

AMERICAN UNIVERSITY OF BEIRUT

A TRIO APPROACH USING WHOLE EXOME SEQUENCING
TO DISCOVER SHARED MUTATIONS IN TWO FIRST-
DEGREE COUSINS WITH RHABDOMYOSARCOMA

by
AMAL MOHAMMAD HALWANI

A thesis
submitted in partial fulfillment of the requirements
for the degree of Master of Science
to the Department of Biochemistry and Molecular Genetics
of the Faculty of Medicine
at the American University of Beirut

Beirut, Lebanon
July 2020

AMERICAN UNIVERSITY OF BEIRUT

A TRIO APPROACH USING WHOLE EXOME SEQUENCING
TO DISCOVER SHARED MUTATIONS IN TWO FIRST-
DEGREE COUSINS WITH RHABDOMYOSARCOMA

by
AMAL MOHAMMAD HALWANI

Approved by:

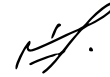
[Signature]

Dr. Pierre Khoueiry, Assistant Professor, PhD (Bioinformatics, Structural Biology and Genomics)

Dept. of Biochemistry and Molecular Genetics

[Signature]

Advisor



Dr. Nadine Darwiche, Professor, PhD (Biochemistry and Molecular Biology)

Dept. of Dept. of Biochemistry and Molecular Genetics

Member of Committee

Nadine Darwiche

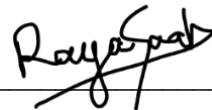
[Signature]

Dr. Raya Saab, Associate Professor, MD (Board Certified Pediatrics and Pediatric Hematology/Oncology)

Dept. of Pediatrics and Adolescent Medicine

Member of Committee

[Signature]



Date of thesis defense: August 25, 2020

AMERICAN UNIVERSITY OF BEIRUT

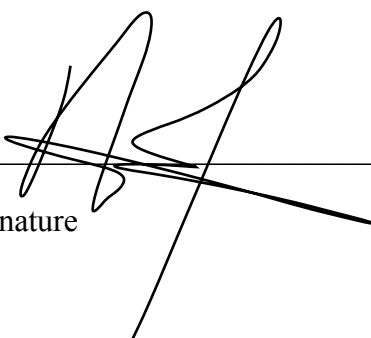
THESIS RELEASE FORM

Student Name: _____ Halwani _____ Amal _____ Mohammad _____
Last First Middle

I authorize the American University of Beirut, to: (a) reproduce hard or electronic copies of my thesis; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes:

- As of the date of submission
- One year from the date of submission of my thesis.
- Two years from the date of submission of my thesis.
- Three years from the date of submission of my thesis.

*(to check a box above, right click on the box, choose properties, then select "checked".
DELETE THIS WHEN DONE)*



Signature

February 8, 2021 _____
Date

ACKNOWLEDGEMENTS

I would like to thank every person who has contributed in this work and who has helped me in accomplishing it. I really appreciate your guidance, patience and support throughout this journey.

A million thanks...

ABSTRACT OF THE THESIS OF

Amal Mohammad Halwani

for

Master of Science
Major: Biochemistry

Title: A Trio Approach Using Whole Exome Sequencing To Discover Shared Mutations In Two First-Degree Cousins With Rhabdomyosarcoma

Rhabdomyosarcoma (RMS) is a rare soft tissue sarcoma that arises in cells committed to the myogenic lineage but unable to achieve full differentiation. It occurs mainly in children and younger adults and is histologically classified into mainly the embryonal subtype (ERMS) and the alveolar subtype (ARMS). ARMS accounts for almost 30% of RMS cases overall, yet it is more aggressive and associated with worse prognosis due to two characteristic translocations of the FOXO1 gene with either the PAX3 or PAX7 genes. Such cases are described as fusion positive (FP).

Although the vast majority of cases arise sporadically, almost 10% of children or adolescents diagnosed with RMS are considered genetically predisposed to develop this type of sarcoma. While a number of cancer predisposition syndromes have been previously reported in the literature as increasing the risk of RMS (e.g.: Li-Fraumeni Syndrome, Costello Syndrome...), many others are yet to be discovered. These syndromes are usually the result of point mutations inherited in an autosomal dominant manner.

Although rare, familial cases of RMS can greatly inform about the biology and genetic basis of the sporadic disease, and this clears the floor for the discovery of new genes and/or gene variants associated with RMS specifically and cancer generally. Such a case has been recently reported whereby two first-degree cousins from a consanguineous family have been diagnosed with FP-ARMS at an early age. In order to identify the potential inherited single nucleotide variant(s) responsible for the RMS tumor growth in these two relatives, if any, whole-exome sequencing was performed on two trios each comprised of the patient and their respective parents.

Sequencing outcome resulted in a large number of variants that were filtered according to a number of independent variables in order to narrow down the possible hits. Among forty-seven variants of interest, three were of particular relevance and were highly suspected of being responsible for the development of RMS in the two pediatric patients; the FANCL gene variant (rs62020347), the BRCA1 variant (rs1799950) and the NUMBL variant (rs536916726). Further research must be carried on each variant independently in order to determine its specific role in the etiology of RMS.

TABLE OF CONTENTS

| | |
|---|----|
| ACKNOWLEDGEMENTS | 1 |
| ABSTRACT..... | 2 |
| ILLUSTRATIONS..... | 5 |
| TABLES..... | 6 |
| ABBREVIATIONS | 7 |
| INTRODUCTION | 8 |
| A. Rhabdomyosarcoma..... | 8 |
| B. Molecular Diagnostic Tools..... | 12 |
| C. DNA Sequencing Timeline | 13 |
| D. NGS Breakthrough and Workflow | 14 |
| E. Depth and Breadth of Coverage | 15 |
| F. Whole Exome Sequencing..... | 16 |
| G. Population-Based Databases and Trio-WES Analyses..... | 18 |
| HYPOTHESIS, AIMS, AND SIGNIFICANCE..... | 20 |
| MATERIALS AND METHODS..... | 23 |
| A. Sample Collection..... | 23 |
| B. Sample Evaluation using Whole Exome Sequencing | 23 |
| C. Identification of Common Variants Using Bioinformatics Analysis..... | 24 |

| | |
|---|-----------|
| D. Variants Search in Cancer Databases | 25 |
| E. Pedigree Construction | 25 |
| RESULTS | 26 |
| A. Quality Checks of Sequenced Reads | 26 |
| B. Whole Exome Sequencing on the Two Trios | 27 |
| C. Variant Calling on The Six Samples | 29 |
| D. Variant Filtering | 30 |
| E. FANCL and BRCA1 gene variants as highly likely candidates based on gene function | 33 |
| F. Variants common to patients and carrier parents as eligible causative candidates | 34 |
| G. Variants with unknown clinical significance are more probable to be associated with tumor predisposition as compared to benign/likely benign variants | 37 |
| H. The identification of low allele frequency variants highlights four major gene candidates: RGSL1, CDKAL1, NLRX1 and R3HDM2 | 38 |
| I. Further research is required to study the tumorigenic effect of the filtered variants that have not been previously associated with cancer | 41 |
| DISCUSSION AND LIMITATIONS | 43 |
| CONCLUSION AND FUTURE PROSPECTS | 47 |
| REFERENCES | 50 |

ILLUSTRATIONS

Figure

| | |
|--|----|
| 1. PolyPhen-2 and SIFT scores..... | 12 |
| 2. NGS workflow and stepwise timeline..... | 15 |
| 3. Depth and Breadth of Coverage in NGS..... | 16 |
| 4. Variant calling pipeline..... | 18 |
| 5. Family pedigree..... | 21 |
| 6. Sequence quality histogram..... | 26 |
| 7. Proportions of TSG variants, DRG variants and others..... | 34 |
| 8. Variants common to patients and parents..... | 35 |
| 9. Pathogenicity of the 47 variants (according to Varsome)..... | 38 |
| 10. Association of the 47 variants with different types of cancer..... | 42 |

TABLES

Table

| | |
|---|-------|
| 1. Selected heritable cancer predisposition syndromes associated with RMS.. | 9-10 |
| 2. Read counts..... | 28-29 |
| 3. SNPs, INDELS, and depth of read of each sample..... | 29-30 |
| 4. The 47 variants..... | 32 |
| 5. Variants shared with parents..... | 36 |
| 6. Variants' allele frequencies..... | 39-40 |

ABBREVIATIONS

| | |
|-------|--|
| ARMS | alveolar rhabdomyosarcoma |
| DGGE | denaturing gradient gel electrophoresis |
| ERMS | embryonal rhabdomyosarcoma |
| FA | fanconi anemia |
| FP | fusion-positive |
| INDEL | insertion or deletion |
| MAF | maximum allele frequency |
| MLPA | multiplex ligation-dependent probe amplification |
| NGS | next-generation sequencing |
| PCR | polymerase chain reaction |
| RFLP | restriction fragment length polymorphism |
| RMS | rhabdomyosarcoma |
| SNP | single nucleotide polymorphism |
| SNV | single nucleotide variant |
| SOLiD | sequencing by oligo ligation detection |
| SSCP | single strand conformational polymorphism |
| VUS | variant of unknown significance |
| WES | whole exome sequencing |

CHAPTER I

INTRODUCTION

A. Rhabdomyosarcoma

Rhabdomyosarcoma (RMS) is a rare, malignant soft tissue sarcoma with a certain degree of skeletal muscle differentiation (Cortes Barrantes et al., 2019). It is the most prevalent soft tissue tumor in children and adolescents as it accounts for almost 5% of all pediatric tumors (Egas Bejar and Huh, 2014) with a yearly incidence rate of 4.3 cases per million children (Hassan et al., 2014). Primary tumor sites include either skeletal muscle tissue or hollow organs of the head and neck, extremities and genitourinary tract (Dasgupta et al., 2016). RMS can be divided into two major histopathological subtypes: the embryonal (ERMS) and alveolar (ARMS) tumors. ERMS accounts for almost 70% of RMS cases and is usually diagnosed in children below the age of 10, while ARMS constitutes around 30% of tumors and affects young adults more frequently (Owosho et al., 2016; Roberts and MacDuff, 2018). The alveolar subtype is associated with worse prognosis due to genetic translocations involving important developmental transcription factors and such cases are described as fusion-positive (FP) (Cortes Barrantes et al., 2019).

The majority of RMS cases occur sporadically, which means that they arise in a temporally-irregular and scattered manner. However, around 10% of the manifestations are accounted for by predisposing genetic disorders or familial syndromes inherited primarily in an autosomal dominant fashion. In fact, autopsies of deceased children and adolescents with RMS revealed at least one congenital anomaly in around 30% of the

individuals. The most common abnormalities arose in the gastrointestinal tract, the genitourinary tract as well as the central nervous system (Ruymann et al., 1988). Genetic disorders linked to RMS include neurofibromatosis type I, Li-Fraumeni syndrome and Beckwith-Wiedemann syndrome among others (Table 1). Among Li-Fraumeni syndrome families, RMS was the most commonly detected pediatric cancer (Egas Bejar and Huh, 2014). Breakthroughs in the fields of molecular biology and bioinformatics paved the way for a greater understanding of the genetic and clinical foundations underlying each of these syndromes. Nevertheless, specific protocols to allow accurate assessments of these familial or cancer predisposition syndromes in the management of sarcoma patients are still lacking (Farid and Ngeow, 2016).

