H.S., The Australian Ovarian Cancer,Study Group, Hamilton, A., Mileshkin, L., Au-Yeung, G., Kennedy, C., Hung, J., Chiew, Y., Harnett, P., Friedlander, M., Quinn, M., Pyman, J., Cordner, S., O/'Brien, P., Leditschke, J., Young, G., Strachan, K., Waring, P., Azar, W., Mitchell, C., Traficante, N., Hendley, J., Thorne, H., Shackleton, M., Miller, D.K., Arnau,

G.M., Tothill, R.W., Holloway, T.P., Semple, T., Harliwong, I., Nourse, C., Nourbakhsh, E., Manning, S., Idrisoglu, S., Bruxner, T.J.C., Christ, A.N., Poudel, B., Holmes, O., Anderson, M., Leonard, C., Lonie, A., Hall, N., Wood, S., Taylor, D.F., Xu, Q., Fink, J.L., Waddell, N., Drapkin, R., Stronach, E., Gabra, H., Brown, R., Jewell, A., Nagaraj, S.H., Markham, E., Wilson, P.J., Ellul, J., McNally, O., Doyle, M.A., Vedururu, R., Stewart, C., Lengyel, E., Pearson, J.V., Waddell, N., deFazio, A., Grimmond, S.M. & Bowtell, D.D.L. 2015, "Whole–genome characterization of chemoresistant ovarian cancer", *Nature,* vol. 521, no. 7553, pp. 489-494.

Permuth-Wey, J. & Egan, K. 2009, "Family history is a significant risk factor for pancreatic cancer: results from a systematic review and meta-analysis", *Familial Cancer,* vol. 8, no. 2, pp. 109-117.

Perou, C.M., Sorlie, T., Eisen, M.B., van, d.R., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S.X., Lonning, P.E., Borresen-Dale, A., Brown, P.O. & Botstein, D. 2000, "Molecular portraits of human breast tumours", *Nature,* vol. 406, no. 6797, pp. 747-752.

Pharoah, P.D.P., Day, N.E., Duffy, S., Easton, D.F. & Ponder, B.A.J. 1997, "Family history and the risk of breast cancer: a systematic review and meta-analysis", *International Journal of Cancer,* vol. 71, no. 5, pp. 800-809.

Piek, J., van Diest, P., Zweemer, R.P., Jansen, J.W., Poort-Keesom, R., Menko, F.H., Gille, J., Jongsma, A., Pals, G., Kenemans, P. & Verheijen, R. 2001, "Dysplastic changes in prophylactically removed Fallopian tubes of women predisposed to developing ovarian cancer", *Journal of Pathology,* vol. 195, no. 4, pp. 451-456.

Piskol, R., Ramaswami, G. & Li, J. 2013, "Reliable Identification of Genomic Variants from RNA-Seq Data", *The American Journal of Human Genetics,* vol. 93, no. 4, pp. 641-651.

Quail, M., Smith, M., Coupland, P., Otto, T., Harris, S., Connor, T., Bertoni, A., Swerdlow, H. & Gu, Y. 2012, "A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers", *BMC Genomics,* vol. 13, no. 1, pp. 341.

Quirk, J.T. & Natarajan, N. 2005, "Ovarian cancer incidence in the United States, 1992–1999", *Gynecologic oncology,* vol. 97, no. 2, pp. 519-523.

Rahman, N., Seal, S., Thompson, D., Kelly, P., Renwick, A., Elliott, A., Reid, S., Spanova, K., Barfoot, R., Chagtai, T., Jayatilake, H., McGuffog, L., Hanks, S., Evans, D.G., Eccles, D., The Breast Cancer, S.C., Easton, D.F. & Stratton, M.R. 2006, "PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene", *Nature genetics,* vol. 39, no. 2, pp. 165-167.

Rakha, E., El-Sayed, M., Reis-Filho, J. & Ellis, I. 2009, "Patho-biological aspects of basal-like breast cancer", *Breast cancer research and treatment,* vol. 113, no. 3, pp. 411-422.

