UT GSBS Dissertations and Theses (Open Access)

Graduate School of Biomedical Sciences

8-2018

Omics Approaches to Uncover Germline and Somatic Variation Underlying Inherited Sarcomagenesis

Justin Wong

OMICS APPROACHES TO UNCOVER GERMLINE AND SOMATIC VARIATION UNDERLYING INHERITED SARCOMAGENESIS

by

Justin Wai-Chun Wong, B.S.

APPROVED:

____
Ralf Krahe, Ph.D.
Advisory Professor

____
Ken Chen, Ph.D

____
Guillermina Lozano, Ph.D.

____
Jeffrey Morris, Ph.D.

____
Nicholas Navin, Ph.D.

____
Louise Strong, M.D.

APPROVED:

____

Dean, The University of Texas
MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences

OMICS APPROACHES TO UNCOVER GERMLINE
AND SOMATIC VARIATION UNDERLYING
INHERITED SARCOMAGENESIS

A

DISSERTATION

Presented to the Faculty of

The University of Texas

MD Anderson Cancer Center UTHealth

Graduate School of Biomedical Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

by

Justin Wai-Chun Wong
Houston, Texas

August, 2018

OMICS APPROACHES TO UNCOVER GERMLINE AND SOMATIC VARIATION UNDERLYING INHERITED SARCOMAGENESIS

Justin Wai-Chun Wong, B.S.

Advisory Professor: Ralf Krahe, Ph.D.

Sarcomas are rare mesenchymal tumors, making up 15% of all childhood and 1% of all adult tumors. They account for a disproportionate share of mortality in young adults, and if left untreated, are highly likely to metastasize. However, sarcoma etiology is poorly understood, and having numerous histological subtypes has complicated elucidation. To better understand factors underlying sarcomagenesis, we leveraged two rare inherited cancer predisposition syndromes, Li-Fraumeni Syndrome (LFS), and LFS-like (LFSL), both with a high incidence of sarcomas. LFS is caused by mutations in the tumor suppressor gene *TP53 (p53),* but has variable and incomplete penetrance, suggesting additional acquired somatic mutations are necessary for tumorigenesis. In contrast, LFSL has no known cause, although a 10-Mb region in 1q23 has been mapped by linkage analysis as a putative LFSL locus. Therefore, to better identify genetic variation underlying LFS and LFSL we utilized a 2-pronged approach. First, we evaluated LFSL families for rare, co-segregating, germline mutations, which identified a mutation *in ARHGAP30* that was present in four LFSL families. Moreover, this mutation impacted both proliferation and migration when overexpressed *in vitro*. Subsequent analysis of publicly available data indicates a potential role for ARHGAP30 in sporadic cancers. Secondly, we endeavored to identify somatically acquired drivers of sarcomagenesis. In cancer, passenger events are acquired concomitantly with driver mutations, and distinguishing them remains a key challenge. To best address this, we used a comparative genomics approach, leveraging a “humanized” mouse model of LFS with a hotspot mutation, *Trp53*^{R172H}*,* analogous to *TP53*^{R175H} in humans. Hypothesizing that sarcoma etiology is similar in humans and mice, we then catalogued recurrent changes in the genome, transcriptome, and methlyome. We found little overlap in any of the omics approaches across the human tumors, which came from diverse *p53* mutations and sarcoma types, but found strong overlap in the mouse tumors (fibrosarcomas and osteosarcomas). Recurrent data discovered in the mouse was mirrored in some human sporadic mesenchymal tumors, including novel genes like *MROH2A, and MIR219A2.* Our results emphasize the utility of a model disorder and comparative omics to uncover genes with relevance for both inherited and sporadic tumors.

Acknowledgements

When I first started this project, I wasn't sure where to begin. I came from an analytical, biomedical engineering background with experience doing bench-to-bedside research to drive an existing technology into the clinic. However, given a prior, and personal, interest in Neurofibromatosis I, it made sense to pursue a project that married genetics and functional testing. I have been thankful for the opportunities to learn entirely new skills, ranging from implementing sequencing analysis pipelines, to techniques in molecular genetics.

I would also like to thank Dr. Ralf Krahe for his mentorship during this time; it was of tremendous benefit to me to be able to lean on him for advice personally and professionally, and to see how we were able to use our respective areas of expertise to push the project further along. I was able to apply engineering approaches to the project, while leaning on him to continue to develop a foundation in genetics.

This research would not have been possible without the gracious contributions of numerous clinical collaborators, including Dr. Louise Strong, Phyllis Begin, Jasmina Bojadzieva from M.D. Anderson Cancer Center, Dr. Albert de la Chapelle and Ms. Heather Hampel From Ohio State University, Dr. Henry Lynch and Carrie Snyder from Creighton University, and Dr. Mai Phuong and Dr. Sharon Savage from the National Cancer Institute. We are eternally grateful for their efforts to catalog and collect samples from LFS and LFSL families and for letting us use them.

We are further grateful to Dr. Guillermina Lozano for the use of her *Trp53*^{R175H} mouse model and to Dr. Elizabeth Whitley (Pathogenesis) for diagnosing the pathology of

our tumors. In addition, we had numerous other collaborators at M.D. Anderson Cancer Center. We worked with Dr. Wenyi Wang and Dr. Gang Peng on how to best implement FamSeq in our pipeline, and Dr. Keith Baggerly and Dr. Ying Wang for much of the TCGA and statistical analysis. We are thankful to Dr. Nicholas Navin and Naveen Ramesh for their help with the copy number analysis. Dr. Yue Lu and Marcos Estecio were instrumental in generating and analyzing the mouse methylation data (RRBS). We owe a tremendous thanks to Dr. Lou Ramagli and the SMF for sequencing nearly all of our samples as well as to Dr. Adriana Paulucci for her help with live cell imaging. I am grateful to my summer interns, especially Kevin Tracy and Jeffrey Hsu for their help running experiments and counting cells.

Separately, I would like to thank other members of the Krahe Lab. Dr. Mario Sirito for being an ever-present source of guidance and help, especially for mouse work and other bench top experiments. Dr. Yu (Ruby) Deng and Dr. Qin Chen for their help running Westerns and qPCRs when I had minimal experience and for their encouragement to always be thinking about experimental end points. No list of Krahe Lab members would be complete without mentioning Dr. Linda Bachinski, whose clear, concise communication was always something I aspired to, and upon whose work was the basis of the non-*p53* LFS project.