| Inherited syndrome | Inheritance | Genes | Chief clinical features | Associated sarcomas |
|---|--------------------|--|--|-----------------------------|
| Beckwith-Wiedemann | Sporadic/AD | <i>CDKA1C</i> , <i>KCNQ10</i> , <i>T1</i> , <i>LIT1</i> , <i>IGF2</i> , and <i>H19</i> | Overgrowth syndrome: macroglossia, omphalocele, hemihypertrophy, gigantism | Embryonal RMS |
| Bloom | AR | <i>RECQL3</i> on <i>15q26.1</i> | Progeroid syndrome: growth retardation, sun sensitivity, telangiectasias, and other skin changes | Osteosarcoma, embryonal RMS |
| Constitutional mismatch repair syndrome | AR | <i>PSM2</i> at <i>7p/q22.1</i> | Predisposition to hematologic malignancies, CNS tumors, gastrointestinal tumors and polyps, and other embryonic tumors | Embryonal RMS |
| Costello | AD | <i>HRAS</i> at <i>11q15</i> / <i>12p12.1</i> | RASopathy: coarse facies, short stature, cardiac anomalies, developmental delay, and congenital myopathy | Embryonal RMS |
| Familial pleuropulmonary blastoma (DICER1 syndrome) | AD | <i>DICER1</i> at <i>14q23.13</i> | Predisposition to pleuropulmonary blastomas and other dysplastic/malignant lesions | Embryonal RMS |

| | | | | |
|--|----|--|--|--|
| Gorlin syndrome/nevoid basal cell carcinoma syndrome | AD | <i>PTCH</i> at <i>Xp11.23/9q22</i> | Multiple basal cell carcinomas, odontogenic keratocysts, palmar/plantar pits, calcification of the falx cerebri, rib abnormalities | Embryonal RMS |
| LFS | AD | <i>TP53</i> at <i>17p13.1</i> , <i>CHEK2</i> at <i>22q12</i> | Predisposition to early onset of multiple cancers, most commonly premenopausal breast cancer, STS, CNS tumors, osteosarcomas, adrenocortical carcinomas, and leukemias | Osteosarcomas, RMS, STS |
| Mosaic variegated aneuploidy | AR | <i>BUB1B</i> at <i>15q15</i> | Intrauterine growth restriction, microcephaly, predisposition to cancer (Wilms tumor, hematologic malignancies) | Embryonal RMS |
| NF1 | AD | <i>NF1</i> at <i>17q11.2</i> | Café-au-lait spots, neurofibromas, iris hamartomas (Lisch nodules), optic gliomas, skeletal abnormalities | MPNST, GIST, RMS |
| Nijmegen breakage syndrome | AR | <i>NBS1</i> at <i>8q21.3</i> | Chromosomal instability syndrome associated with microcephaly, growth retardation, immunodeficiency, and tumor predisposition | Embryonal RMS |
| Noonan syndrome | AD | <i>PTPN11</i> at <i>12q24</i> , <i>SOS1</i> at <i>2-22</i> | RASopathy associated with dysmorphic facies, short stature, neck webbing, cardiac anomalies, deafness, and bleeding diathesis | Embryonal RMS, giant cell tumor of bone, granular cell tumor, PVNS |
| Rubinstein-Taybi | AD | <i>CREBBP</i> at <i>16p13.1</i> | Multiple congenital anomalies, developmental delay, microcephaly, and dysmorphic features | Embryonal RMS, LMS |
| Werner | AR | <i>WRN</i> at <i>8p11.2-12</i> | Progeroid syndrome with tight atrophic skin and bird-like facies, early onset atherosclerosis, diabetes, and osteoporosis | Osteosarcoma, embryonal RMS |

Table 1. Selected heritable cancer predisposition syndromes associated with RMS.

This table lists the cancer predisposing syndromes related to RMS. The columns from left to right represent the inherited syndrome, the inheritance pattern, the affected genes, the chief clinical features as well as the associated sarcomas. *Abbreviations: AD, autosomal dominant; APC, adenomatous polyposis coli; AR, autosomal recessive; CNS, central nervous system; FAP, familial adenomatous polyposis; FH, fumarate hydratase; GIST, gastrointestinal stromal tumor; HLRCC, hereditary leiomyomatosis and renal*

cancer; LFS, Li-Fraumeni syndrome; LMS, leiomyosarcoma; MPNST, malignant peripheral nerve sheath tumor; NF1, neurofibromatosis type 1; PDGFRA, platelet-derived growth factor receptor A; PEComa, perivascular epithelioid cell tumor; PVNS, pigmented villonodular synovitis; RCC, renal cell carcinoma; RMS, rhabdomyosarcoma; STS, soft-tissue sarcoma; TSC1, tuberous sclerosis complex 1.

These predisposition disorders usually arise as a result of inherited mutations affecting the coding sequences of tumor suppressor genes or proto-oncogenes (Crucis et al., 2015). Furthermore, pedigree studies revealed that such syndromes are most frequently observed in families linked to a history of cancer along generations and tumorigenesis tends to develop at relatively early ages, during childhood or adolescence (Lupo et al., 2015). The American Society of Clinical Oncology (ASCO) has devised a set of guidelines engulfing hereditary risk assessment and associated interpretations in cancer manifestations. Thus according to the ASCO, an adequate screening of cancer family history would consist of collecting the cancer history of first and second-degree relatives. For each relative with cancer, information regarding primary cancer(s) type, the age at diagnosis of each primary cancer and the relative lineage (maternal or paternal), must be reported (Lu et al., 2014). Results obtained from cancer history taking would then guide the healthcare provider or the oncologist towards appropriate genetic testing and course of treatment. The study by Lupo et al. revealed that having a first-degree relative with cancer was specifically strongly linked with ERMS and more frequent in RMS patients in general compared to controls. Moreover, the younger was the first-degree relative at the age of diagnosis (<30 years of age), the greater was the association with childhood RMS (Lupo et al., 2015).

As far as the treatment procedures, they do not differ remarkably between familial and sporadic RMS manifestations; according to clinicopathologic prognostication schema and therapeutic regimens developed by the Intergroup

Rhabdomyosarcoma Study Group, treatment of RMS is multimodal encompassing surgery, radiation and chemotherapy (Farid and Ngeow, 2016).

B. Molecular Diagnostic Tools

Advances in DNA sequencing technologies can help detect rare and novel mutations linked to familial cancer etiology, by comparing to reference sequences reported in the literature. The availability of annotation tools such as SIFT and PolyPhen-2 scores is highly useful in order to predict the pathogenicity level of amino acid substitutions on protein function that arise as a result of an SNV. Briefly, both scales range from 0.0 to 1.0 but the boundaries have opposite meanings (Figure 1); a SIFT score of 0.0 signifies a deleterious substitution while a PolyPhen score of 0.0 predicts a benign change. On the opposite end of the scale, a SIFT score of 1.0 predicts a tolerated substitution whereas a PolyPhen score of 1.0 refers to a damaging alteration (*PolyPhen-2 Score*).

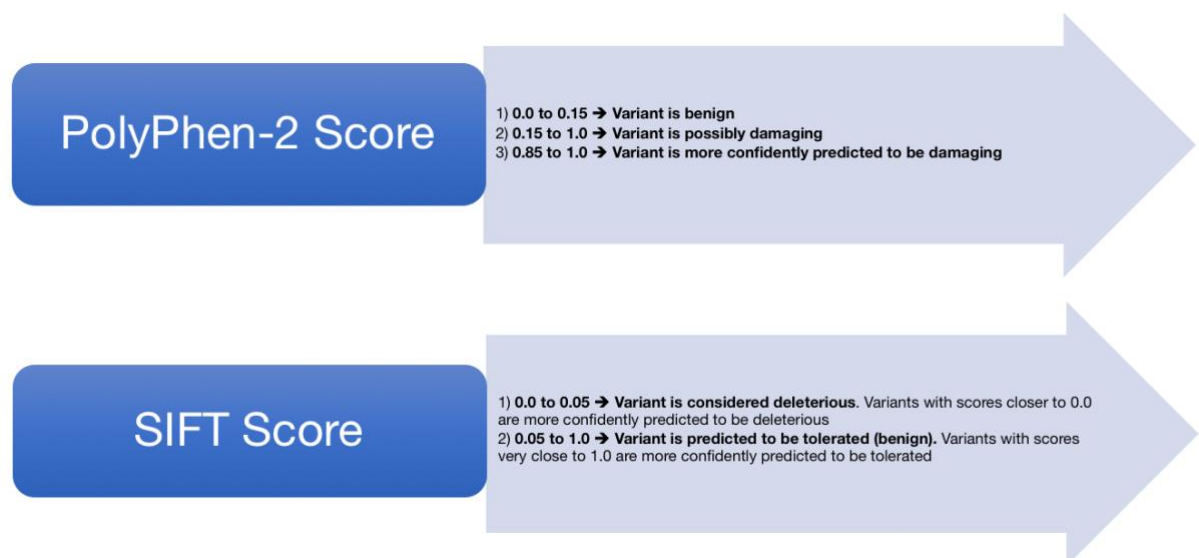


Figure 1. PolyPhen-2 and SIFT scores. Both scales are continuous and range from 0.0 to 1.0 but while a PolyPhen-2 score of 0.0 predicts a benign variant, a SIFT score of 0.0

means that the variant is deleterious. Similarly, a PolyPhen-2 score of 1.0 predicts a damaging variant while a score of 0.0 on the SIFT scale refers to a tolerated variant.

Molecular diagnostic methods for the identification of known mutations include but are not limited to Polymerase Chain Reaction (PCR), DNA microarray and Multiplex ligation-dependent probe amplification (MLPA). As for unknown mutations, geneticists rely on a variety of approaches such as Single Strand Conformational Polymorphism (SSCP), Denaturing Gradient Gel Electrophoresis (DGGE) or Restriction fragment length polymorphism (RFLP). However, Next Generation Sequencing (NGS) is the most recent and efficient sequencing procedure, that provides the highest accuracy of sequencing data (Mahdiah and Rabbani, 2013).

C. DNA Sequencing Timeline

Before dwelling into the latest NGS sequencing technology, it is interesting to shed light on the DNA sequencing methodology timeline. DNA sequencing debuted with the Sanger sequencing method in 1977 (Sanger et al., 1977). Three years later, in 1980, Allan Maxam and Walter Gilbert developed the Maxam-Gilbert chemical cleavage method (Maxam and Gilbert, 1980). Sanger sequencing is a procedure based on chain-termination using fluorescently or radioactively labeled dideoxy nucleotides lacking a 3'-OH group while the Maxam-Gilbert method relies on cleaving the DNA backbone at specific chemically modified nucleotides. These technologies are classified as First Generation Sequencing methods. Further down the line, Second Generation Sequencing was developed as a means to respond to the need of high throughput sequencing of large genomes. Second Generation methodologies involve parallel sequencing, target short DNA (or RNA) sequences and usually rely on the use of a solid support with micro channels where sequencing takes place. Next Generation sequencing

falls under the umbrella of second generation sequencing. Eventually, third generation sequencing approaches were established as cost-effective methods to massively sequence long DNA molecules in parallel, increase sequencing throughput and rates as well as simplifying the techniques used for sample preparation (Schadt et al., 2010; Slatko et al., 2018).

D. NGS Breakthrough and Workflow

NGS holds many advantages over conventional sequencing technologies; libraries are prepared in cell-free systems and thus the need to clone bacterial fragments is eliminated, the need for electrophoresis is eliminated as well since sequencing output can be detected through cyclic base interrogation. Last but not least, thousands to millions of sequencing reactions take place in parallel which translates into an immense number of reads (van Dijk et al., 2014). Major NGS technologies have been developed during the years, the first platform to see the light was released in 2005, followed by Solexa/Illumina, sequencing by Oligo Ligation Detection (SOLiD) and Ion Torrent's Personal Genome Machine respectively (Liu et al., 2012). Although NGS's platforms are diverse, they all share one common factor: massive parallel sequencing of millions of DNA fragments (Behjati and Tarpey, 2013). The first step in every NGS workflow is the library preparation, consisting of the fragmentation of DNA (or RNA) and ligation of the fragments to platform-specific adapters (Figure 2). Only fragments of the desired size fused to adapters at both ends are selected for via PCR, which also serves as an amplification procedure for sequencing. Sequencing primer sites present on the adapters allow thereafter for sequencing to take place either as single-end sequencing whereby only one end of the fragment is processed or paired-end sequencing whereby both ends

of the fragments are processed and this depends on the platform used. Furthermore, when multiple libraries are sequenced in parallel, index primers are inserted in order to distinguish between them (van Dijk et al., 2014). High depth is provided for by the multiple sequencing of all the bases of the genome (Behjati and Tarpey, 2013).

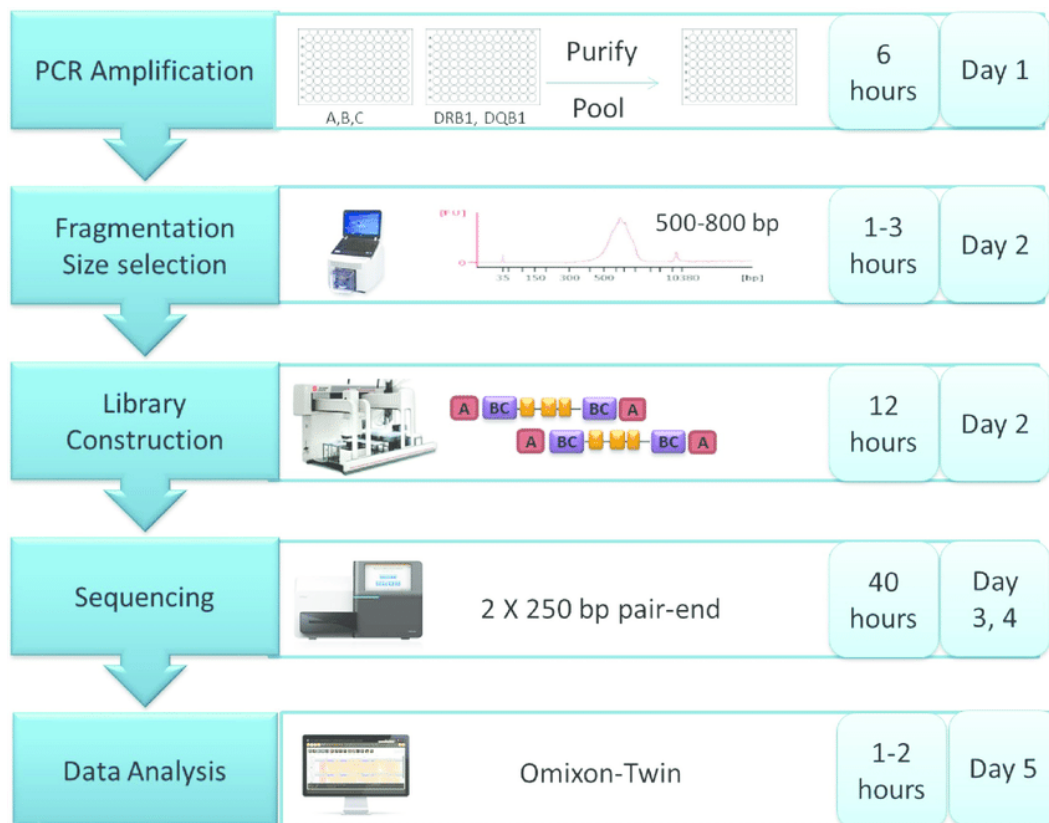


Figure 2. NGS workflow and stepwise timeline. Day 1: PCR amplification (around 6 hours). Day 2: Fragmentation and size selection (1-3 hours) and library construction (12 hours). Days 3, 4: Sequencing (40 hours). Day 5: Data analysis (1-2 hours). Each NGS experiment is performed along a period of 5 days (an average of 61.5 hours in total).