Ramos, P., Karnezis, A.N., Craig, D.W., Sekulic, A., Russell, M.L., Hendricks, W.P., Corneveaux, J.J., Barrett, M.T., Shumansky, K. & Yang, Y. 2014, "Small cell carcinoma of the ovary, hypercalcemic type, displays frequent

inactivating germline and somatic mutations in SMARCA4", *Nature genetics,* vol. 46, no. 5, pp. 427-429.

Rapley, E.A., Hockley, S., Warren, W., Johnson, L., Huddart, R., Crockford, G., Forman, D., Leahy, M.G., Oliver, D.T., Tucker, K., Friedlander, M., Phillips, K., Hogg, D., Jewett, M.A.S., Lohynska, R., Daugaard, G., Richard, S., Heidenreich, A., Geczi, L., Bodrogi, I., Olah, E., Ormiston, W.J., Daly, P.A., Looijenga, L.H.J., Guilford, P., Aass, N., Fosså, ,S.D., Heimdal, K., Tjulandin, S.A., Liubchenko, L., Stoll, H., Weber, W., Einhorn, L., Weber, B.L., McMaster, M., Greene, M.H., Bishop, D.T., Easton, D. & Stratton, M.R. 2004, "Somatic mutations of KIT in familial testicular germ cell tumours", *British journal of cancer,* vol. 90, no. 12, pp. 2397-2401.

Reddy, E.P., Reynolds, R.K., Santos, E. & Barbacid, M. 1982, "A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene", *Nature,* vol. 300, no. 5888, pp. 149-152.

Ripperger, T., Gadzicki, D., Meindl, A. & Schlegelberger, B. 2008, "Breast cancer susceptibility: current knowledge and implications for genetic counselling", *European Journal of Human Genetics,* vol. 17, no. 6, pp. 722-731.

Roberts, C.W.M. & Orkin, S.H. 2004, "The SWI/SNF complex mdash] chromatin and cancer", *Nature reviews Cancer,* vol. 4, no. 2, pp. 133-142.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R. & Delaney, A. 2007, "Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing", *Nature methods,* vol. 4, no. 8, pp. 651-657.

Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G. & Mesirov, J.P. 2011, "Integrative genomics viewer", *Nature Biotechnology,* vol. 29, no. 1, pp. 24-26.

Roth, A., Ding, J., Morin, R., Crisan, A., Ha, G., Giuliany, R., Bashashati, A., Hirst, M., Turashvili, G., Oloumi, A., Marra, M.A., Aparicio, S. & Shah, S.P. 2012, "JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data", *Bioinformatics,* vol. 28, no. 7, pp. 907-913.

Rouleau, M., Patel, A., Hendzel, M.J., Kaufmann, S.H. & Poirier, G.G. 2010, "PARP inhibition: PARP1 and beyond", *Nature reviews Cancer,* vol. 10, no. 4, pp. 293-301.

Rowley, J.D. 1973, "A New Consistent Chromosomal Abnormality in Chronic Myelogenous Leukaemia identified by Quinacrine Fluorescence and Giemsa Staining", *Nature,* vol. 243, no. 5405, pp. 290-293.

Samuels, D.C., Han, L., Li, J., Quanghu, S., Clark, T.A., Shyr, Y. & Guo, Y. 2013, "Finding the lost treasures in exome sequencing data", *Trends in genetics,* vol. 29, no. 10, pp. 593-599.

Sangha N., Wu R., Kuick R., Powers S., Mu D., Fiander D., Yuen K., Katabuchi H., Tashiro H., Fearon E. R. & Cho K. R. 2008, "Neurofibromin 1 (NF1) defects are common in human ovarian serous carcinomas and co-occur with TP53 mutations", *Neoplasia,* vol. 101372, pp. 1362-1372.