I am also indebted to my committee members: Dr. Louise Strong, Dr. Nick Navin, Dr. Guillermina Lozano, Dr. Paul Scheet, Dr. Jeffrey Morris, and Dr. Ken Chen (with of course Dr. Ralf Krahe), for serving on my committee over several years, and for their valuable feedback and critiques. This project would not have been possible without them. I am further grateful to Dr. Gil Cote, Dr. Joya Chandra, Dr. Jeffrey Morris, Dr. Nick Navin, and Dr. Louise Strong for serving on my examining committee, and helping me get through what was the most stressful times I've ever experienced.

This work was supported by an NIH training grant (T32 Training Program in Molecular Genetics of Cancer Pre-doctoral Fellow (T32 CA009299)) and a CPRIT grant (RP130054)

Lastly, but certainly not least, I would also like to acknowledge my family for their love and support throughout the process.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

1KG – 1000 genomes project

BWA – Burrows-wheeler aligner

CCLE – Cancer cell line encyclopedia

ChIP – Chromatin immunoprecipitation

CN – copy number

CNA – Copy number aberrations

E(-rich) – glutamic acid(-rich)

ESP – Exome sequencing project

FFPE – formalin fixed paraffin-embedded

FG6 – Fugene 6

FR – Femur

FS – Fibrosarcoma

GATK – Genome analysis toolkit

GOF – Gain of function

HLOD – Heterogeneity logarithm of the odds

hOB – human osteoblasts

HS2000/4000 – Hi-seq 2000/4000

IR – Irradiation

L-ARHGAP30 – long isoform of ARHGAP30

L/N – license number

lcWGS – low-coverage whole genome sequencing

LFS – Li-Fraumeni Syndrome

LFSL – Li-Fraumeni Syndrome-like

LMS – Leiomyosarcoma

LOF – Loss of function

LOH – loss of heterozygosity

LPF – Lipofectamine 3000

LPS – Liposarcoma

MAF – Minor allele frequency

NGS – Next generation sequencing

OS – Osteosarcoma

PBL – Peripheral blood leukocytes

PKD – Pyruvate kinase deficiency

PP2 – Polyphen2

QC – quality control

qRT-PCR –

ROH – retention of heterozygosity

RPPA – Reverse phase protein array

S-ARHGAP30 – short isoform of ARHGAP30

SIFT – Sorting Intolerant from Tolerant

SNP – Single Nucleotide Polymorphism

SNV – Single nucleotide variant

STLMS – Soft-tissue leiomyosarcoma

STS – Soft-tissue sarcoma

TCGA – The Cancer Genome Atlas

TSG – tumor suppressor gene

ULMS – gynecologic/uterine leiomyosarcoma

UTR – Untranslated region

VL – Vastus lateralis

WES – Whole exome sequencing

WGS – Whole genome sequencing

WHO – World Health Organization

WT – wild-type

TCGA cancer types

BLCA - Bladder Urothelial Carcinoma

BRCA - Breast invasive carcinoma

CESC - Cervical squamous cell carcinoma and endocervical adenocarcinoma

COAD - Colon adenocarcinoma

DLBC - Lymphoid Neoplasm Diffuse Large B-cell Lymphoma

GBM - Glioblastoma multiforme

HNSC - Head and Neck squamous cell carcinoma

KICH - Kidney Chromophobe

KIRC - Kidney renal clear cell carcinoma

KIRP - Kidney renal papillary cell carcinoma

LGG - Brain Lower Grade Glioma

LIHC - Liver hepatocellular carcinoma

LUAD - Lung adenocarcinoma

LUSC - Lung squamous cell carcinoma

OV - Ovarian serous cystadenocarcinoma

PAAD - Pancreatic adenocarcinoma

PCPG - Pheochromocytoma and Paraganglioma

PRAD - Prostate adenocarcinoma

READ - Rectum adenocarcinoma

SARC - Sarcoma

SKCM - Skin Cutaneous Melanoma

STAD - Stomach adenocarcinoma

THCA - Thyroid carcinoma

UCEC - Uterine Corpus Endometrial Carcinoma

UCS - Uterine Carcinosarcoma

Pedigree-specific

AML – Acute myeloid leukemia

BL – Bladder cancer

BR – Breast cancer

CA – Carcinoma

H&N – Head and neck cancer

HD – Hodgkin’s disease

LU – Lung cancer

NHL – Non-Hodgkins Lymphoma

OST – Osteosarcoma

OV – Ovarian cancer

RCC – Renal cell carcinoma

RMS – rhabdomyosarcoma

SC – Sarcomatoid carcinoma

WT – Wilms’ Tumor

1 Introduction

1.1 Introduction to cancer

Cancer is a multifactorial genetic and epigenetic disease characterized by abnormal, unchecked cell growth. Cells typically progress gradually through a series of steps, from hyperplasia, to dysplasia, to neoplasia, and sometimes malignancy, usually driven by genetic changes, that disrupt core processes that normally help regulate cell growth. These processes have been categorized into ten distinct functions, known as "the Hallmarks of Cancer." Initially, Hanahan and Weinberg postulated that cancer cells might have additional capabilities across six different categories: sustained growth signaling, evasion of growth suppressors, resistance to cell death, induction of angiogenesis, enabling of replicative immortality, and activation of invasion and metastasis (Figure 1) (1).