E. Depth and Breadth of Coverage

Every NGS protocol performed is characterized by two important parameters that determine the confidence level related to sequencing each base by itself and to the percentage of the genome that has been covered during the sequencing process (Figure

3). The average number of times that a specific base has been aligned to its reference counterpart in the reference genome is referred to as coverage depth (*Sequencing Coverage for NGS Experiments*). It is calculated by dividing the number of aligned reads that contain the particular base by the length of the genome (*Coverage depth – Metagenomics*). On the other hand, coverage breadth denotes the proportion of bases in a reference genome that have been covered at a specific depth. The higher the depth and breadth of coverage, the greater is the degree of confidence associated with variant discovery.

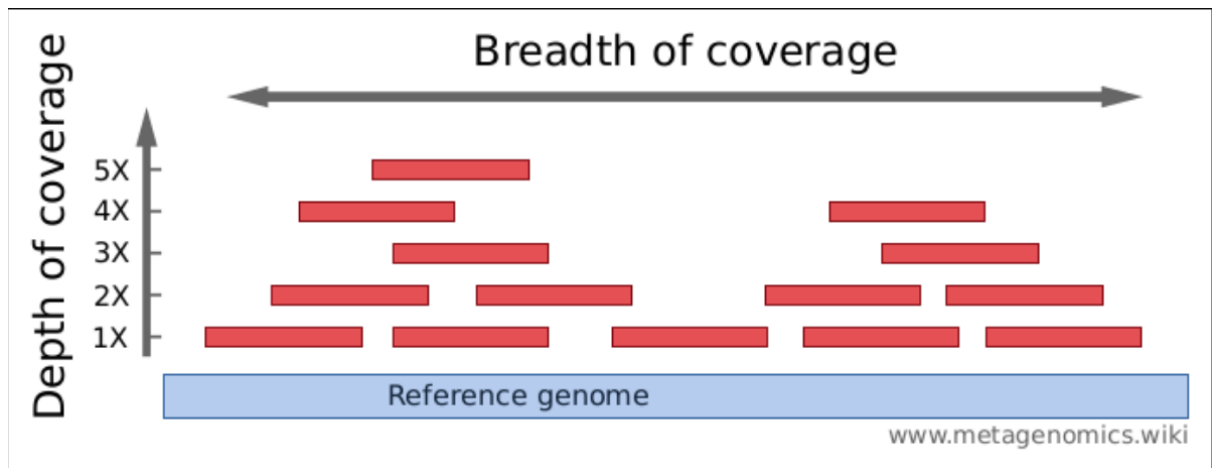


Figure 3. Depth and Breadth of Coverage in NGS. The depth of coverage is the number of times a base has been sequenced and is expressed as 1X, 2X, 3X... (as 1,2 or 3 times the coverage). The breadth of coverage is the fraction of the genome that has been sequenced at a specific depth.

F. Whole Exome Sequencing

One innovation of NGS is the possibility to sequence specific regions of the genome, especially in cases where sequencing the whole genome is unnecessary and time-consuming. Since 85% of disease-causing mutations (Mendelian diseases) occur in the protein-coding sequences of the genome or exonic regions, it is sometimes useful to

focus on sequencing these areas specifically, which comprise collectively only 1-2% of the whole genome (Choi et al., 2009). Interestingly, whole-exome sequencing (WES) offers itself as a particularly attractive and cost-efficient diagnostic tool (Mueller et al., 2018). This sequencing approach relies on the use of oligonucleotide probes that capture the protein-coding areas of DNA fragments (van Dijk et al., 2014). It allows for the identification of new variants, which can lead to the discovery of new genetic disorders or the identification of alternative forms of a previously reported disease. Every WES workflow involves three major procedures: sample preparation, target-enrichment and sequencing (Teer and Mullikin, 2010) (Figure 4). The sample is prepared via DNA extraction, purification and quality control (*Principles and Workflow*). This is followed by library preparation, whereby DNA is either physically or enzymatically fragmented and read length is selected for based on the sequencing platform used. Next, exonic regions are captured by hybridization using pre-designed oligonucleotide probes via a process known as target-enrichment, involving the use of specific exome capture kits. Subsequently, a washing and elution step is performed in order to isolate the exome, which will be sequenced using an NGS platform. The final step consists of data analysis with the help of bioinformatics; the obtained reads are aligned, variants identified, annotated and filtered and finally visualized for downstream analyses.

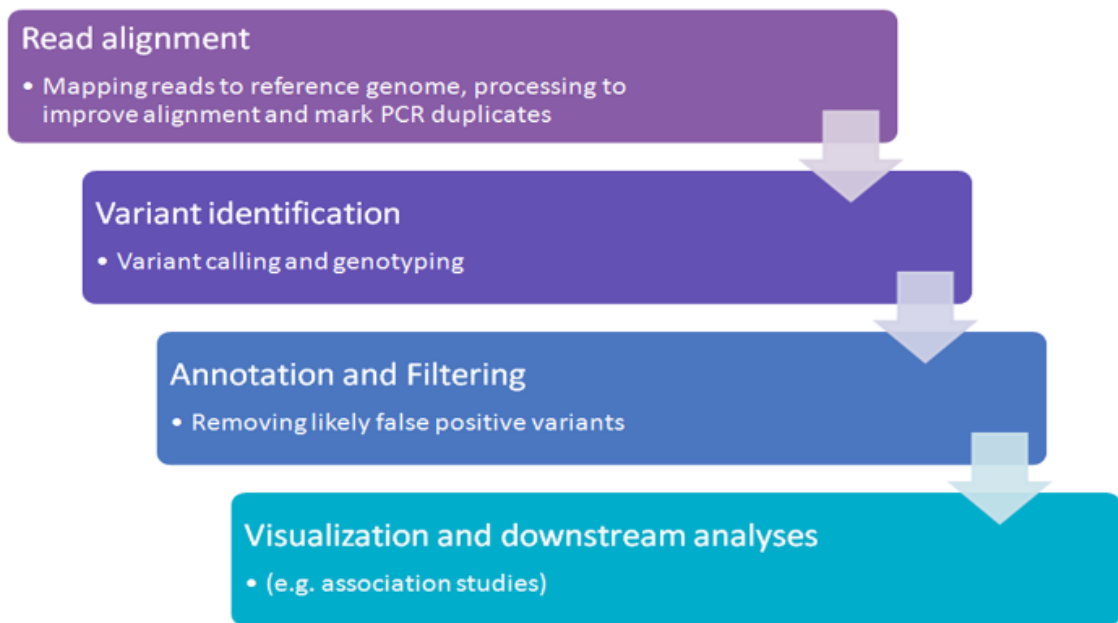


Figure 4. Variant calling pipeline. Raw reads are mapped to the reference genome, variants are identified (variant calling), annotated (researched for pathogenicity predictions and association to disease), filtered and finally visualized for analysis (involvement in disease of interest).

G. Population-Based Databases and Trio-WES Analyses

The breakthrough of NGS paved the way for the assembly and aggregation of whole genome and exome sequencing data obtained from multiple extensive sequencing projects, into large databases. Population-based studies were of particular interest as they revealed significant differences in allele frequencies among genetically diverse populations. These comparative analyses allow for the establishment of reference genomes for each population or ethnicity and subsequently for the management of Mendelian diseases in populations, based on the incidence rates of potentially pathogenic genetic variants. These reference genomes and their allele frequencies are clustered in online databases, most importantly gnomAD (encompassing ExAC) and the 1000 Genomes Project. It is worthy to note that slight differences in sequencing data rates can be observed as a result of the random sampling carried by each of the competing projects.

The unprecedented success of whole-genome and whole-exome sequencing has paved the way for the development of highly effective diagnostic strategies for the discovery of novel genetic variants in Mendelian and multigene disorders (Ewans et al., 2018). The complex field of hereditary cancer predisposition syndromes and hereditary risk assessment of pediatric cancer patients and their relatives for example, has become more accessible thanks to the increasingly popular WES parent-child trio approach (Kuhlen et al., 2018). In fact, one study performed on a cohort of children with cancer revealed that trio clinical whole-exome sequencing proved to be more effective than proband exome sequencing in the detection of causative pathogenic germline mutations and possibly de novo cancer-predisposing genes (Diets et al., 2018). Thus the leverage that family trio sequencing holds over individual sequencing in the area of pediatric oncology is its capacity to provide insight into inheritance patterns, types of mutations, key cancer pathway disruptions as well as compound heterozygosity and unidentified rare variants. This would greatly aid in the establishment of adequate individualized therapy regimens and treatment plans (Kuhlen et al., 2018).

CHAPTER II

HYPOTHESIS, AIMS, AND SIGNIFICANCE

Given the fact that RMS is a rare cancer by nature, the importance of studying familial RMS manifestations lies in the possibility of uncovering new tumor suppressor and proto-oncogenes involved in cancer biology and in the process of tumorigenesis. Thus the unexplained occurrence of familial cancers warrants investigation of the affected individual and his/her family for known and potentially novel genetic predisposition. This leads to postulate about the existence of hitherto undetected “high-risk” inherited variants that are responsible for the development of RMS in such aforementioned instances.

Two children with RMS, within the same family, that do not exhibit any of the previously reported cancer predisposition syndromes, have been identified and treated at the Children’s Cancer Institute at AUBMC. More specifically, they are first cousins (3rd degree relatives) belonging to a consanguineous family. The two pediatric patients have been diagnosed with FP-ARMS and exhibited complementary temporal profiles of diagnosis and relapse (Figure 5). The male patient was diagnosed at five years of age while his female cousin was fourteen at diagnosis.