Santen, G.W.E., Aten, E., Vulto-van Silfhout, A.T., Pottinger, C., van Bon, B.W.M., van Minderhout, I.J.H.M., Snowdowne, R., van, d.L., Boogaard, M., Linssen, M.M.L., Vijfhuizen, L., van, d.W., Vollebregt, M.J.(., the Coffin-Siris consortium, Breuning, M.H., Kriek, M., van Haeringen, A., den Dunnen, J.T., Hoischen, A., Clayton-Smith, J., de Vries, B.B.A., Hennekam, R.C.M., van Belzen, M.J., Almureikhi, M., Baban, A., Barbosa, M., Ben-Omran, T., Berry, K., Bigoni, S., Boute, O., Brueton, L., van, d.B., Canham, N., Chandler, K.E., Chrzanowska, K., Collins, A.L., de Toni, T., Dean, J., den Hollander, N.S., Flore, L.A., Fryer, A., Gardham, A., Graham, J.M., Harrison, V., Horn, D., Jongmans, M.C., Josifova, D., Kant, S.G., Kapoor, S., Kingston, H., Kini, U., Kleefstra, T., Krajewska-Walasek, M., Kramer, N., Maas, S.M., Maciel, P., Mancini, G.M.S., Maystadt, I., McKee, S., Milunsky, J.M., Nampoothiri, S., Newbury-Ecob, R., Nikkel, S.M., Parker, M.J., Pérez-Jurado, L.A., Robertson, S.P., Rooryck, C., Shears, D., Silengo, M., Singh, A., Smigiel, R., Soares, G., Splitt, M., Stewart, H., Sweeney, E., Tassabehji, M., Tuysuz, B., van Eerde, A.M., Vincent-Delorme, C., Wilson, L.C. & Yesil, G. 2013, "Coffin–Siris Syndrome and the BAF Complex: Genotype–Phenotype Study in 63 Patients", *Human mutation,* vol. 34, no. 11, pp. 1519-1528.

Sathirapongsasuti, J.F., Lee, H., Horst, B., Brunner, G., Cochran, A.J., Binder, S., Quackenbush, J. & Nelson, S.F. 2011, "Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV", *Bioinformatics,* vol. 27, no. 19, pp. 2648-2654.

Saunders, C.T., Wong, W.S.W., Swamy, S., Becq, J., Murray, L.J. & Cheetham, R.K. 2012, "Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs", *Bioinformatics,* vol. 28, no. 14, pp. 1811-1817.

Schneppenheim, R., FrÃ¼hwald, M.C., Gesk, S., Hasselblatt, M., Jeibmann, A., Kordes, U., Kreuz, M., Leuschner, I., Subero, J.I.M., Obser, T., Oyen, F., Vater, I. & Siebert, R. 2010, "Germline nonsense mutation and somatic inactivation of SMARCA4/BRG1 in a family with rhabdoid tumor predisposition syndrome", *The American Journal of Human Genetics,* vol. 86, no. 2, pp. 279-284.

Schnitt, S.J. 2010, "Classification and prognosis of invasive breast cancer: from morphology to molecular taxonomy", *Modern pathology,* vol. 23, pp. S60-S64.

Schwartzentruber, J., Korshunov, A., Liu, X., Jones, D.T., Pfaff, E., Jacob, K., Sturm, D., Fontebasso, A.M., Quang, D.K. & Tönjes, M. 2012, "Driver mutations in histone H3. 3 and chromatin remodelling genes in paediatric glioblastoma", *Nature,* vol. 482, no. 7384, pp. 226-231.

Schwarz, R.F., Ng, C.K., Cooke, S.L., Newman, S., Temple, J., Piskorz, A.M., Gale, D., Sayal, K., Murtaza, M. & Baldwin, P.J. 2015, "Spatial and temporal heterogeneity in high-grade serous ovarian cancer: a phylogenetic analysis", *PLoS Med,* vol. 12, no. 2, pp. e1001789.

Schwarz, J.M., Rodelsperger, C., Schuelke, M. & Seelow, D. 2010, "MutationTaster evaluates disease-causing potential of sequence alterations", *Nature Methods,* vol. 7, no. 8, pp. 575-576.

Scully, R.E. 1987, "Classification of human ovarian tumors", *Environmental health perspectives,* vol. 73, pp. 15-24.

Scully, R. & Livingston, D.M. 2000, "In search of the tumour-suppressor functions of BRCA1 and BRCA2", *Nature,* vol. 408, no. 6811, pp. 429-432.