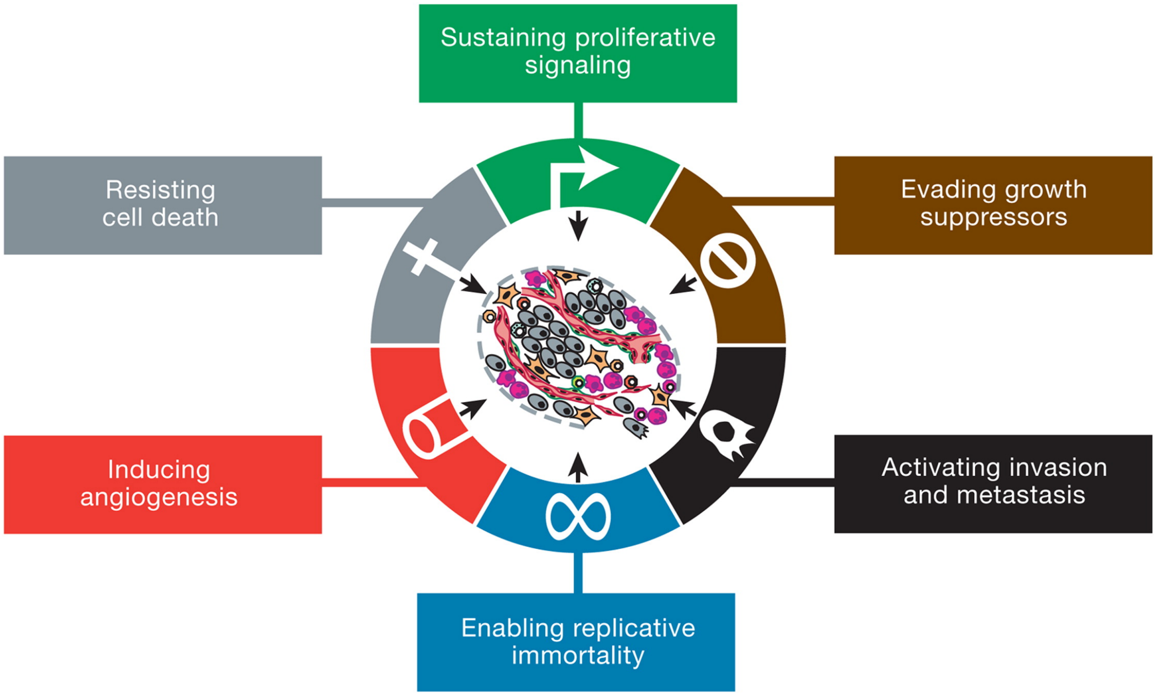

Sustaining proliferative signaling
Evading growth suppressors
Activating invasion and metastasis
Enabling replicative immortality
Inducing angiogenesis
Resisting cell death

Figure 1: Hallmarks of Cancer. Original six canonical ways that cancer cells have added abilities (2). Used with permission: Hanahan, D., and R. A. Weinberg. 2011. Hallmarks of cancer: the next generation. *Cell* 144: 646-674. L/N 4294491470430.

Subsequent research has strongly indicated two other hallmarks: deregulating cellular energetics, and avoiding immune destruction, and two enabling characteristics: genome instability and mutation, and tumor-prone inflammation (Figure 2) (2).

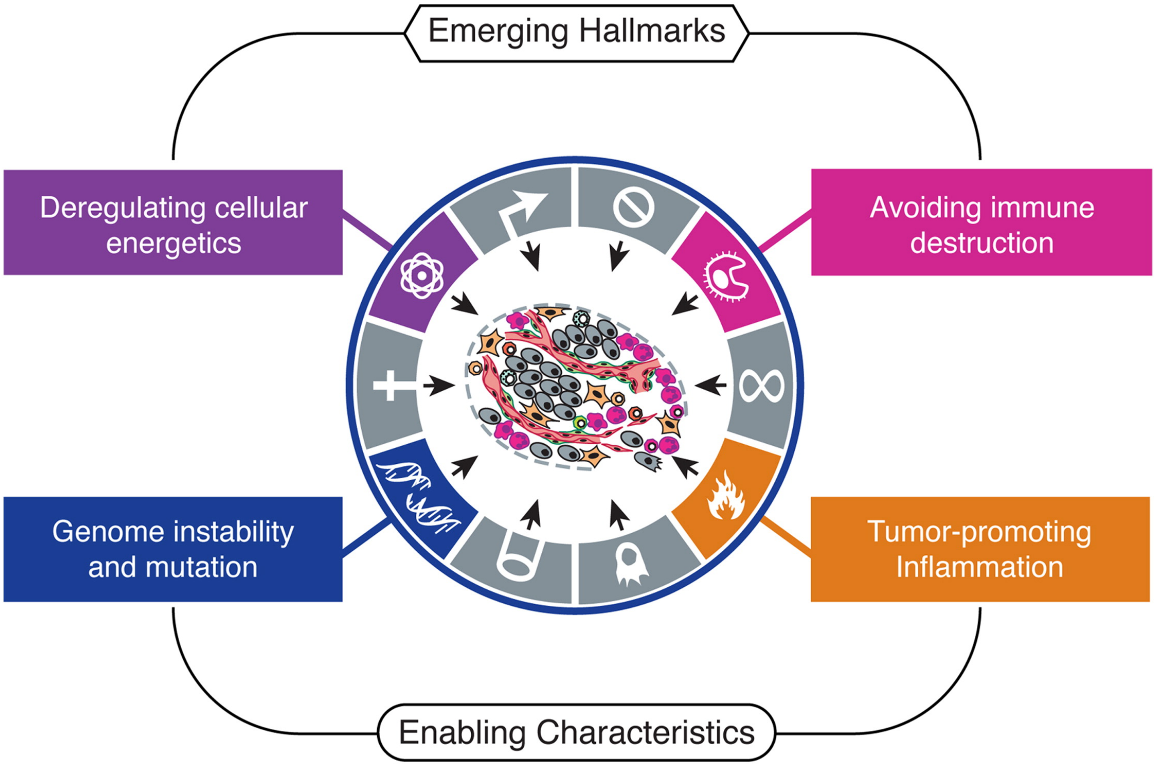

Emerging Hallmarks
Deregulating cellular energetics
Avoiding immune destruction
Genome instability and mutation
Tumor-promoting Inflammation
Enabling Characteristics

Figure 2: Hallmarks of Cancer, updated. In 2011, the hallmarks were updated to include two new emerging hallmarks, and two enabling characteristics (2). Used with permission: Hanahan, D., and R. A. Weinberg. 2011. Hallmarks of cancer: the next generation. *Cell* 144: 646-674. L/N 4294491470430.

These ideas provide a framework for understanding how tumors might arise, and how they differ from normal cells.

The changes necessary for tumorigenesis can be acquired in a variety of ways. In cancer predisposition syndromes, germline mutations can be inherited from the parents. Alternatively, changes in the genome can be acquired somatically, either from errors during cell division, or from exposure to environmental mutagens such as smoking or UV rays. For example, in microsatellite instability syndromes, such as Lynch Syndrome, during replication, defects in mismatch repair genes lead to DNA polymerase slippage, frameshift mutations and non-functional proteins (3). However, not all such changes may be cancer-causing; some alterations that are acquired may instead be passengers, with limited or even no functional impact. Distinguishing cancer-causing driver mutations from passenger events remains a key challenge in understanding cancer etiology (4).