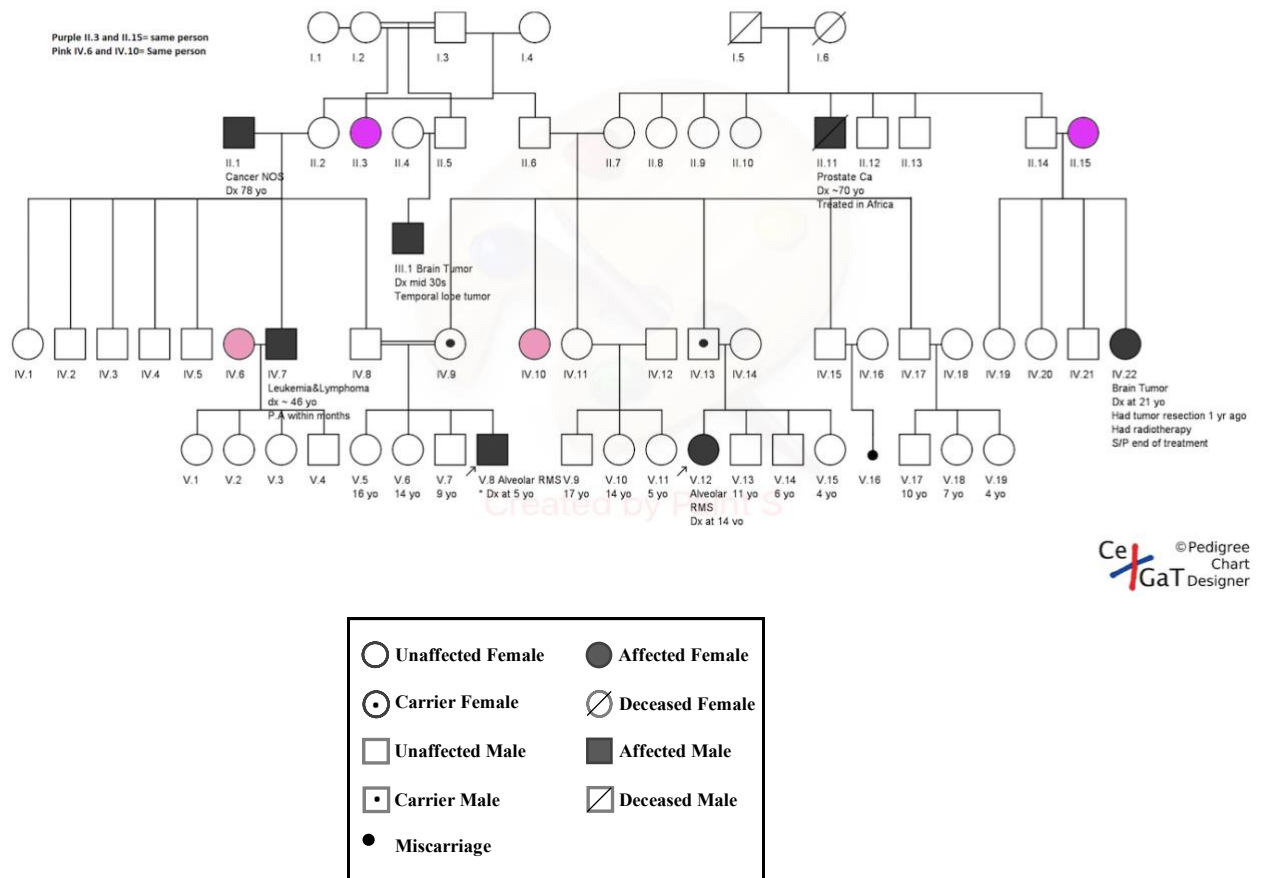


Figure 5. Family Pedigree. The two pediatric patients (V.8 and V.12) belong to a consanguineous family with an appreciable anticipation of the age of cancer onset along generations. The carrier parents (IV.9 and IV.13) are siblings. The potential carrier parent (IV.8) is a first-degree relative to the carrier parents and the non-carrier parent (IV.14) is unrelated to the members of this family.

The majority of acquired cancer-predisposing conditions develop usually as a result of single nucleotide variants in exonic regions of either DNA repair or tumor suppressor genes. Thus we propose that the two ARMS-positive relatives harbor identical potential causative mutation(s) in the coding sequence of a tumor suppressor or a DNA repair gene from their carrier parents. These potential variants are assumed to be absent in non-carrier parents.

This study focuses on identifying point mutations (more specifically missense variants) in coding regions that are common to the cousins as well as their carrier parents and absent however in the non-carrier parents. Furthermore, it aims at investigating whether the detected and filtered variants have been reported in the literature in relation to RMS or cancer in general as well as describing novel inherited mutations (if found) that contribute to the development of RMS in the before-mentioned or other similar cases.

CHAPTER III

MATERIALS AND METHODS

A. Sample Collection

Tissue samples were acquired from the first and second tumors of the two patients, and germline DNA was collected from peripheral blood of the patients as well as their parents (within a familial cancer research biorepository study). These samples are currently stored in the IRB-approved pediatric center biorepository operated by Dr. Raya Saab, main PI on this study. Both patients and their parents have consented to research on genetic testing for possible hereditary cancer (familial cancer study), under an AUBMC IRB-approved biorepository protocol. It is worthy to note that among the four parents, three are related while the female patient's mother does not belong to this family's pedigree. Thus, she was considered as a non-carrier and analysis was carried out accordingly in order to decrease the number of hits. Samples were assigned distinct numbers in order to identify them as follows: the male patient is FC-30, the female patient is FC-35, the first carrier parent (mother of the male patient) is FC-31, the second carrier parent (father of the female patient and brother of the first carrier parent) is FC-37, the first non-carrier parent (father of the male patient) is FC-47 and finally the second non-carrier parent (the unrelated mother of the female patient) is FC-42.

B. Sample Evaluation using Whole Exome Sequencing

Sequencing was performed on the germline exonic DNA of the two patients and their parents in order to identify inherited coding variants. Libraries were constructed

and captured using the SureSelectXT Human All Exon V5 technology according to the manufacturer's protocol and the generated amplified fragments were sequenced using Illumina HiSeq, following the manufacturer's instructions. After the amplification step, the sequencing libraries were quantified and diluted to working solutions and then pooled for template preparation following the manufacturer's protocol. Read length was set to 100 base pairs (bps) per read.

C. Identification of Common Variants Using Bioinformatics Analysis

Data analysis was performed on sequenced samples to map the individual reads to the human reference genome in order to assemble the obtained fragments and thus detect variants common to the cousins as well as their carrier parents (absent in non-carrier parents). Sequences were aligned against the human reference genome hg38 assembly. Base calling and sequence alignment were performed using bwa mem software and genetic variants were identified accordingly. The raw reads were obtained in the form of FastQ files that were subsequently submitted for quality control and preprocessed before the alignment step that converted the reads into BAM files. Mapped reads were further processed in order to remove duplicate ones and variant calling was performed as to identify SNVs and INDELS. Germline variants were annotated and initially filtered according to maximum allele frequency (MAF) and deleterious prediction.

D. Variants Search in Cancer Databases

An in-depth literature review was carried out to extract reported information, if any, regarding the identified variants of interest and their putative roles in RMS specifically or any other type of cancer or disease in general. This variant annotation procedure aimed at determining the estimated pathogenicity and predicted effect of the germline variants on the amino acid sequence, structure and therefore function of the associated proteins. It was performed using a number of biological or epidemiological prediction or evidence-based tools such as SIFT, PolyPhen and public databases. Databases were researched by plugging in the rsID of each variant and these mainly included Varsome, ClinVar, dbSNP, TCGA (GDC Data Portal), OncoMX, cBioPortal, COSMIC, George Washington's HIVE LAB and UniProtKB.

E. Pedigree Construction

The two patients and their respective parents were interrogated about their family history. In line with that, data was collected in regards to relationships between blood-related relatives, kinship associations and most importantly prevalence of cancer and disease among the members of this family's generations. Cancer history was traced back along five prior generations. Accordingly, the pedigree was generated by plotting the obtained information into the "kinship2" R package.

CHAPTER IV

RESULTS

A. Quality Checks of Sequenced Reads

Prior to read alignment, a quality control step was performed, a procedure crucial for assessing sequencing error and confidence level at each cycle of base calling or insertion. Quality scores are especially important when calling single nucleotide variants (SNVs) to detect legitimate alterations in the genome that are not due to experimental errors. The higher the score, the higher the confidence. In our study, the Phred score was used to compute a mean quality value across each base position of all reads for each sample (Figure 6).

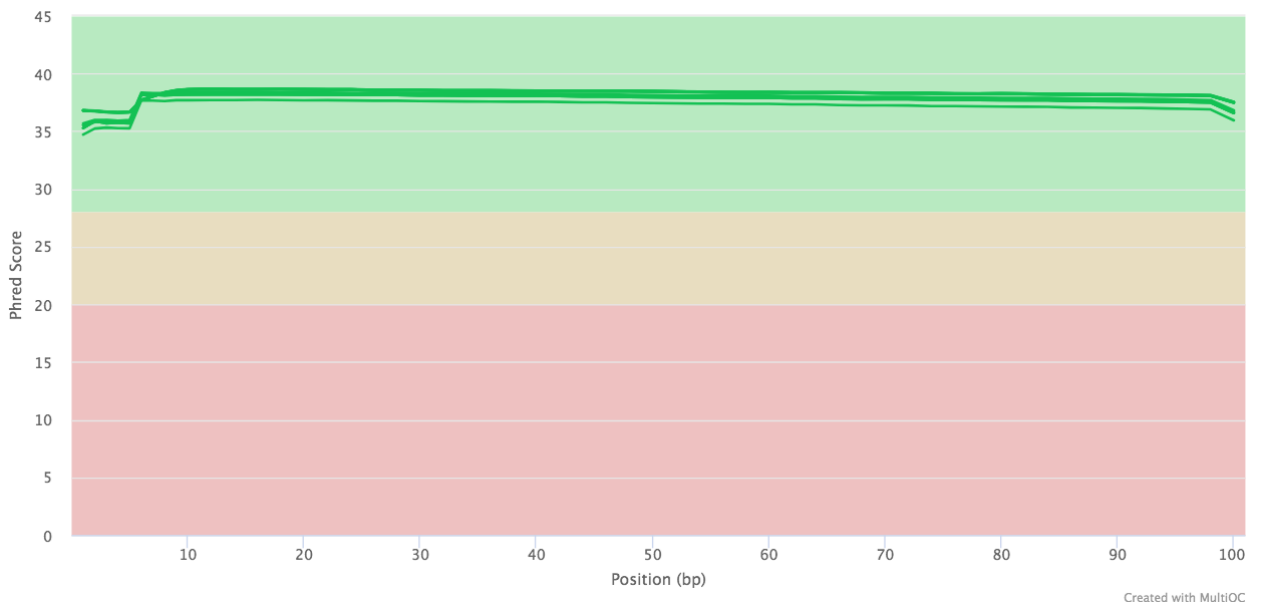


Figure 6. Sequence quality histogram: mean quality value across each base position in the read. This graph exhibits the Phred score (y-axis) for each base position (x-axis) along the 100 bp-long reads. The higher the score, the greater is the confidence in each base call or insertion.

All six samples exhibited a similar score pattern, interpreted by the overlap of the lines observed on the plot. The Phred scale ranges from 0 to 45 and the score per base is obtained algorithmically as a function of the base-calling error probabilities. A score of 30 or above translates into a base call accuracy of 99.9% or higher respectively and is considered as a good quality score (*Phred-scaled quality scores*). All samples demonstrated a mean quality score of around 35 at the first position and showed a further increase to a value of 38 starting position six, plateauing at around this score along the rest of the base positions.

B. Whole Exome Sequencing on the Two Trios

The sequencing process generated a digital output displaying the ordered sequence of nucleotides, a process referred to as base calling. This step provided a significant number of raw paired-end reads for each sample (Table 1) with read length of 100 bp. The reads for the six samples were aligned to the Human reference genome (hg38 assembly) and duplicated reads were marked using MarkDuplicates from GATK (Poplin et al. 2018). The total number of reads were highest in FC-47 followed by FC-30 and FC-42 respectively. On the other hand, FC-37 obtained the lowest output of total reads. Percentages of mapped reads were almost 100% for all samples (Table 1). As for the rates of duplication, no sample exceeded a rate of 25% but the highest degree of duplication was observed in FC-47. In parallel, the lowest level of duplication was observed in FC-37. A slightly greater rate of duplication was observed in FC-30 relatively to FC-35 (23.0972% vs. 21.0097%). The average rate of duplication however was found to be around 22% (Table 1). Read duplication is an inevitable fallacy that takes place during the PCR amplification step of sequencing. Duplication occurs when

two copies of the same original molecule are amplified and the resulting read duplicates can affect quality and confidence of mapping and variant calling. This problem is usually solved by removing or at least marking these duplicates (*How PCR duplicates arise in next-generation sequencing*, 2012). As specified by The American Health Information Management Association (AHIMA), an optimal duplication rate should not exceed 5% of mapped reads (*How to Measure Duplicate Rates*, 2013). High duplication rates (30% or higher) are the consequence of mainly two events; either there is considerable discrepancy in fragment size or the starting material for PCR is too little, which would require greater amplification of the library (*How PCR duplicates arise in next-generation sequencing*, 2012). In the presented case, the amount of DNA extracted from the samples was not abundant enough in order to obtain a minimal level of library amplification during PCR, translating into the quasi-significant duplication rate that was observed. This consequence can lead to a decrease in the efficiency of sequencing coverage (Bansal, 2017).

| Sample | Total Reads | Mapped Reads | Duplicate Reads | % mapped | % duplicate |
|---------------|--------------------|---------------------|------------------------|-----------------|--------------------|
| FC-30 | 151976773 | 151754085 | 35102380 | 99.85% | 23.0972% |
| FC-31 | 137557715 | 137278032 | 29382025 | 99.80% | 21.3598% |
| FC-35 | 136701949 | 136509205 | 28720661 | 99.86% | 21.0097% |
| FC-37 | 133065117 | 132858771 | 27810615 | 99.84% | 20.9% |
| FC-42 | 140236526 | 140045953 | 29817465 | 99.86% | 21.2623% |

| | | | | | |
|--------------|-----------|-----------|----------|--------|----------|
| | | | | | |
| FC-47 | 171836433 | 171576008 | 41597341 | 99.85% | 24.2075% |

Table 2: Read counts. The rows display the six members that form the two trios, comprised of the two pediatric patients and their relative parents. The columns present (from left to right) the total read count, the number of mapped reads, the number of duplicate reads, the percentage of mapped reads and the percentage of duplicate reads for each of the six samples.