Shah, S.P., Köbel, M., Senz, J., Morin, R.D., Clarke, B.A., Wiegand, K.C., Leung, G., Zayed, A., Mehl, E., Kalloger, S.E., Sun, M., Giuliany, R., Yorida, E., Jones, S., Varhol, R., Swenerton, K.D., Miller, D., Clement, P.B., Crane, C., Madore, J., Provencher, D., Leung, P., DeFazio, A., Khattra, J., Turashvili, G., Zhao, Y., Zeng, T., Glover, J.N.M., Vanderhyden, B., Zhao, C., Parkinson, C.A., Jimenez-Linan, M., Bowtell, D.D.L., Mes-Masson, A., Brenton, J.D., Aparicio, S.A., Boyd, N., Hirst, M., Gilks, C.B., Marra, M. & Huntsman, D.G. 2009a, "Mutation of FOXL2 in Granulosa-Cell Tumors of the Ovary", *New England Journal Med,* vol. 360, no. 26, pp. 2719-2729.

Shah, S.P., Morin, R.D., Khattra, J., Prentice, L., Pugh, T., Burleigh, A., Delaney, A., Gelmon, K., Guliany, R., Senz, J., Steidl, C., Holt, R.A., Jones, S., Sun, M., Leung, G., Moore, R., Severson, T., Taylor, G.A., Teschendorff, A.E., Tse, K., Turashvili, G., Varhol, R., Warren, R.L., Watson, P., Zhao, Y., Caldas, C., Huntsman, D., Hirst, M., Marra, M.A. & Aparicio, S. 2009b, "Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution", *Nature,* vol. 461, no. 7265, pp. 809-813.

Sharma, S., Stumpo, D.J., Balajee, A.S., Bock, C.B., Lansdorp, P.M., Brosh, R.M. & Blackshear, P.J. 2006, "RECQL, a Member of the RecQ Family of DNA Helicases, Suppresses Chromosomal Instability", *Molecular and cellular biology,* vol. 27, no. 5, pp. 1784-1794.

Shi, Y. & Majewski, J. 2013, "FishingCNV: a graphical software package for detecting rare copy number variations in exome sequencing data", *Bioinformatics,* vol.29, no. 11, pp.1461-1462.

Siegel, R., Ma, J., Zou, Z. & Jemal, A. 2014, "Cancer statistics, 2014", *CA: A Cancer Journal for Clinicians,* vol. 64, no. 1, pp. 9-29.

Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. 2005, "ROCR: visualizing classifier performance in R", *Bioinformatics,* vol. 21, no. 20, pp. 3940-3941.

Smith, M.E., Cimica, V., Chinni, S., Jana, S., Koba, W., Yang, Z., Fine, E., Zagzag, D., Montagna, C. & Kalpana, G.V. 2011, "Therapeutically targeting cyclin D1 in primary tumors arising from loss of Ini1", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 108, no. 1, pp. 319-324.

Smith, M.J., O'Sullivan, J., Bhaskar, S.S., Hadfield, K.D., Poke, G., Caird, J., Sharif, S., Eccles, D., Fitzpatrick, D., Rawluk, D., du Plessis, D., Newman, W.G. & Evans, D.G. 2013, "Loss-of-function mutations in SMARCE1 cause an inherited disorder of multiple spinal meningiomas", *Nature genetics,* vol. 45, no. 3, pp. 295-298.

Smith, M., Wallace, A., Bowers, N., Rustad, C., Woods, C.G., Leschziner, G., Ferner, R. & Evans, D.G. 2012, "Frequency of SMARCB1 mutations in familial and sporadic schwannomatosis", *Neurogenetics,* vol. 13, no. 2, pp. 141-145.

Sotiriou, C., Neo, S.Y., McShane, L.M., Korn, E.L., Long, P.M., Jazaeri, A., Martiat, P., Fox, S.B., Harris, A.L. and Liu, E.T., 2003, “Breast cancer classification and prognosis based on gene expression profiles from a population-based study“, *Proceedings of the National Academy of Sciences*, vol. 100, no. 18, pp.10393-10398.