Beyond just looking for genetic changes in the genome, we increasingly find that changes in the transcriptome and epigenome can also lead to disease, including cancer (5-7). Moreover, this makes integrative approaches that seek to completely profile tumors across the genome, transcriptome, and epigenome uniquely powerful to quantify, and evaluate as many changes as possible (7). For example, pairing whole genome sequencing (WGS) and profiling by RNA-seq in breast cancer allows scientists to determine if somatic point mutations are being highly expressed (8). Our goal is to identify any such changes that may be important in the genesis of cancer, and in sarcomas in particular, in the hopes that they may prove useful therapeutically.

1.2 Introduction to Sarcomas

Sarcomas are a relatively rare mesenchymal cancer, making up ~1% of all adult tumors and 15% of all childhood tumors (9-11). They are most often found in the arms or legs (60%), and chest or abdomen (30%), but they can be either soft tissue or bone depending on the location (11).

The large majority of soft tissue and bone sarcomas do not have a known causative factor. There are some isolated reports that suggest that soft tissue sarcomas may arise near scar tissue, or at fracture sites, or that sarcomas of the bone may be more likely to arise from bone infarctions or radiation injury, but these appear to be the exception rather than the norm. The foremost known causes are related to genetic susceptibility (which still represent a relatively small number of tumors) such as individuals with germline *TP53 (p53), RB1,* or *NF1* mutations, but for the most part, the etiology of sarcomas, genetic or otherwise, remains a relative mystery (11).

1.2.1 Sarcomas and clinical outcomes

Treatment for patients with advanced-stage sarcomas has not changed dramatically in the past 30 years (12). The primary options for treatment are surgery, radiation, chemotherapy, and targeted therapy. Surgery remains the most commonly accepted therapeutic option, but requires removal of healthy tissue with margins as wide a 1 cm because even microscopically small portions have been associated with recurrence, metastasis and death (11). Under ideal conditions with low grade sarcomas, and removal of all tumorigenic tissue, local control is about 93% (9). Clinicians are still searching for a consensus as to the best chemotherapy and radiotherapy options (11), raising the possibility that better molecular profiling to inform treatment options could improve patient outcomes.

Without treatment, sarcomas are highly likely to metastasize and are considered highly aggressive, accounting for a disproportionate share of mortality in young adults. (SEER Cancer Statistics Review 1975-2008; http://www.seer.cancer.gov/csr/1975_2008/)(11, 13). In fact, the mean age of onset for sarcomas, in part due to hereditary syndromes, is earlier than for many other types of cancer (14). Therefore, identification of sarcoma-related genes, either germline variants, or acquired somatic variants, is considered vitally important for diagnostic testing to help identify at-risk individuals, and for uncovering and developing potential therapeutic avenues.

1.2.2 Challenges in understanding sarcoma etiology

One distinct challenge in identifying drivers of sarcomagenesis is the sheer number of histological subtypes of sarcomas. According to the World Health Organization (WHO), there are over 80 subtypes of soft tissue sarcomas, and over 60 subtypes of bone sarcomas (11). Because sarcomas are both rarer tumor types, and because they comprise such a heterogeneous grouping of tumors, it is challenging to isolate enough tumors of similar subtyping to generate sufficient statistical power to identify drivers, often requiring extensive collaborations. Moreover, these data are further complicated because not only are sarcomas very diverse (numerous subtypes), but also each sarcoma is heterogeneous in composition. Sarcomas are frequently composed of bone, cartilage, and fat (13), which may further obfuscate and hinder identification of important genes for tumorigenesis, particularly if a driver gene need only be present in one of these tissues.

Two prominent papers were recently published on soft tissue sarcomas (STS). In a *Cell* paper, out of the Cancer Genome Atlas network (TCGA), the authors sequenced six types of soft tissue sarcomas, the majority of which are leiomyosarcomas (LMS) and liposarcomas (LPS) (10). LMS were again a focus in the second paper, from Chudasma et al. and published in *Nature Communications* (15). Both studies found that adult STSs have heterogeneous mutational profiles with copy number aberrations (CNA) consistent with chromothripsis being a common occurrence. CNAs tended to be deletions, rather than amplifications, and the authors found relatively few point mutations (10, 15).

In particular, the data from TCGA suggested that some alterations may be sarcoma-subtype specific. These differentiating features can range anywhere from genomic changes (e.g. CNAs and point mutations), to changes in the transcriptome, methlyome, or protein levels. For example, most LMS have elevated signaling in PI3K/AKT signaling, and over 70% have at least shallow deletions in *TP53, RB1,* and *PTEN*. The majority of LMS also have elevated miR-143, and miR-145. However, two subtypes of LMS, gynecologic LMS (ULMS), and soft-tissue LMS (STLMS) have distinct differences in the methylome and in reverse phase protein array (RPPA) analyses. ULMS showed hypomethylation of ESR1 target genes, something not seen in STLMS. RPPA showed that the DNA damage pathway was more active in ULMS over STLMS, but that the HIF1 inflammation pathway was more active in STLMS over ULMS (10).

Taken together, these data implicate CNAs as a key player in both LMS and LPS, and indicate sarcoma subtypes have distinct molecular profiles, which may drive therapeutic approaches for clinicians in the future. Ultimately, the data suggest that all sarcomas are not likely to share the same etiology and that sequencing of similar sarcomas may be necessary to avoid confounding from multiple types of sarcomas.

However, the authors do note a role for point mutations in soft tissues sarcomas. In addition to CNAs frequently occurring somatically in the MDM2-p53 and p16-CDK4-RB1 pathways, the most recurrently mutated genes across sarcoma types in both studies were a triumvirate composed of *TP53, ATRX,* and *RB1*, further suggesting that these genes (and their pathways) play important roles in sarcomagenesis (10, 15).