C. Variant Calling on The Six Samples

One essential feature of NGS is variant calling, in other words the action of identifying nucleotide alterations (SNVs) once reads are aligned to the reference genome. While sequencing yields FASTQ files, read mapping produces BAM or CRAM files that can be analyzed for base alterations. In our study, sequencing data was evaluated to extract the number of single nucleotide polymorphisms (SNPs) and base insertions or deletions (INDELs) for each sample (Table 2). It is worthy to note that SNVs are referred to as SNPs when prevalent in 1% or more of the population (*Types of variants*, 2018). Table 2 further displays the variant average depth for each of the six samples. Allele depth refers to the number of reads that contain a specific variant allele across all mapped reads and the average is computed for each sample by combining all the depth values obtained for every variant.

| Sample | # of SNPs | # of INDELs | Mean Depth |
|---------------|------------------|--------------------|-------------------|
| FC-30 | 45500 | 4912 | 131.0281 |
| FC-31 | 46713 | 4890 | 118.3734 |
| FC-35 | 47173 | 4923 | 115.7991 |

| | | | |
|-------|-------|------|----------|
| FC-37 | 46665 | 4905 | 114.7013 |
| FC-42 | 47813 | 4905 | 118.9184 |
| FC-47 | 46735 | 4919 | 143.8434 |

Table 3. SNPs, INDELs, and depth of read of each sample. The rows of this table represent the six studied samples. The first column demonstrates the number of SNPs, the second provides the INDEL count and the third shows the average allele depth, for each of the six samples.

It can be observed that differences among the numbers of SNPs and INDELs are not significant between the patients and their parents. SNPs counts range between 45500 SNPs detected in FC-30 and 47813 SNPs in FC-42. As compared to FC-30, FC-35 demonstrated a higher number of SNPs (47173). Similarly, INDEL counts across the samples were analogous and FC-35 had a slightly higher count as compared to FC-30 (4923 vs. 4912). Whether these minor discrepancies are due to gender, to another variable or just to chance is yet to be determined. As for the average depth of variants, differences across samples were not substantial either and ranged from 114.7013 in FC-37 to 143.8434 in FC-47. A higher average depth was observed in FC-30 relatively to FC-35 (131.0281 vs. 115.7991).

D. Variant Filtering

Among the thousands of obtained exonic SNPs for each sample, a filtering step was conducted in order to limit the number of hits and this was crucial given the lack of a reference genome for the Lebanese population. SNV filtering was based on three distinct criteria: allele frequency, SIFT and PolyPhen scores. Thus among thousands of variants, only those that had a maximum allele frequency of 0.5, as well as projected to

be deleterious by SIFT and probably damaging by PolyPhen were selected for, variants that did not meet these requirements were filtered out of the study. All filtered variants were assigned to be missense mutations. This bottleneck process generated forty-seven variants out of the thousands obtained initially from the six samples (Table 3). All variants are missense SNVs that belong to protein coding genes.

| chr | var_position | var_id | mut | depth | MAF | gene_name | AAs | SIFT | PolyPhen |
|-------|--------------|-------------|-----|-------|-------|-----------|-----|-------------------|--------------------------|
| chr1 | 169704697 | rs2229569 | G>A | 236 | 0.329 | SELL | P/S | deleterious(0.02) | probably_damaging(0.999) |
| chr1 | 171185820 | rs2020870 | A>G | 166 | 0.151 | FMO2 | D/G | deleterious(0.01) | probably_damaging(0.94) |
| chr1 | 180178877 | rs17855475 | G>C | 120 | 0.267 | QSOX1 | G/A | deleterious(0.01) | probably_damaging(0.954) |
| chr1 | 182471338 | rs61759906 | G>C | 106 | 0.01 | RGSL1 | G/R | deleterious(0) | probably_damaging(0.996) |
| chr1 | 223628681 | rs71644745 | G>A | 122 | 0.112 | CAPN8 | A/V | deleterious(0.01) | probably_damaging(0.999) |
| chr2 | 27037601 | rs1124649 | G>A | 137 | 0.467 | TMEM214 | V/M | deleterious(0.01) | probably_damaging(0.991) |
| chr2 | 128268653 | rs3958533 | G>T | 94 | 0.328 | HS6ST1 | R/S | deleterious(0) | probably_damaging(0.995) |
| chr2 | 128318303 | rs200979099 | G>T | 119 | 0.472 | HS6ST1 | D/E | deleterious(0) | probably_damaging(0.991) |
| chr4 | 74750631 | rs11938093 | A>T | 86 | 0.329 | BTC | L/M | deleterious(0) | probably_damaging(0.982) |
| chr4 | 146867557 | rs10013280 | C>T | 46 | 0.474 | TTC29 | A/T | deleterious(0.02) | probably_damaging(0.998) |
| chr5 | 155015898 | rs17116710 | G>A | 46 | 0.264 | KIF4B | R/H | deleterious(0.02) | probably_damaging(0.985) |
| chr6 | 4087715 | rs13200786 | A>T | 125 | 0.079 | C6orf201 | D/V | deleterious(0) | probably_damaging(0.996) |
| chr6 | 20546466 | rs111739077 | G>A | 157 | 0.01 | CDKAL1 | R/Q | deleterious(0.01) | probably_damaging(0.997) |
| chr6 | 29440193 | rs2074469 | T>C | 172 | 0.283 | OR10C1 | F/L | deleterious(0.03) | probably_damaging(0.995) |
| chr6 | 33173503 | rs2855430 | G>A | 241 | 0.21 | COL11A2 | P/L | deleterious(0.02) | probably_damaging(0.995) |
| chr6 | 33286888 | rs14398 | A>G | 88 | 0.214 | WDR46 | V/A | deleterious(0) | probably_damaging(0.973) |
| chr6 | 112136181 | rs1050349 | G>C | 91 | 0.36 | LAMA4 | P/R | deleterious(0) | probably_damaging(0.988) |
| chr7 | 142065775 | rs4507684 | A>T | 107 | 0.121 | MGAM | M/L | deleterious(0) | probably_damaging(0.96) |
| chr7 | 143053050 | rs10245778 | C>T | 153 | 0.129 | OR6V1 | S/F | deleterious(0) | probably_damaging(1) |
| chr9 | 128721272 | rs11539209 | T>A | 189 | 0.109 | ZDHHC12 | N/I | deleterious(0) | probably_damaging(0.997) |
| chr9 | 132499377 | rs7047726 | G>A | 204 | 0.485 | CFAP77 | G/R | deleterious(0.03) | probably_damaging(1) |
| chr11 | 1017898 | rs781135233 | T>G | 742 | 1E-04 | MUC6 | T/P | deleterious(0.02) | probably_damaging(0.991) |
| chr11 | 4849039 | rs12417164 | A>T | 151 | 0.339 | OR51S1 | I/N | deleterious(0) | probably_damaging(0.985) |
| chr11 | 59422439 | rs1453547 | G>A | 136 | 0.286 | OR5A2 | P/L | deleterious(0.03) | probably_damaging(0.993) |
| chr11 | 67634891 | rs3758938 | T>G | 252 | 0.453 | TBX10 | K/T | deleterious(0.01) | probably_damaging(0.991) |
| chr11 | 119174663 | rs199828804 | C>A | 190 | 0.003 | NLRX1 | L/I | deleterious(0.01) | probably_damaging(0.996) |
| chr11 | 123906595 | rs17127947 | T>G | 122 | 0.284 | OR8D4 | L/R | deleterious(0) | probably_damaging(0.992) |
| chr12 | 10010776 | rs1359082 | C>A | 72 | 0.462 | CLEC12B | T/N | deleterious(0.02) | probably_damaging(0.998) |
| chr12 | 48330001 | rs2291483 | C>T | 110 | 0.291 | H1FNT | S/F | deleterious(0.01) | probably_damaging(0.925) |
| chr12 | 51890884 | rs12368048 | C>A | 260 | 0.261 | ANKRD33 | T/N | deleterious(0) | probably_damaging(0.952) |
| chr12 | 52433824 | rs2232387 | C>T | 231 | 0.209 | KRT75 | A/T | deleterious(0) | probably_damaging(0.999) |
| chr12 | 57272472 | rs143617893 | G>A | 93 | 0.004 | R3HDM2 | P/L | deleterious(0.03) | probably_damaging(0.995) |
| chr12 | 71139754 | rs3763978 | C>G | 108 | 0.415 | TSPAN8 | G/A | deleterious(0.03) | probably_damaging(0.989) |
| chr15 | 53615751 | rs17730281 | G>A | 132 | 0.485 | WDR72 | L/F | deleterious(0) | probably_damaging(0.96) |
| chr15 | 89260719 | rs62020347 | C>T | 61 | 0.108 | FANCI | P/L | deleterious(0.01) | probably_damaging(0.992) |
| chr16 | 28592334 | rs1059491 | T>G | 145 | 0.445 | SULT1A2 | N/T | deleterious(0) | probably_damaging(0.999) |
| chr16 | 71649815 | rs61733127 | A>G | 157 | 0.175 | PHLPP2 | L/S | deleterious(0) | probably_damaging(0.935) |
| chr16 | 78432540 | rs3764340 | C>G | 215 | 0.108 | WWOX | P/A | deleterious(0.04) | probably_damaging(0.92) |
| chr17 | 35984351 | rs16971802 | T>C | 83 | 0.154 | CCL14 | K/E | deleterious(0.03) | probably_damaging(0.953) |
| chr17 | 43094464 | rs1799950 | T>C | 273 | 0.081 | BRCA1 | Q/R | deleterious(0.02) | probably_damaging(0.944) |
| chr19 | 19302283 | rs17751061 | C>T | 236 | 0.201 | SUGP1 | R/H | deleterious(0.01) | probably_damaging(0.999) |
| chr19 | 40668105 | rs536916726 | G>T | 79 | 0.001 | NUMBL | P/H | deleterious(0.05) | probably_damaging(0.997) |
| chr19 | 40876571 | rs75152309 | T>A | 144 | 0.179 | CYP2A7 | K/M | deleterious(0.01) | probably_damaging(0.954) |
| chr19 | 44908822 | rs7412 | C>T | 58 | 0.11 | APOE | R/C | deleterious(0) | probably_damaging(1) |
| chr19 | 47375601 | rs61751860 | C>T | 86 | 0.155 | DHX34 | R/C | deleterious(0) | probably_damaging(0.973) |
| chr20 | 31820503 | rs34396614 | C>G | 216 | 0.028 | MYLK2 | P/A | deleterious(0) | probably_damaging(0.994) |
| chr22 | 30366151 | rs740223 | G>A | 263 | 0.318 | CCDC157 | D/N | deleterious(0) | probably_damaging(1) |

Table 4: The 47 variants common to the two patients. This excel sheet provides basic information regarding the 47 variants, obtained after the filtering step. The columns, from left to right, are as follows: chromosome number, variant position on the chromosome, variant ID, reference and alternate alleles, allele depth, maximum allele frequency, gene name, reference and alternate amino acids, SIFT score and PolyPhen-2 score.

As detailed below, the process of variant filtering and elimination was subsequently carried further by integrating new variables related to gene function, presence/absence of variants in carrier parents, degree of pathogenicity, allele frequency as well as whether a variant has been reported in cancer databases.

E. FANCL and BRCA1 gene variants as highly likely candidates based on gene function

Forty-seven potential causative variants were established in forty-seven different genes. We decided to focus on the function of the different proteins encoded by these genes to enhance the process of putative gene selection. More specifically, DNA repair and tumor suppressor genes were of particular interest given that these families of genes play major roles in the cell cycle. Subsequently, the 47 filtered variants were classified according to their gene function; tumor suppressors and DNA repairs were categorized separately given that they were of particular interest whereas genes encoding all other cellular functions were grouped into one large category (Figure 7). Furthermore, we hypothesized that the putative inherited germline variant(s) involved in the development of RMS in the two cousins are related to the coding sequence of tumor suppressor genes. Two interesting variants belonged to the BRCA1 (ID#: rs179950) and the FANCL (ID#: rs62020347) genes, which play the dual role of DNA repair and tumor suppression.