Song, H., Cicek, M.S., Dicks, E., Harrington, P., Ramus, S.J., Cunningham, J.M., Fridley, B.L., Tyrer, J.P., Alsop, J., Jimenez-Linan, M., Gayther, S.A., Goode, E.L. & Pharoah, P.D.P. 2014, "The contribution of deleterious germline mutations in BRCA1, BRCA2 and the mismatch repair genes to ovarian cancer in the population", *Human molecular genetics,* vol. 23, no. 17, pp. 4703-4709.

Spencer, D.H., Sehn, J.K., Abel, H.J., Watson, M.A., Pfeifer, J.D. & Duncavage, E.J. 2013, "Comparison of clinical targeted next-generation sequence data from formalin-fixed and fresh-frozen tissue specimens", *The Journal of Molecular Diagnostics,* vol. 15, no. 5, pp. 623-633.

Stewart, C.J.R., Alexiadis, M., Crook, M.L. & Fuller, P.J. 2013, "An immunohistochemical and molecular analysis of problematic and unclassified ovarian sex cord–stromal tumors", *Human pathology,* vol. 44, no. 12, pp. 2774-2781.

Stratton, J.F., Pharoah, P., Smith, S.K., Easton, D. & Ponder, B.A. 1998, "A systematic review and meta‐analysis of family history and risk of ovarian cancer", *British Journal of Obstetrics & Gynaecology,* vol. 105, no. 5, pp. 493-499.

Stratton, M.R., Campbell, P.J. & Futreal, P.A. 2009, "The cancer genome", *Nature,* vol. 458, no. 7239, pp. 719-724.

Struewing, J.P., Abeliovich, D., Peretz, T., Avishai, N., Kaback, M.M., Collins, F.S. & Brody, L.C. 1995, "The carrier frequency of the BRCA1 185delAG mutation is approximately 1 percent in Ashkenazi Jewish individuals", *Nature genetics,* vol. 11, no. 2, pp. 198-200.

Sun, A., Tawfik, O., Gayed, B., Thrasher, J.B., Hoestje, S., Li, C. & Li, B. 2007, "Aberrant expression of SWI/SNF catalytic subunits BRG1/BRM is associated with tumor development and increased invasiveness in prostate cancers", *The Prostate,* vol. 67, no. 2, pp. 203-213.

Sun, J., Wang, Y., Xia, Y., Xu, Y., Ouyang, T., Li, J., Wang, T., Fan, Z., Fan, T., Lin, B., Lou, H. & Xie, Y. 2015, "Mutations in *RECQL* gene are sssociated with predisposition to creast cancer", *PLoS Genetics,* vol. 11, no. 5, pp. e1005228.

Tabin, C.J., Bradley, S.M., Bargmann, C.I., Weinberg, R.A., Papageorge, A.G., Scolnick, E.M., Dhar, R., Lowy, D.R. & Chang, E.H. 1982, "Mechanism of activation of a human oncogene", *Nature,* vol. 300, no. 5888, pp. 143-149.

Tabor, H.K., Risch, N.J. & Myers, R.M. 2002, "Candidate-gene approaches for studying complex genetic traits: practical considerations", *Nature reviews Genetics,* vol. 3, no. 5, pp. 391-397.

The Cancer Genome Atlas Research Network 2011, "Integrated genomic analyses of ovarian carcinoma", *Nature,* vol. 474, no. 7353, pp. 609-615.

The ENCODE Project Consortium 2012, "An integrated encyclopedia of DNA elements in the human genome", *Nature,* vol. 489, no. 7414, pp. 57-74.

Thompson, E.R., Doyle, M.A., Ryland, G.L., Rowley, S.M., Choong, D.Y., Tothill, R.W., Thorne, H., kConFab, B.D., Li, J. & Ellul, J. 2012, "Exome sequencing identifies rare deleterious mutations in DNA repair genes FANCC and BLM as potential breast cancer susceptibility alleles", *PLoS Genetics,* vol. 8, no. 9, pp. e1002894.

Tischkowitz, M., Sabbaghian, N., Hamel, N., Pouchet, C., Foulkes, W.D., Mes-Masson, A., Provencher, D.M. & Tonin, P.N. 2013, "Contribution of the PALB2 c.2323C>T p.Q775X] Founder mutation in well-defined breast and/or ovarian cancer families and unselected ovarian cancer cases of French Canadian descent", *BMC Medical Genetics,* vol. 14, no. 1, pp. 1-7.