1.2.3 Sarcomas and multiple germline drivers

The thrust of the previous works by the TCGA (10) and Chudasama et al. (15) focus on acquired somatic changes. A third, slightly older study, by Ballinger et al. looked at germline genetic risk factors for sarcomas (16). These included a mix of both sporadic and familial sarcomas, across a broad spectrum of sarcoma types, including soft tissue and bone sarcomas, with publically available Caucasian data used as controls to eliminate polymorphic alleles from consideration. Using targeted exon sequencing of 72 genes, selected for known impact in cancer, Ballinger et al. found that risk generally fell into two groups: classic monogenic variation (~80%), such as p53 (1% of all sarcomas), and polygenic rare variation (~20%). Individuals classified as having polygenic rare variation had comparable tumor-free survival to monogenic *p53* variants, suggesting that two so-called weaker effects can make up for one big one (16).

Given the limited gene set from targeted sequencing, Ballinger et al. do not identify any specifically novel genes, but do observe 2% of sarcoma-patients to have rare germline mutations predicted to be damaging in *ERCC2*. ERCC2 is a helicase involved in base excision repair; the authors argue that it should now be considered a sarcoma susceptibility gene (16). Ultimately, more than half of the patients had variants predicted to be deleterious, and 40% of these (one-fifth overall) had mutations with known pathogenicity, suggesting that additional genetic risk factors for sarcomagenesis have yet to be discovered.

1.3 Introduction to *p53*

p53 is arguably the most important tumor suppressor gene and has been called the "guardian of the genome" (17), and more recently, the "guardian of the epigenome" (18). Over half of all cancers have alterations in it, and it is considered the most mutated gene in human cancers (19). It has roles in numerous cancer-related processes, including cell-cycle arrest and apoptosis. Moreover, germline mutations and deletions in *p53* lead to Li-Fraumeni Syndrome (LFS), a rare cancer predisposition syndrome that has a high incidence of sarcomas, suggesting that *p53* may play a role in sarcomagenesis (20, 21). Probably in part due to LFS, it is considered to be the strongest monogenic driver of sarcomas (16). Conversely, based on publicly available data from cBio, *p53* alterations are present in less than 60% of sporadic sarcomas, and as low as 20%, (10, 15, 22-25) implying that disrupted *p53* may not be required for their formation, and that sarcoma etiology may be considerably varied. More specifically, sarcomagenesis may occur by two divergent mechanisms, one that is *p53*-mediated, and one that is *p53*-independent, and consideration of only one of these mechanisms, such as in LFS tumors, with germline *p53* drivers, may improve detection of additional sarcoma risk factors. Normally, *p53* abrogates tumor growth in part by helping cells to sense cellular stresses such as hypoxia and DNA damage, and limit cell proliferation under conditions where genomic integrity is likely to be compromised. However, when *p53* is lost, the loss of these protective aspects can lead to the accumulation of oncogenic mutations, as well as unchecked cell proliferation, leading to a positive feedback loop where these populations expand more rapidly, resulting in tumorigenesis (17, 26).

1.3.1 *p53* and *MDM2*

p53 is known to be regulated by a variety of regulators, including the negative-regulator *MDM2.* MDM2 is an E3-ubiquitin ligase that specifically ubiquitinates *p53*, exporting it out of nucleus, and marking it for degradation (27-29). Moreover, *MDM2*, as a *p53*-inducible gene, is closely correlated with *p53* levels in normal cells. When working appropriately, these combine to form an auto-regulatory loop, designed to maintain low levels of *p53* in the absence of stress. On the flip-side, during periods of stress and DNA damage, *p53* and *MDM2* are both phosphorylated, preventing their interaction with each other, thus stabilizing *p53* (30-34). In some cases, this stabilization has been shown to be an important step in tumorigenesis, particularly when there is mutant *p53*, such as in LFS (35).

In addition, work by several groups has implicated a polymorphism in *MDM2* (SNP309) as a risk factor across several cancer types, including colorectal, breast, lung, and brain among many others (36-39). In conjunction with LFS or in sporadic sarcomas, the presence of the *MDM2 SNP309* polymorphism appears to accelerate tumor formation (37, 40-42). The results of meta-analyses have only continued to affirm that there is evidence for association between *MDM2 SNP309* and the *p53 R72P* polymorphism, suggesting that *MDM2* may act as a modifier gene for tumorigenesis (29, 37).

1.3.2 *p53* gain-of-function mutations

Although the majority of mutations in *p53* are loss-of-function (LOF), several groups have demonstrated that, contrary to expectation based on other tumor suppressor genes, which only have LOF mutations, some *p53* mutations result in gain-of-function (GOF) (43). This idea was first noted when some tumors with point mutations in *p53* were found to have elevated levels of *p53* in cancer cells relative to controls (43, 44), suggesting that *p53* must have acquired additional properties capable of aiding, rather than hindering tumor progression. The majority of results have been established by overexpressing mutant *p53* in *p53-null* cells, with GOF features associated with elevated resistance to apoptosis (45-48), cell migration and invasion (49, 50), or alternatively with cancer in animal models (43, 51, 52).

1.3.3 *p53* mutation incidence

The majority of tumor-relevant mutations in *p53* occur in the DNA binding domain (DBD), which comprises exons 5-8 of the gene. The World Health Organization (WHO) based IARC database, which collects and compiles published data with *p53* mutations at the germline, and at the somatic level, show several hotspots (53).