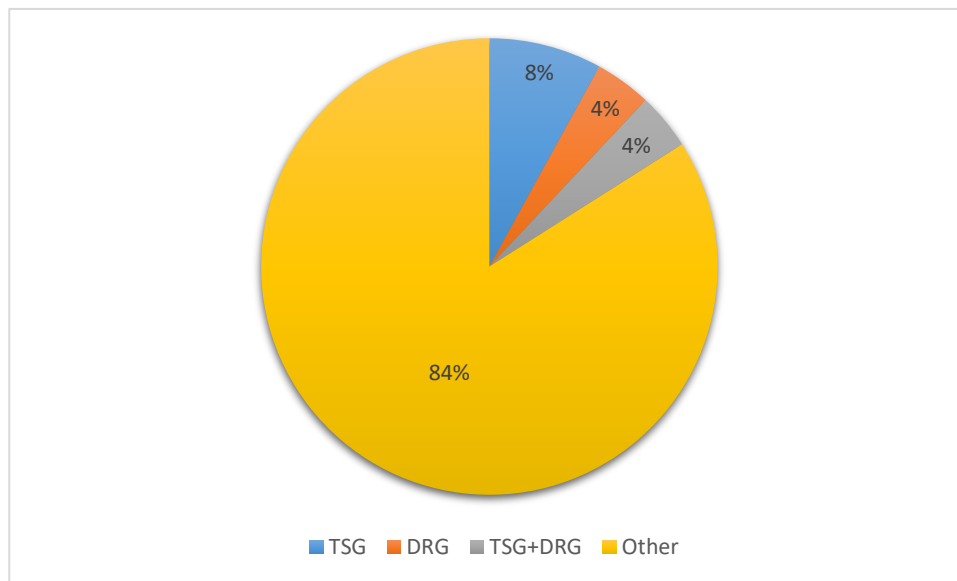


Figure 7. Proportions of TSG variants, DRG variants and others. This pie chart divides the 47 filtered variants into those belonging to either tumor suppressor genes (TSGs), DNA repair genes (DRGs), both TSGs and DRGs or other protein-coding genes.

F. Variants common to patients and carrier parents as eligible causative candidates

The carrier parents that have transferred the potential causative mutation(s) to the two RMS-positive cousins (FC-30 and FC-35) were assumed to be the two sibling parents (FC-31 and FC-37). Accordingly, the second pair of parents (FC-42 and FC-47) were presumed to be non-carriers.

Thus among the forty-seven potential variants, those that were present simultaneously in the two patients as well as their carrier parents were of particular interest. Interestingly, only ten variants were not detected in the carrier parents (Figure 8). Furthermore, sequencing data revealed that none of the forty-seven variants were carried by FC-42 (the unrelated mother) and that among those present in the carrier parents, fifteen were also detected in FC-47 (parent related to sibling parents given that

the family is consanguineous). The fact that these variants are present in three out of the four parents increases their potential of being involved in the cause of tumor predisposition in the two patients. Additionally, two variants out of the fifteen were particularly attractive, the first was affiliated with FANCL (rs62020347), a gene that plays a role in DNA repair and the other (rs61751860) affected the DHX34 gene, an established tumor suppressor.

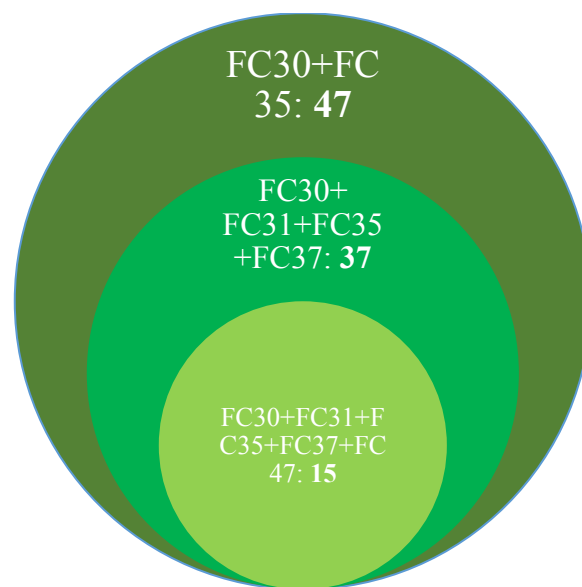


Figure 8. Variants common to patients and parents. This diagram exhibits the number of variants common to the two patients (FC30 and FC35), those common to the patients as well as their carrier parents (FC30, FC31, FC35 and FC37) and the number of variants shared by the patients, the carrier parents and the potential carrier parent (FC30, FC31, FC35, FC37 and FC47).

We observed that carrier parent FC-31 shares less variants with the patients as compared to carrier parent FC-37 (Table 4). Furthermore, the potential carrier FC-47 shares the least amount of common variants with the two cousins.

| Parent (FC ID) | Shared Variants (rs IDs) |
|-------------------------|---|
| Carrier FC-31 | rs2229569, rs2020870, rs17855475, rs61759906, rs71644745, rs1124649, rs3958533, rs10013280, rs13200786, rs111739077, rs2855430, rs14398, rs1050349, rs4507684, rs10245778, rs11539209, rs7047726, rs12417164, rs1453547, rs3758938, rs199828804, rs17127947, rs1359082, rs2291483, rs12368048, rs2232387, rs143617893, rs3763978, rs62020347, rs61733127, rs3764340, rs16971802, rs1799950, rs17751061, rs536916726, rs75152309, rs7412, rs61751860 (40 in total) |
| Carrier FC-37 | rs2229569, rs2020870, rs17855475, rs61759906, rs71644745, rs1124649, rs200979099, rs11938093, rs10013280, rs17116710, rs13200786, rs111739077, rs2074469, rs2855430, rs14398, rs1050349, rs4507684, rs10245778, rs11539209, rs7047726, rs12417164, rs1453547, rs3758938, rs199828804, rs17127947, rs1359082, rs2291483, rs12368048, rs2232387, rs143617893, rs3763978, rs17730281, rs62020347, rs1059491, rs61733127, rs3764340, rs16971802, rs1799950, rs17751061, rs536916726, rs75152309, rs7412, rs61751860, rs34396614, rs740223 (45 in total) |
| Potential Carrier FC-47 | rs2020870, rs17855475, rs71644745, rs1124649, rs3958533, rs200979099, rs11938093, rs10013280, rs17116710, rs2074469, rs1050349, rs7047726, rs12417164, rs1453547, rs17127947, rs2291483, rs3763978, rs17730281, rs62020347, rs1059491, rs17751061, rs61751860, rs34396614, rs740223 (24 in total) |

Table 5. Variants shared with parents. This table displays the variants shared between the two patients and carrier parent FC-31, carrier parent FC-37 and potential carrier parent FC-47, respectively. It can be observed that the number of variants common to the patients and their carrier parents (40 and 45) is close to the total number of variants (47). The highlighted variants refer to those shared between the two carrier parents (FC-31 and FC-37). Carrier FC-37 has five additional variants common to the patients as compared to carrier FC-31.

G. Variants with unknown clinical significance are more probable to be associated with tumor predisposition as compared to benign/likely benign variants

Next, we attempted to determine the clinical significance of the variants and thus their relationship to human health, or in other words to predict the degree of pathogenicity of the variants of interest (benign, pathogenic or of unknown significance). This was possible due to a number of databases that report pathogenicity verdicts based on a combination of supporting evidence extracted from a number of variables such as in silico algorithms, epidemiological and biological data. The strength of predictions varies from weak evidence associated with in silico algorithms to strong evidence offered by the biological analyses. Thus for this purpose, two important databases were explored; Varsome and ClinVar. Verdicts reported in these databases regarding the variants of this study mainly overlapped, with a benign/likely benign majority (Figure 9). 15% of variants were described as having an unknown significance (VUS) while none of the SNVs were reported to have a pathogenic effect. Notably, the pathogenicity outcome diverged between the two databanks in regards to two mutations, the KRT75 gene variant (ID#: rs2232387) reported as benign in Varsome but as a risk factor in ClinVar and the APOE gene variant (ID#: rs7412), described as benign in Varsome but as pathogenic in ClinVar. The KRT75 gene's function is essential for hair and nail formation while the APOE gene produces an apoprotein involved in the catabolism of triglyceride-rich lipoproteins. None of the implicated genes plays a role in DNA repair or in tumor suppression and the differences observed in the reported pathogenicity can be explained from discrepancies in the prediction algorithms used, as well as the populations and cases that have been examined by each database.

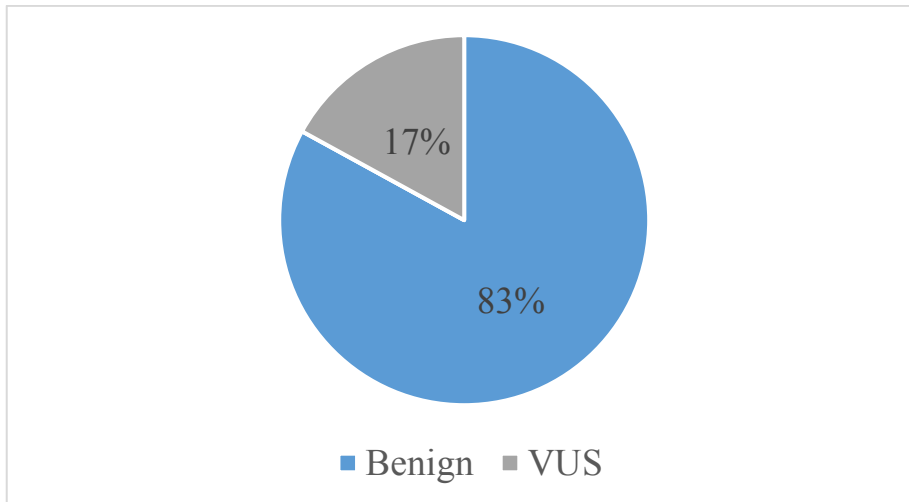


Figure 9. Pathogenicity of the 47 variants (according to Varsome). This pie chart represents the fractions of benign variants, pathogenic variants and variants of unknown significance (VUS). None of the variants were described as pathogenic.

H. The identification of low allele frequency variants highlights four major gene candidates: RGSL1, CDKAL1, NLRX1 and R3HDM2

It was reasonable to assume that the potential causative variant(s) responsible for the predisposition to RMS tumors in the aforementioned case study are very rare, or in other words characterized by very low allele frequencies among populations. It was possible to investigate the allele frequency of each of the forty-seven variants of interest using online databases such as GnomAD or 1000 Genomes (Table 5). Incidence rates observed in different databases were highly consistent but minor inconsistencies in percentages detected between databases can be explained as a result of the random selection of individuals among populations. Variants with allele frequencies inferior to 1% were more appealing since the lower the frequency of a mutation, the greater its probability of being involved in a pathogenic process. Interestingly, four variants were identified to have an average allele frequency lower than 1%; the RGSL1 gene variant

(rs1759906), the CDKAL1 gene variant (rs111739077), the NLRX1 gene variant (rs199828804) and the R3HDM2 gene variant (rs143617893). All four variants were predicted to be of unknown significance and none of the associated genes was a DNA repair or a tumor suppressor gene. The NLRX1 variant had the lowest frequency among the four mutations.

| Variant ID | Gene Name | AF gnomAD | AF ExAC | AF 1000 |
|-------------------|------------------|------------------|----------------|----------------|
| rs2229569 | SELL | 0.1838 | N/A | 0.244 |
| rs2020870 | FMO2 | 0.0678 | 0.08149 | 0.092 |
| rs17855475 | QSOX1 | 0.1331 | 0.13604 | 0.099 |
| rs61759906 | RGSL1 | 0.0038 | 0.0016 | 0.001 |
| rs71644745 | CAPN8 | 0.0629 | 0.0666 | 0.081 |
| rs1124649 | TMEM214 | 0.3422 | 0.31613 | 0.303 |
| rs3958533 | HS6ST1 | N/A | 0.3522 | N/A |
| rs200979099 | HS6ST1 | N/A | 0.4625 | N/A |
| rs11938093 | BTC | 0.2338 | 0.2237 | 0.2 |
| rs10013280 | TTC29 | 0.3561 | 0.4552 | 0.408 |
| rs17116710 | KIF4B | 0.1706 | 0.18761 | 0.179 |
| rs13200786 | C6orf201 | 0.0711 | 0.0581 | 0.047 |
| rs111739077 | CDKAL1 | 0.0059 | 0.00537 | 0.003 |
| rs2074469 | OR10C1 | 0.1558 | 0.18997 | 0.17 |
| rs2855430 | COL11A2 | 0.1072 | 0.12304 | 0.102 |
| rs14398 | WDR46 | 0.0909 | 0.12521 | 0.098 |
| rs1050349 | LAMA4 | 0.215 | N/A | 0.21046 |
| rs4507684 | MGAM | 0.0524 | 0.03956 | 0.048 |
| rs10245778 | OR6V1 | 0.0891 | 0.0803 | 0.082 |
| rs11539209 | ZDHHC12 | 0.0524 | 0.1147 | 0.056 |

| | | | | |
|-------------|---------|--------|---------|-------|
| rs7047726 | CFAP77 | 0.1041 | 0.1268 | 0.186 |
| rs781135233 | MUC6 | N/A | 0.015 | N/A |
| rs12417164 | OR51S1 | 0.131 | 0.17949 | 0.143 |
| rs1453547 | OR5A2 | 0.1992 | 0.20902 | 0.126 |
| rs3758938 | TBX10 | 0.2423 | 0.25685 | 0.217 |
| rs199828804 | NLRX1 | 0.0003 | 0.00084 | N/A |
| rs17127947 | OR8D4 | 0.1936 | 0.17719 | 0.187 |
| rs1359082 | CLEC12B | 0.3923 | 0.35266 | 0.312 |
| rs2291483 | H1FNT | 0.2329 | 0.2198 | 0.206 |
| rs12368048 | ANKRD33 | N/A | N/A | 0.165 |
| rs2232387 | KRT75 | 0.1253 | 0.12308 | 0.143 |
| rs143617893 | R3HDM2 | 0.0006 | 0.0015 | 0.001 |
| rs3763978 | TSPAN8 | 0.2931 | 0.33319 | 0.227 |
| rs17730281 | WDR72 | 0.2307 | 0.2549 | 0.267 |
| rs62020347 | FANCI | 0.0441 | 0.05456 | 0.046 |
| rs1059491 | SULT1A2 | 0.3308 | 0.3126 | 0.225 |
| rs61733127 | PHLPP2 | 0.1124 | N/A | 0.087 |
| rs3764340 | WVOX | 0.0733 | 0.07435 | 0.092 |
| rs16971802 | CCL14 | 0.0671 | 0.04665 | 0.063 |
| rs1799950 | BRCA1 | 0.052 | 0.04407 | 0.022 |
| rs17751061 | SUGP1 | N/A | 0.11968 | 0.069 |
| rs536916726 | NUMBL | N/A | N/A | N/A |
| rs75152309 | CYP2A7 | 0.0854 | 0.07308 | 0.061 |
| rs7412 | APOE | 0.0827 | 0.0718 | 0.075 |
| rs61751860 | DHX34 | 0.1185 | 0.1679 | 0.087 |
| rs34396614 | MYLK2 | 0.0127 | N/A | 0.008 |
| rs740223 | CCDC157 | 0.1881 | 0.19611 | 0.159 |