Tonin, P.N., Mes-Masson, A., Futreal, P.A., Morgan, K., Mahon, M., Foulkes, W.D., Cole, D.E.C., Provencher, D., Ghadirian, P. & Narod, S.A. 1998, "Founder BRCA1 and BRCA2 Mutations in French Canadian Breast and Ovarian Cancer Families", *The American Journal of Human Genetics,* vol. 63, no. 5, pp. 1341-1351.

Trapnell, C. & Salzberg, S.L. 2009, "How to map billions of short reads onto genomes", *Nature biotechnology,* vol. 27, no. 5, pp. 455-457.

Van Allen, E.,M., Wagle, N., Stojanov, P., Perrin, D.L., Cibulskis, K., Marlow, S., Jane-Valbuena, J., Friedrich, D.C., Kryukov, G., Carter, S.L., McKenna, A., Sivachenko, A., Rosenberg, M., Kiezun, A., Voet, D., Lawrence, M., Lichtenstein, L.T., Gentry, J.G., Huang, F.W., Fostel, J., Farlow, D., Barbie, D., Gandhi, L., Lander, E.S., Gray, S.W., Joffe, S., Janne, P., Garber, J., MacConaill, L., Lindeman, N., Rollins, B., Kantoff, P., Fisher, S.A., Gabriel, S., Getz, G. & Garraway, L.A. 2014, "Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine", *Nature medicine,* vol. 20, no. 6, pp. 682-688.

Vaughan, S., Coward, J.I., Bast, R.C., Berchuck, A., Berek, J.S., Brenton, J.D., Coukos, G., Crum, C.C., Drapkin, R., Etemadmoghadam, D., Friedlander, M., Gabra, H., Kaye, S.B., Lord, C.J., Lengyel, E., Levine, D.A., McNeish, I.A., Menon, U., Mills, G.B., Nephew, K.P., Oza, A.M., Sood, A.K., Stronach, E.A., Walczak, H., Bowtell, D.D. & Balkwill, F.R. 2011, "Rethinking ovarian cancer: recommendations for improving outcomes", *Nature Reviews Cancer,* vol. 11, no. 10, pp. 719-725.

Venkatraman, E.S. & Olshen, A.B. 2007, "A faster circular binary segmentation algorithm for the analysis of array CGH data", *Bioinformatics,* vol. 23, no. 6, pp. 657-663.

Venkitaraman, A.R. 2004, "Tracing the network connecting brca and fanconi anaemia proteins", *Nat Rev Cancer,* vol. 4, no. 4, pp. 266-276.

Venkitaraman, A.R. 2002, “Cancer susceptibility and the functions of BRCA1 and BRCA2”, *Cell*, vol. 108, no. 2, pp. 171-182.

Vézina, H., Durocher, F., Dumont, M., Houde, L., Szabo, C., Tranchant, M., Chiquette, J., Plante, M., Laframboise, R., Lépine, J., Nevanlinna, H., Stoppa-Lyonnet, D., Goldgar, D., Bridge, P. & Simard, J. 2005, "Molecular and genealogical characterization of the R1443X BRCA1 mutation in high-

risk French-Canadian breast/ovarian cancer families", *Human genetics,* vol. 117, no. 2-3, pp. 119-132.

Walsh, C.S., 2015, “Two decades beyond BRCA1/2: homologous recombination, hereditary cancer risk and a target for ovarian cancer therapy”, *Gynecologic oncology*, vol. 137, no. 2, pp.343-350.

Walsh, T., Casadei, S., Lee, M.K., Pennil, C.C., Nord, A.S., Thornton, A.M., Roeb, W., Agnew, K.J., Stray, S.M., Wickramanayake, A., Norquist, B., Pennington, K.P., Garcia, R.L., King, M. & Swisher, E.M. 2011, "Mutations in 12 genes for inherited ovarian, fallopian tube, and peritoneal carcinoma identified by massively parallel sequencing", *Proceedings of the National Academy of Sciences,* vol. 108, no. 44, pp. 18032-18037.