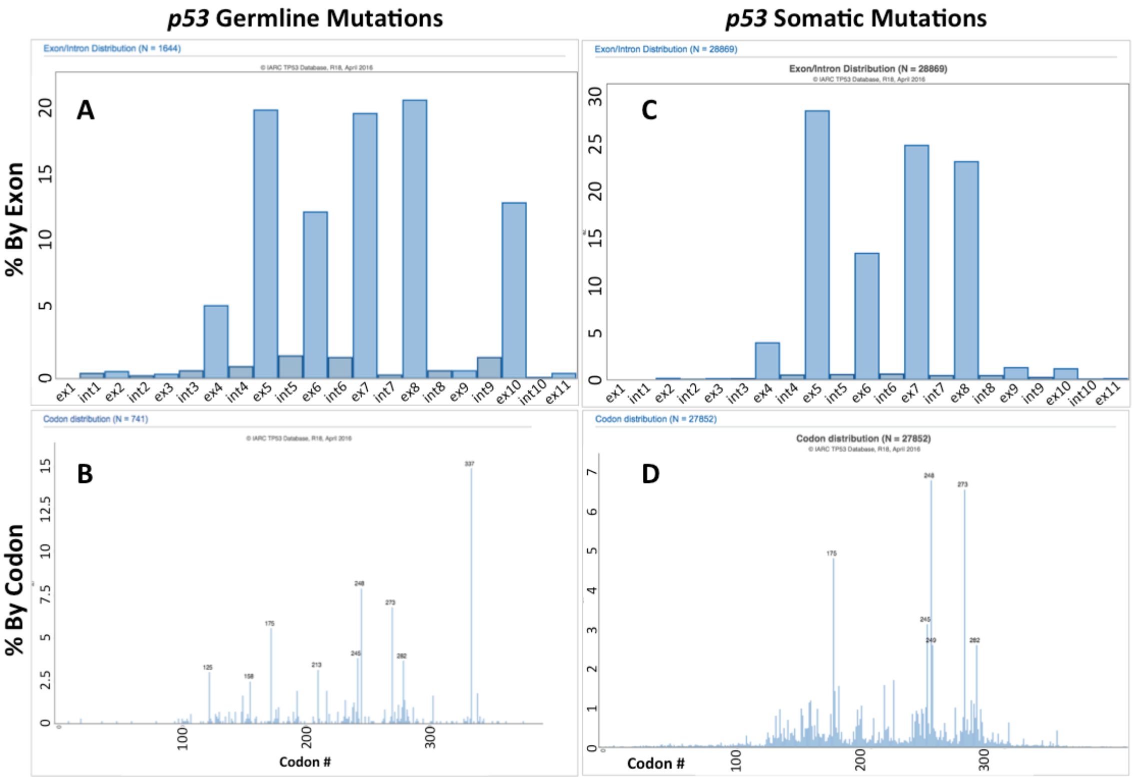


Figure 3: Compilation of published data containing *p53* mutations, as generated by the IARC database (53). Data on the left (A,B) composed of pedigrees with germline *p53* mutations. Data on the right (C, D) represents individuals with somatic mutations in *p53*. Data indicate that for both germline and somatic variants, most occur between exons 5-8, and appear to have similar hotspots to one another.

1.3.4 Hotspot mutations in *p53*

These data also suggest one other way in which *p53* differs from most other tumor suppressor genes. In general, the mutational landscape for tumor suppressor genes and oncogenes are considered to be different; tumor suppressor genes tend to have flat mutational profiles, while oncogenes tend to have profiles with distinct, sharp peaks, known as hotspot mutations (54). For example, in common tumor suppressor genes (TSG) like *RB1* and *NF1*, the profiles are relatively flat (Figure 4). In contrast, for oncogenes such as *BRAF*, there are often hotspot mutations, i.e. nucleotides that are frequently mutated (Figure 4). Thus, these data do not rule out the possibility that other cancer predisposing TSG will not have mutation hot spots like *TP53,* or that some of these variants may be GOF mutations.

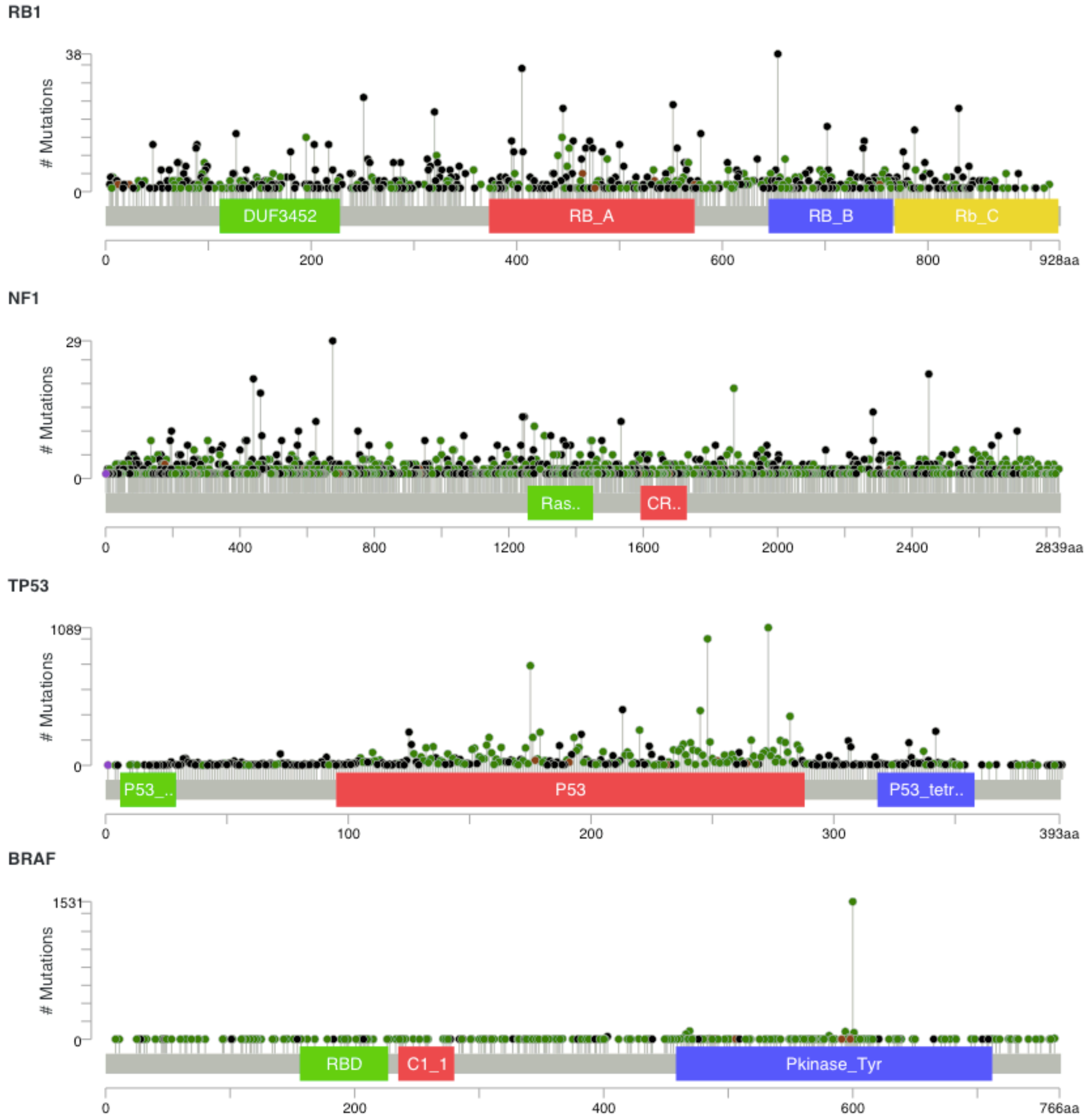


Figure 4: Lollipop diagrams from cBio (22) depicting the mutation frequency across four genes. Three are tumor suppressors (*RB1, NF1,* and *TP53*) and one is an oncogene *(BRAF)*. The profiles for the first two TSG are representative of most TSG. The third TSG, *TP53*, appears to have mutation hot spots that are more consistent with a classic oncogene, such as *BRAF.* Note that the scales are considerably different across the four genes.