Table 6. Variants' allele frequencies. This table provides the reported allele frequencies (AF) in the GnomAD, ExAC and 1000 Genomes databases. The columns, from left to right, represent variant ID, gene name, GnomAD reported frequency, ExAC reported frequency and 1000 Genomes reported frequency for each of the 47 variants.

I. Further research is required to study the tumorigenic effect of the filtered variants that have not been previously associated with cancer

To better understand potential contribution to phenotype, we searched genomic cancer databases (including COSMIC, TCGA and ICGC) for the SNVs of interest to identify whether they have been reported to be linked to RMS or to other types of cancer. The forty-seven variants could be classified into six major categories based on data retrieved from the public databases. These databases use specific algorithms in order to harvest and sort variants from different online-published cancer studies and thus associate each variant to the cancer type in which it was detected. Each of the forty-seven variants was searched using its rsID and most were indeed detected in cancer cases, more specifically five cancer types and thus were grouped accordingly (Figure 10). The five cancer types in which variants were reported were as follows: colorectal cancer, esophageal cancer, breast cancer, thyroid carcinoma, and skin cancer. The rest of the variants have not been reported in any cancer types; these were grouped into category number six (not found or NF). The majority of variants detected in cancer cases were associated with colorectal cancer. While the SNVs not previously reported in other cancers might be exclusively linked to RMS tumors, further research needs to be performed to identify whether they are indeed detected in other cases of RMS. On the other hand, a few variants have been described in esophageal or breast cancers as well as thyroid carcinoma.

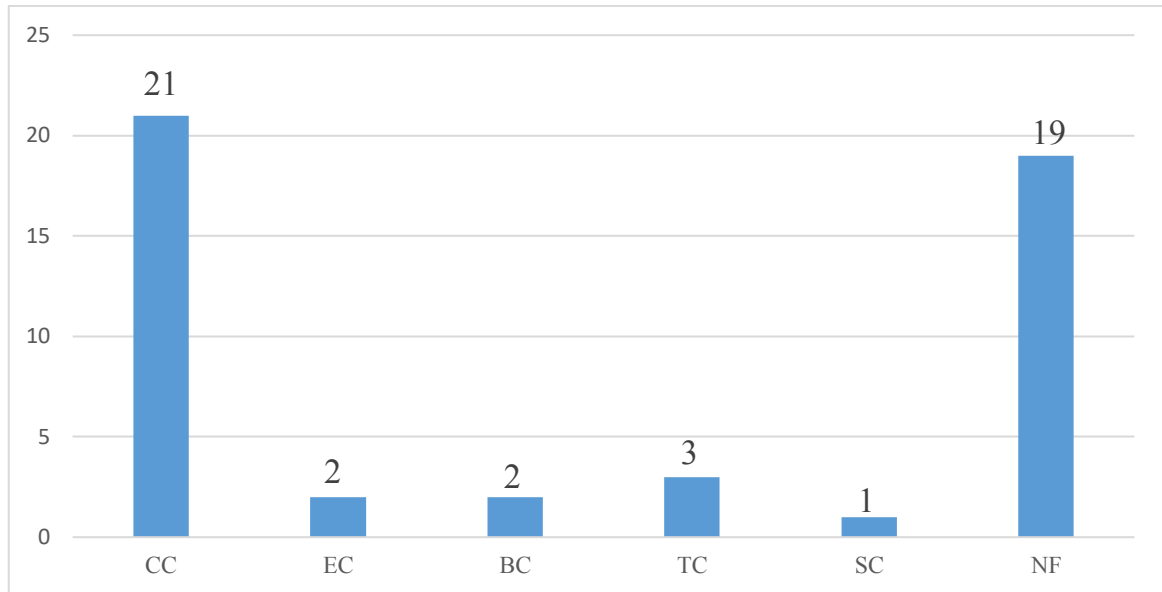


Figure 10. Association of the 47 variants with different types of cancer. This bar plot groups the 47 variants into six categories: those that have been reported in cases of colorectal cancer (CC), esophageal cancer (EC), breast cancer (BC), thyroid carcinoma (TC), skin cancer (SC), and those that have not been reported in cancer (NF). The six categories are presented on the x-axis and the scale for variant counts is displayed on the y-axis.

Three variants were relatively more attractive candidates as they combined a number of interesting criteria such as being rare (with an allele frequency very close to 1%), being present in the carrier parents, and affecting the exonic region of either DNA repair or tumor suppressor genes. These variants included the variant of the FANCL gene (ID#: rs62020347), that of BRCA1 (ID#: rs1799950) and lastly the NUMBL gene variant (ID#: rs536916726).

CHAPTER V

DISCUSSION AND LIMITATIONS

Sarcomas encompass a class of rare and heterogeneous tumors. In fact, only 7% of pediatric tumors are soft tissue sarcomas, half of which are classified as RMS (Spunt and Pappo, 2006). RMS is mostly sporadic; however, around 5% of cases have been associated with familial cancer predisposition syndromes (Hemminki and Li, 2001). Children with established hereditary conditions such as Li-Fraumeni syndrome characterized by a germline TP53 mutation, neurofibromatosis type I or retinoblastoma with inherited RB1 mutations witness an increased prevalence of RMS (Malkin et al., 1990; Sung et al., 2004; Farid and Ngeow, 2016). The previously mentioned study by Lupo et al. further revealed that a family history of cancer is highly correlated with cancer predisposition syndromes, which has also been linked to early onset childhood RMS. In fact, it showed that having a first-degree relative with cancer was more prevalent in the case of RMS patients and that risk of RMS was strongly positively correlated with having a first-degree relative with cancer who was diagnosed at a relatively younger age (Lupo et al., 2015). Thus although little research has been carried out in regards to familial RMS occurrences, inherited genetic susceptibility and familial history of cancer are important factors that should be taken into consideration since it may allow for improved diagnosis, prognosis and treatment in the clinical setting (Chan et al., 2017).

A number of familial syndromes contributing to the development of childhood RMS have been already reported. These conditions usually arise as a result of inherited germline mutations that often manifest phenotypically into growth anomalies. Few

hereditary mutations however, have been described to increase the risk of RMS without translating into physical aberrations and the fact that many others are yet to be discovered is not unlikely. The case of the two first-cousins in a family with cancer history, that developed the same subtype of RMS at a very early age fits this assumption. It led us to postulate that they have inherited causative mutations from their carrier parents that were involved in the sarcoma tumor growth. The approach followed was to perform whole-exome sequencing on the two trios composed of the patients and their parents and to filter the variants in concordance with specific criteria. The more damaging, health-compromising and rare a variant is, the more it qualifies to be a causative agent in the development of RMS in the two patients. Moreover, SNPs occurring in the coding regions of tumor suppressor genes had greater chances of playing a role in the patients' malignancies.

Germline mutations of FANCL, a gene essential for DNA repair and chromosomal stability, constitute a main cause behind congenital Fanconi anemia (FA) (Bogliolo and Surrallés, 2015). FA is a rare condition that manifests with bone marrow failure, an early aging onset, developmental abnormalities as well as in increased susceptibility to a number of different cancers. In fact, FA patients are at a high risk of developing head and neck, gastrointestinal, vulvar and esophageal cancers (Bagby, 2003). Furthermore, a study performed on a large Asian cohort showed that pediatric sarcoma patients harbored pathogenic FANCL germline mutations, establishing an association between inherited FANCL mutations and early sarcoma onset (Chan et al., 2017).

BRCA1 is an established tumor suppressor and DNA double-strand break repair gene (O'Donovan and Livingston, 2010). It is a breast and ovarian cancer-associated

gene whereby BRCA1 inherited germline mutations have been detected in 20%-40% of familial breast cancer cases (Couch et al., 2014). Although the association between BRCA1 inherited variants and childhood RMS has not been reported in the literature, the BRCA1 SNV qualifies as a potential causative agent in this family.

The third identified gene of potential interest is NUMBL. NUMBL is a recently recognized tumor suppressor gene, as it has been observed to be downregulated in human lung cancer cell lines, inducing Notch pathway activation and subsequently tumorigenesis (Yingjie et al., 2013). An increasing number of NUMBL SNVs are being reported to be in association with multiple carcinomas, however no previous links have been made between NUMBL germline mutations and sarcoma tumor growth. Variants of other genes could as well be involved but this requires further research.

This study has a number of limitations, starting with the lack of adequate control. Germline DNA was extracted from the peripheral blood of the patients, given that inherited SNVs play a major role in cancer etiology. Due to the incomplete penetrance of the tumor phenotype, it was not adequate to use family members as good controls and thus the use of population controls would have been an ideal scenario. Another limitation lies in the absence of a reliable reference genome for the Lebanese population; the presence of a reference genome would have helped filtering out insignificant variants that may have low average allele frequencies but that may be relatively prevalent among the Lebanese population. Instead, frequencies deriving from European, Latino and African populations were taken into consideration in order to compute incidence rates of the obtained germline variants of interest.

Besides the aforementioned epidemiological limitations, our research project further presented with a methodological limitation regarding the average rate of

duplication. A considerable duplication rate is inevitable and even expected for WES and decreases when paired-end sequencing is performed as per our study. The average rate was reported at 22%, a value that was possibly a consequence of the small amount of extracted DNA. This could have had an unfavorable impact on the efficiency of sequencing coverage of the experiment, and would call for enhancing the complexity of the library (Bansal, 2017).

Another obstacle of this study involved the statistical analysis of obtained results, as some hits could have been missed. One explanation is attributed to errors in exon-intron recognition especially in regards to pairs of genes with short intergenic regions or genes with long introns. This is correlated with the identification of gene splice sites and highlights the role of alternative splicing in variant identification. Another reason could be errors of sequencing in the analyzed reads (Koonin and Galperin, 2011). Furthermore, large studies are required in order to convey adequate statistical power to disease-associated low-frequency polymorphisms (Stitzel et al., 2011).

While our study focused on mutations affecting the coding sequences that comprise only 2% of the human genome (Elkon and Agami, 2017), it failed to encompass those that affect the non-coding DNA. In fact, it is possible that polymorphisms potentially associated with RMS arise in non-exonic sequences such as introns, regulatory sequences or tandem repeats. Another set of mutations that was not tackled in our project were the epigenetic aberrations whose role in cancer etiology is growing (Nebbioso et al., 2018). More comprehensive approaches, although costlier, entail the use of whole genome or chromatin immunoprecipitation (ChIP) sequencing

CHAPTER VI

CONCLUSION AND FUTURE PROSPECTS

Previous studies on sarcoma patients have revealed that the majority of cases do not present with a remarkable family history of cancer. Nevertheless, sarcomas are strongly associated with cancer predisposition syndromes, the latter being manifestations of heritable, germline genetic mutations (Abha Gupta, 2021). It is only recently that oncologists grasped the role and significance of family history and genetic predisposition in cancer etiology. Thus the frequency of children with soft-tissue sarcomas “having a genetic predisposition to malignancy” is most probably underrated (DeVita, 2008). Oncologists are delving into the search for new genes and tumor clusters in order to establish “genotype:phenotype correlations in sarcoma patients” (Abha Gupta, 2021).