Wang, Y., Klijn, J.G., Zhang, Y., Sieuwerts, A.M., Look, M.P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M.E. & Yu, J. 2005, "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer", *The Lancet,* vol. 365, no. 9460, pp. 671-679.

Wang, K., Li, M. & Hakonarson, H. 2010, "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data", *Nucleic Acids Res,* vol. 38, no. 16, pp. e164-e164.

Wang, Q., Jia, P., Li, F., Chen, H., Ji, H., Hucks, D., Dahlman, K., Pao, W. & Zhao, Z. 2013, "Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers", *Genome Medicine,* vol. 5, no. 10, pp. 91.

Wang, W. & Lai, Y. 2014, "Molecular pathogenesis in granulosa cell tumor is not only due to somatic FOXL2 mutation", *Journal of Ovarian Research,* vol. 7, pp. 88.

Wetmore, C., Bendel, A. & Gajjar, A. 2014, "Activity of alisertib (MLN8237) as single agent in recurrent atypical teratoid rhabdoid tumor (AT/RT) in four children: a single patient treatment plan pilot study", *Cancer Genetics,* vol. 207, no. 9, pp. 453.

Whittemore, A.S. 1994, "Characteristics Relating to Ovarian Cancer Risk: Implications for Prevention and Detection", *Gynecologic oncology,* vol. 55, no. 3, pp. S15-S19.

Williams, C., Pontén, F., Moberg, C., Söderkvist, P., Uhlén, M., Pontén, J., Sitbon, G. & Lundeberg, J. 1999, "A high frequency of sequence alterations is due to formalin fixation of archival specimens", *The American Journal of Pathology,* vol. 155, no. 5, pp. 1467-1471.

Wilm, A., Aw, P.P.K., Bertrand, D., Yeo, G.H.T., Ong, S.H., Wong, C.H., Khor, C.C., Petric, R., Hibberd, M.L. & Nagarajan, N. 2012, "LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets", *Nucleic acids research,* vol. 40, no. 22, pp. 11189-11201.

Witkowski, L., Mattina, J., Schönberger, S., Murray, M.J., Huntsman, D.G., Reis-Filho, J., McCluggage, W.G., Nicholson, J.C., Coleman, N., Calaminus, G., Schneider, D.T., Arseneau, J., Stewart, C.J.R. & Foulkes, W.D. 2013,

"DICER1 hotspot mutations in non-epithelial gonadal tumours", *British journal of cancer,* vol. 109, no. 10, pp. 2744-2750.

Witkowski, L., Carrot-Zhang, J., Albrecht, S., Fahiminiya, S., Hamel, N., Tomiak, E., Grynspan, D., Saloustros, E., Nadaf, J., Rivera, B., Gilpin, C., Castellsague, E., Silva-Smith, R., Plourde, F., Wu, M., Saskin, A., Arseneault, M., Karabakhtsian, R.G., Reilly, E.A., Ueland, F.R., Margiolaki, A., Pavlakis, K., Castellino, S.M., Lamovec, J., Mackay, H.J., Roth, L.M., Ulbright, T.M., Bender, T.A., Georgoulias, V., Longy, M., Berchuck, A., Tischkowitz, M., Nagel, I., Siebert, R., Stewart, C.J.R., Arseneau, J., McCluggage, W.G., Clarke, B.A., Riazalhosseini, Y., Hasselblatt, M., Majewski, J. & Foulkes, W.D. 2014, "Germline and somatic SMARCA4 mutations characterize small cell carcinoma of the ovary, hypercalcemic type", *Nature genetics,* vol. 46, no. 5, pp. 438-443.

Witkowski, L. & Foulkes, W.D. 2015, "In Brief: Picturing the complex world of chromatin remodelling families", *The Journal of pathology,* vol.237, no. 4, pp. 403-406.