2 Li-Fraumeni Syndrome (LFS) and LFS-like (LFSL)

2.1 LFS

In order to better understand the underlying genetics and epigenetics behind sarcomagenesis, we propose to use LFS as a model disorder. LFS is a rare, inherited, heterogeneous, cancer predisposing syndrome caused by mutations in the tumor suppressor gene *p53* (70-80% of cases) (55, 56) with a high prevalence of sarcomas (20, 56, 57, 58{, 59). In a classic LFS pedigree (Figure 5), we see characteristic patterns of autosomal dominant inheritance (cancer is observed in every generation) and anticipation (age of onset for cancer gets younger for each generation, ranging from a lung cancer at age 61 in the oldest generation, to a variety of cancers from roughly 30-50 years old in the second generation, to sarcomas in the first two decades of life in the third generation).

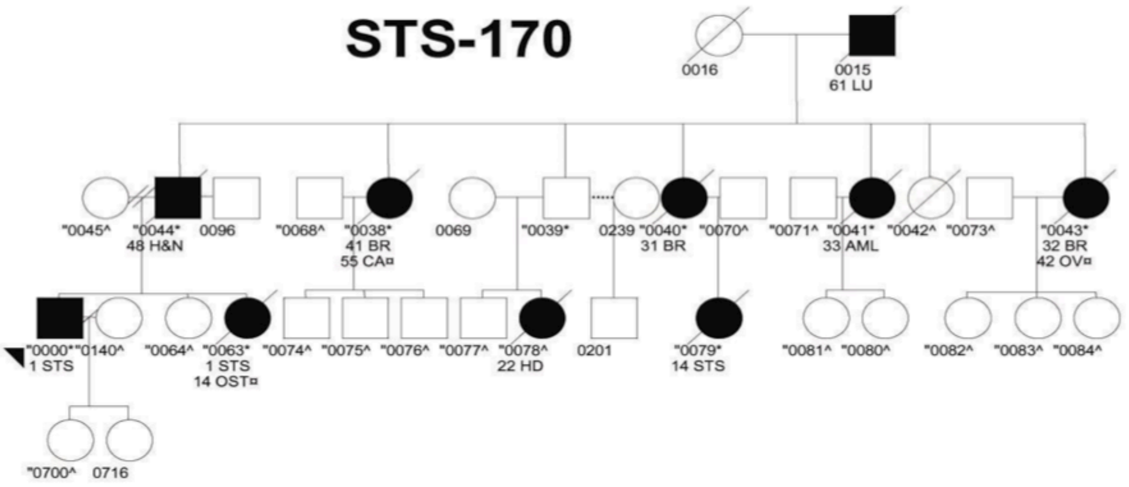


Figure 5: Canonical LFS pedigree with a *TP53 M133T* mutation. The pedigree shows characteristic patterns of autosomal dominant inheritance, anticipations, a broad tumor spectrum, and a high prevalence of sarcomas, including a soft tissue sarcoma in the proband, denoted by an arrow. Shaded circles represent cancer, with the age of diagnosis

and cancer type noted below. *p53* mutation carriers are denoted by a “*”, and individuals with *WT p53* are denoted by an “^”.

LFS is further identified by the high prevalence of sarcomas, which account for about 25% of all LFS-tumors (Figure 6) (53). The remainder of tumors seen in LFS covers a diverse spectrum, including breast cancer, brain cancer, lung cancer, and adrenocorticoid cancer (Figure 6) (53, 56). Given that *p53* is considered to be the most preeminent cancer gene (17, 19), it is hardly a surprise that *p53* germline mutations predispose to so many types of cancer.

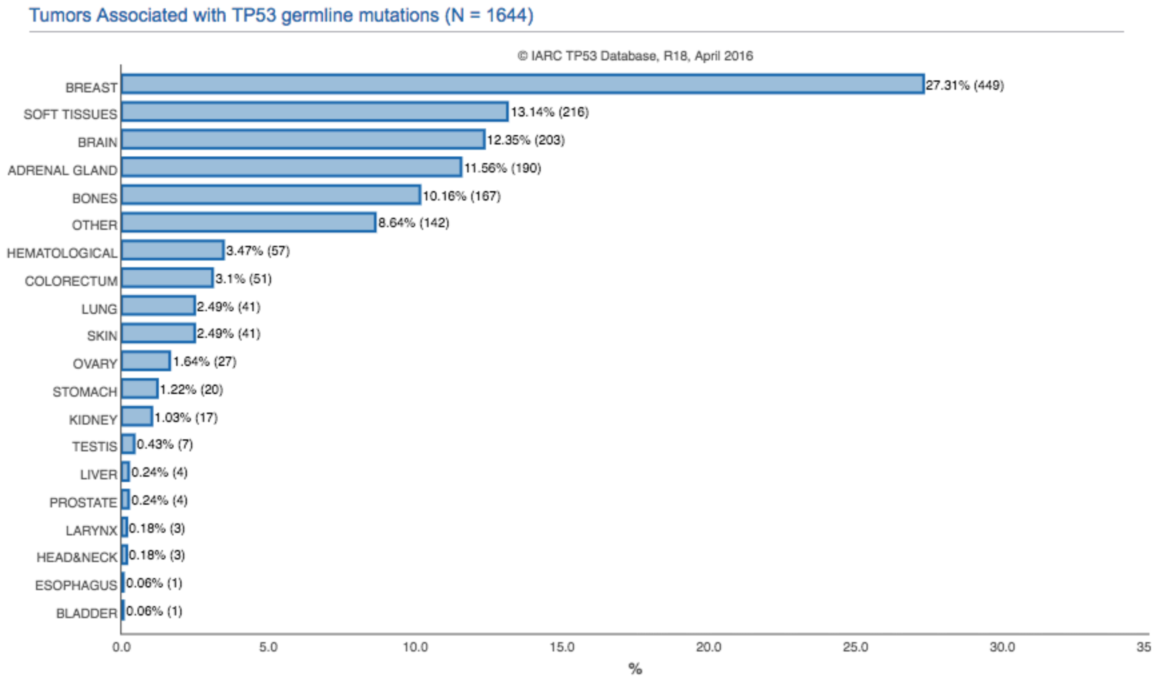


Figure 6: Tumor incidence by type as compiled by the IARC *p53* database for individuals with germline *p53* mutations (53).