Our study tackled RMS, a rare soft-tissue sarcoma characterized by an early-age onset. We postulated that although a number of cancer predisposition syndromes caused by well-identified germline mutations have been already associated with RMS, other causative variants are still to be discovered. This conjecture arose when two first cousins, a 5-year-old male and a 14-year-old female were diagnosed with the alveolar subtype of RMS with no prior diagnosis of a cancer predisposition syndrome. Our results revealed 47 genetic missense variants common to the two relatives, a number of which have been reported in relation to different classes of carcinomas but not sarcomas. Three SNVs were of particular interest (the FANCL, BRCA1 and NUMBL gene variants), as they met a number of common criteria including their role in tumor

suppression, their relatively low allele frequencies and their detection in the carrier parents.

This study warrants one step forward towards understanding the biology of RMS, especially in regards to the setting of familial disease. An essential action moving forward would be to carry out in vitro functional analysis studies of the identified variants. This procedure allows for adequate investigation of the effect of each missense variant on the functional properties of the protein it encodes. Over twenty functional assays have been developed during the past few years; these include the homologous recombination assay in human cells and the yeast recombination assay (Guidugli et al., 2013). Another technique, the embryonic stem cell-based functional assay relies on introducing a human wild-type allele in otherwise lethal mouse embryonic stem cells in order to rescue the phenotype. Variants that do not succeed to rescue lethality are presumed pathogenic (Kuznetsov et al., 2008). The syngeneic human cancer variant knockout cell line model is also one good procedure for functional analysis. It enables the characterization of missense variants through their insertion into a suitable human cancer cell line (Hucl et al., 2008). The protein-protein interaction-based assays are a common valuable approach to assess variant function; the latter can be inferred from the variant protein's interaction with other proteins. Some biochemical methods include pull downs, co-immunoprecipitation and yeast two-hybrid strategies (Guidugli et al., 2013).

Further research needs to be carried whereby the process of variant filtering can be improved by including more specific criteria that would help in eliminating insignificant competing SNPs. This would at least help in narrowing down the number of potential hits. Another useful approach would be to perform a wide Lebanese cohort

study of familial childhood RMS cases that descend from families with a history of cancer, in order to sequence germline exonic SNVs and compare them among patients. Common variants would have greater chances in playing a role in the development of RMS in these pediatric patients.

REFERENCES

- Abha Gupta, M. (2021). Sarcomas and Cancer Predisposition Syndromes.
- Bagby, G. (2003). Genetic basis of Fanconi anemia. *Current Opinion In Hematology* 10, 68-76.
- Bansal, V.** (2017). A computational method for estimating the PCR duplication rate in DNA and RNA-seq experiments. *BMC Bioinformatics* 18.
- Behjati, S., and Tarpey, P. (2013). What is next generation sequencing?. *Archives Of Disease In Childhood - Education & Practice Edition* 98, 236-238.
- Bogliolo, M., and Surrallés, J. (2015). Fanconi anemia: a model disease for studies on human genetics and advanced therapeutics. *Current Opinion In Genetics & Development* 33, 32-40.
- Chan, S., Lim, W., Ishak, N., Li, S., Goh, W., Tan, G., Lim, K., Teo, M., Young, C., and Malik, S. et al. (2017). Germline Mutations in Cancer Predisposition Genes are Frequent in Sporadic Sarcomas. *Scientific Reports* 7, 10660.
- Choi, M., Scholl, U., Ji, W., Liu, T., Tikhonova, I., Zumbo, P., Nayir, A., Bakkaloğlu, A., Özen, S., and Sanjad, S. et al. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings Of The National Academy Of Sciences* 106, 19096-19101.
- Cortes Barrantes, P., Jakobiec, F., and Dryja, T. (2019). A Review of the Role of Cytogenetics in the Diagnosis of Orbital Rhabdomyosarcoma. *Seminars In Ophthalmology* 34, 243-251.
- Couch, F., Nathanson, K., and Offit, K. (2014). Two Decades After BRCA: Setting Paradigms in Personalized Cancer Care and Prevention. *Science* 343, 1466-1470.
- Coverage depth - Metagenomics. (n.d.).
- Crucis, A., Richer, W., Brugières, L., Bergeron, C., Marie-Cardine, A., Stephan, J., Girard, P., Corradini, N., Munzer, M., and Lacour, B. et al. (2015). Rhabdomyosarcomas in children with neurofibromatosis type I: A national historical cohort. *Pediatric Blood & Cancer* 62, 1733-1738.
- Dasgupta, R., Fuchs, J., and Rodeberg, D. (2016). Rhabdomyosarcoma. *Seminars In Pediatric Surgery* 25, 276-283.
- DeVita, V.T. (2008). *Cancer: principles and practice of oncology* (Philadelphia, PA: Lippincott Williams & Wilkins).
- Diets, I., Waanders, E., Ligtenberg, M., van Bladel, D., Kamping, E., Hoogerbrugge, P., Hopman, S., Olderode-Berends, M., Gerkes, E., and Koolen, D. et al. (2018). High

Yield of Pathogenic Germline Mutations Causative or Likely Causative of the Cancer Phenotype in Selected Children with Cancer. *Clinical Cancer Research* 24, 1594-1603.

Egas Bejar, D., and Huh, W. (2014). Rhabdomyosarcoma in adolescent and young adult patients: current perspectives. *Adolescent Health, Medicine And Therapeutics*, 115-125.

Elkon, R., and Agami, R. (2017). Characterization of noncoding regulatory DNA in the human genome. *Nature Biotechnology* 35, 732-746.

Ewans, L., Schofield, D., Shrestha, R., Zhu, Y., Gayevskiy, V., Ying, K., Walsh, C., Lee, E., Kirk, E., and Colley, A. et al. (2018). Whole-exome sequencing reanalysis at 12 months boosts diagnosis and is cost-effective when applied early in Mendelian disorders. *Genetics In Medicine* 20, 1564-1574.

Farid, M., and Ngeow, J. (2016). Sarcomas Associated With Genetic Cancer Predisposition Syndromes: A Review. *The Oncologist* 21, 1002-1013.

Great Valley Publishing Company, I. (2021). How to Measure Duplicate Rates.

Guidugli, L., Carreira, A., Caputo, S., Ehlen, A., Galli, A., Monteiro, A., Neuhausen, S., Hansen, T., Couch, F., and Vreeswijk, M. (2013). Functional Assays for Analysis of Variants of Uncertain Significance inBRCA2. *Human Mutation* 35, 151-164.

Hassan, W., Alfaar, A., Bakry, M., and Ezzat, S. (2014). Orbital tumors in USA: Difference in survival patterns. *Cancer Epidemiology* 38, 515-522.

Hemminki, K., and Li, X. (2001). A population-based study of familial soft tissue tumors. *Journal Of Clinical Epidemiology* 54, 411-416.

How PCR duplicates arise in next-generation sequencing. (2012).

Hucl, T., Rago, C., Gallmeier, E., Brody, J., Gorospe, M., and Kern, S. (2008). A Syngeneic Variance Library for Functional Annotation of Human Variation: Application to BRCA2. *Cancer Research* 68, 5023-5030.

Koonin, E. V. and Galperin, M. Y. (2011). *Sequence - evolution - function: computational approaches in comparative genomics*. New York: Springer.

Kuhlen, M., Taeubner, J., Brozou, T., Wiczorek, D., Siebert, R., and Borkhardt, A. (2018). Family-based germline sequencing in children with cancer. *Oncogene* 38, 1367-1380.

Kuznetsov, S., Liu, P., and Sharan, S. (2008). Mouse embryonic stem cell-based functional assay to evaluate mutations in BRCA2. *Nature Medicine* 14, 875-881.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of Next-Generation Sequencing Systems. *Journal Of Biomedicine And Biotechnology* 2012, 1-11.

Lu, K., Wood, M., Daniels, M., Burke, C., Ford, J., Kauff, N., Kohlmann, W., Lindor, N., Mulvey, T., and Robinson, L. et al. (2014). American Society of Clinical Oncology Expert Statement: Collection and Use of a Cancer Family History for Oncology Providers. *Journal Of Clinical Oncology* 32, 833-840.

Lupo, P., Danysh, H., Plon, S., Curtin, K., Malkin, D., Hettmer, S., Hawkins, D., Skapek, S., Spector, L., and Papworth, K. et al. (2015). Family history of cancer and childhood rhabdomyosarcoma: a report from the Children's Oncology Group and the Utah Population Database. *Cancer Medicine* 4, 781-790.

Mahdieh, N., and Rabbani, B. (2013). An overview of mutation detection methods in genetic disorders. *Iranian Journal Of Pediatrics* 23, 375-388.

Malkin, D., Li, F., Strong, L., Fraumeni, J., Nelson, C., Kim, D., Kassel, J., Gryka, M., Bischoff, F., and Tainsky, M. et al. (1990). Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science* 250, 1233-1238.

Maxam, A., and Gilbert, W. (1980). Sequencing end-labeled DNA with base-specific chemical cleavages. *Methods In Enzymology* 65, 499-560.

Mueller, J., Schlappe, B., Kumar, R., Olvera, N., Dao, F., Abu-Rustum, N., Aghajanian, C., DeLair, D., Hussein, Y., and Soslow, R. et al. (2018). Massively parallel sequencing analysis of mucinous ovarian carcinomas: genomic profiling and differential diagnoses. *Gynecologic Oncology* 150, 127-135.

Nebbioso, A., Tambaro, F., Dell'Aversana, C., and Altucci, L. (2018). Cancer epigenetics: Moving forward. *PLOS Genetics* 14, e1007362.

O'Donovan, P., and Livingston, D. (2010). BRCA1 and BRCA2: breast/ovarian cancer susceptibility gene products and participants in DNA double-strand break repair. *Carcinogenesis* 31, 961-967.

Owosho, A., Huang, S., Chen, S., Kashikar, S., Estilo, C., Wolden, S., Wexler, L., Huryn, J., and Antonescu, C. (2016). A clinicopathologic study of head and neck rhabdomyosarcomas showing FOXO1 fusion-positive alveolar and MYOD1 -mutant sclerosing are associated with unfavorable outcome. *Oral Oncology* 61, 89-97.

Phred-scaled quality scores. (n.d.).

PolyPhen-2 score. (n.d.).

Poplin, R., Chang, P., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., and Afshar, P. et al. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology* 36, 983-987.

Principles and Workflow of Whole Exome Sequencing – CD Genomics. (n.d.).

Roberts, F., and MacDuff, E. (2018). An update on mesenchymal tumors of the orbit with an emphasis on the value of molecular/cytogenetic testing. *Saudi Journal Of Ophthalmology* 32, 3-12.

Ruymann, F., Maddux, H., Ragab, A., Soule, E., Palmer, N., Beltangady, M., Gehan, E., and Newton, W. (1988). Congenital anomalies associated with rhabdomyosarcoma: An autopsy study of 115 cases. A report from the intergroup rhabdomyosarcoma study committee (representing the children's cancer study group, the pediatric oncology group, the United Kingdom children's cancer study group, and the pediatric intergroup statistical center). *Medical And Pediatric Oncology* 16, 33-39.

Sanger, F., Nicklen, S., and Coulson, A. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings Of The National Academy Of Sciences* 74, 5463-5467.

Schadt, E., Turner, S., and Kasarskis, A. (2010). A window into third-generation sequencing. *Human Molecular Genetics* 19, R227-R240.

Sequencing Coverage for NGS Experiments. (n.d.).

Slatko, B., Gardner, A., and Ausubel, F. (2018). Overview of Next-Generation Sequencing Technologies. *Current Protocols In Molecular Biology* 122, e59.

Spunt, S., and Pappo, A. (2006). Childhood Nonrhabdomyosarcoma Soft Tissue Sarcomas Are Not Adult-Type Tumors. *Journal Of Clinical Oncology* 24, 1958-1959.

Stitzel, N. O., Kiezun, A. and Sunyaev, S. (2011). Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome Biology* 12, 227.

Sung, L., Anderson, J., Arndt, C., Raney, R., Meyer, W., and Pappo, A. (2004). Neurofibromatosis in children with Rhabdomyosarcoma: a report from the intergroup Rhabdomyosarcoma study IV. *The Journal Of Pediatrics* 144, 666-668.

Teer, J., and Mullikin, J. (2010). Exome sequencing: the sweet spot before whole genomes. *Human Molecular Genetics* 19, R145-R151.

Types of variants | Garvan Institute of Medical Research. (2018, November 14).

van Dijk, E., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends In Genetics* 30, 418-426.

Yingjie, L., Jian, T., Changhai, Y., and Jingbo, L. (2013). Numbl like regulates proliferation, apoptosis, and invasion of lung cancer cells. *Tumor Biology* 34, 2773-2780.