Witkowski, L., Lalonde, E., Zhang, J., Albrecht, S., Hamel, N., Cavallone, L., May, S.T., Nicholson, J.C., Coleman, N., Murray, M.J., Tauber, P.F., Huntsman, D.G., Schönberger, S., Yandell, D., Hasselblatt, M., Tischkowitz, M.D., Majewski, J. & Foulkes, W.D. 2013, "Familial rhabdoid tumour 'avant la lettre'—from pathology review to exome sequencing and back again", *The Journal of pathology,* vol. 231, no. 1, pp. 35-43.

Wong, A.K.C., Shanahan, F., Chen, Y., Lian, L., Ha, P., Hendricks, K., Ghaffari, S., Iliev, D., Penn, B., Woodland, A., Smith, R., Salada, G., Carillo, A., Laity, K., Gupte, J., Swedlund, B., Tavtigian, S.V., Teng, D.H. & Lees, E. 2000, "BRG1, a component of the SWI-SNF complex, Is mutated in multiple human tumor cell lines", *Cancer research,* vol. 60, no. 21, pp. 6171-6177.

Wu, Y. & Brosh, R.M. 2010, "Distinct roles of RECQ1 in the maintenance of genomic stability", *DNA repair,* vol. 9, no. 3, pp. 315-324.

Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., Kamiyama, M., Hruban, R.H., Eshleman, J.R., Nowak, M.A., Velculescu, V.E., Kinzler, K.W., Vogelstein, B. & Iacobuzio-Donahue, C. 2010, "Distant metastasis occurs late during the genetic evolution of pancreatic cancer", *Nature,* vol. 467, no. 7319, pp. 1114-1117.

Yan, X., Xu, J., Gu, Z., Pan, C., Lu, G., Shen, Y., Shi, J., Zhu, Y., Tang, L., Zhang, X., Liang, W., Mi, J., Song, H., Li, K., Chen, Z. & Chen, S. 2011, "Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia", *Nature genetics,* vol. 43, no. 4, pp. 309-315.

Yates, L.R. & Campbell, P.J. 2012, "Evolution of the cancer genome", *Nature reviews Genetics,* vol. 13, no. 11, pp. 795-806.

Yost, S.E., Alakus, H., Matsui, H., Schwab, R.B., Jepsen, K., Frazer, K.A. & Harismendy, O. 2013, "Mutascope: sensitive detection of somatic mutations from deep amplicon sequencing", *Bioinformatics,* vol. 29, no. 15, pp. 1908-1909.

Young, R.H., Oliva, E. & Scully, R.E. 1994, "Small cell carcinoma of the ovary, hypercalcemic type. A clinicopathological analysis of 150 cases", *The American Journal of Surgical Pathology,* vol. 18, no. 11, pp. 1102-1116.

Zhang, J., Shi, Y., Lalonde, E., Li, L., Cavallone, L., Ferenczy, A., Gotlieb, W., Foulkes, W. & Majewski, J. 2013, "Exome profiling of primary, metastatic and recurrent ovarian carcinomas in a BRCA1-positive patient", *BMC Cancer,* vol. 13, no. 1, pp. 146.

Zhang, S., Royer, R., Li, S., McLaughlin, J.R., Rosen, B., Risch, H.A., Fan, I., Bradley, L., Shaw, P.A. & Narod, S.A. 2011, "Frequencies of BRCA1 and BRCA2 mutations among 1,342 unselected patients with invasive ovarian cancer", *Gynecologic Oncology,* vol. 121, no. 2, pp. 353-357.

Zhao, S.G., Evans, J.R., Kothari, V., Sun, G., Larm, A., Mondine, V., Schaeffer, E.M., Ross, A.E., Klein, E.A., Den, R.B., Dicker, A.P., Karnes, R.J., Erho, N., Nguyen, P.L., Davicioni, E. & Feng, F.Y. 2015, "The landscape of prognostic outlier genes in high-risk prostate cancer", *Clinical cancer research,* doi:10.1158/1078-0432.CCR-15-1250.

Zhou, B., Yuan, T., Liu, M., Liu, H., Xie, J., Shen, Y. & Chen, P. 2012, "Overexpression of the structural maintenance of chromosome 4 protein is associated with tumor de-differentiation, advanced stage and vascular invasion of primary liver cancer", *Oncology reports,* vol. 28, no. 4, pp. 1263-1268.