2.1.1 Clinical criteria for LFS

Several different clinical criteria have emerged for LFS over the years.

1. Classic LFS, first defined in 1988 (60), requires three criteria to be met:
    - Proband with a sarcoma before the age of 45
    - First-degree relative with cancer before the age of 45
    - Additional first degree relative with cancer before the age of 45, or a sarcoma at any age
2. Chompret LFS (61, 62) requires one of the following to be met. Unless otherwise stated, tumors in the LFS spectrum are considered to be: soft tissue sarcoma, osteosarcoma, pre-menopausal breast cancer, brain tumor, adrenal cortical carcinoma, leukemia, or lung cancer
    - Tumor belonging to LFS spectrum before the age of 46 AND at least one first-degree or second-degree family member with an LFS-related tumor (except breast cancer if the individual has breast cancer), before the age of 56, or with multiple tumors
    - A person with multiple tumors, two of which belong to the LFS spectrum (excluding multiple breast cancers), the first of which occurs before the age of 46.
3. Eeles definition (63)

- Two first-degree or second-degree relatives LFS-related tumors at any age

4. Birch definition (64) requires three criteria to be met:

- Proband with childhood cancer, sarcoma, brain tumor, or adrenal cortical tumor before the age of 45
- First- or second-degree relative with a tumor in the LFS spectrum at any age
- First- or second degree relative with any cancer before the age of 60

Individuals and families meeting these criteria turn out to have genetic alterations in *p53 (p53*-LFS) in about 70-80% of cases (55).

2.2 Evidence for a Li-Fraumeni Syndrome-like disorder

However, several families that meet the clinical criteria for LFS appear to lack mutations or alterations in *p53* (56, 65), suggesting the presence of one or more additional cancer/sarcoma-predisposition gene(s). Families meeting the LFS criteria are tested across a barrage of tests to definitively rule out alterations in *p53,* including sequencing and testing for copy number aberrations. Notably, despite phenotypic similarities, the general consensus seems to be that by definition LFS and LFSL should be considered distinct from one another, that is, LFS-carriers must contain a *p53* mutation (56, 66).

2.3 Alternative risk factors for LFS/LFSL

2.3.1 CHEK2

To date, no other mutations have been definitively associated with LFSL (56). In 1999, Bell et al. advanced the idea that *CHEK2* could be a second LFS gene (67). They identified a specific *CHEK2* mutation (*CHEK2 1100delC*) that co-segregated with disease in a family that met the criteria for classical LFS, but lacked a *p53* mutation, as well as two other families with different alterations in *CHEK2* (67). Moreover, CHEK2 was known to be a cell cycle checkpoint kinase that interacted with, and stabilized p53, and the mutation itself, which resulted in a premature stop codon, was found to abolish the kinase function of the CHEK2. Combined, these initial data argued strongly for the probability that *CHEK2* could be an LFSL cancer predisposition gene.

However, subsequent data accumulation on LFSL families did not find *CHEK2 1100delC* to be a common cause of LFSL (68). Sequencing of additional LFSL families revealed few families with *CHEK2* alterations anywhere in the gene (69). Additional studies failed to detect any mutations in *CHEK2* in LFSL families across 48 total families (70, 71). A fourth study, by Lee et al. in 2001 found three missense variants across 10 LFS and 49 LFSL pedigrees, including a polymorphism (Ile157Thr), and two that were somewhat rarer (Arg145Trp in a patient with breast cancer and a sarcoma, and Arg3Trp in a patient with brain cancer) (72). Sodha et al. found *CHEK2* variants in 3 of 26 families, but these included a synonymous variant, an intronic variant which does not appear to impact splice sites, and a 3-bp deletion in exon 3, thus continuing to suggest *CHEK2* may not be an LFSL cancer predisposition gene (69).

Expanding the pedigrees tested beyond LFS/LFSL for *CHEK2* mutations, including in familial breast cancer cohorts, did find excess risk for several cancers, including prostate (73), colon (74), kidney (73), and breast cancer (74, 75), of which only breast is a canonical LFS/LFSL tumor (56). However, they did not find similar upticks in sarcomas and adrenal cortical tumors (73). The variant is fairly common in the general population (MAF is about 1%), and therefore present in some unaffected women, leading to the premise that it is a low-penetrance breast cancer risk allele. Moreover, the deletion is enriched in breast cancer families that are *BRCA1/BRCA2*-negative (74). Therefore, despite some ambiguity in the literature, the general consensus has emerged that *CHEK2* should no longer be considered a cause of LFS/LFSL (76).

2.3.2 p53 UTR

The majority of LFS-related research has focused on coding mutations in *p53*. However, recently, Macedo et al. reported a rare germline mutation (rs78378222) in the 3'UTR of *TP53* that was found in 7 LFSL probands (5.4%) and was correlated with reduced expression of *p53* (77).

2.3.3 Linkage indicates 1q23 contains an LFSL locus

An additional locus for LFSL was mapped by linkage to 1q23 (65). In this study, linkage analysis using microsatellite markers was completed across 62 constitutive DNA samples over four LFSL pedigrees, and mapped a 10-Mb region with a significant positive LOD score (Figure 7). Moreover, although the authors assumed that these four families did not necessarily have the same predisposing, locus, a heterogeneity LOD score, the highest seen across the genome, suggested that two families (STS200 and STS027) contributed in this region, with the STS200 family showing stronger linkage. As of December 2017, according to the UCSC genome browser (hg19) (78), this region contains 148 genes and 5 miRNAs.

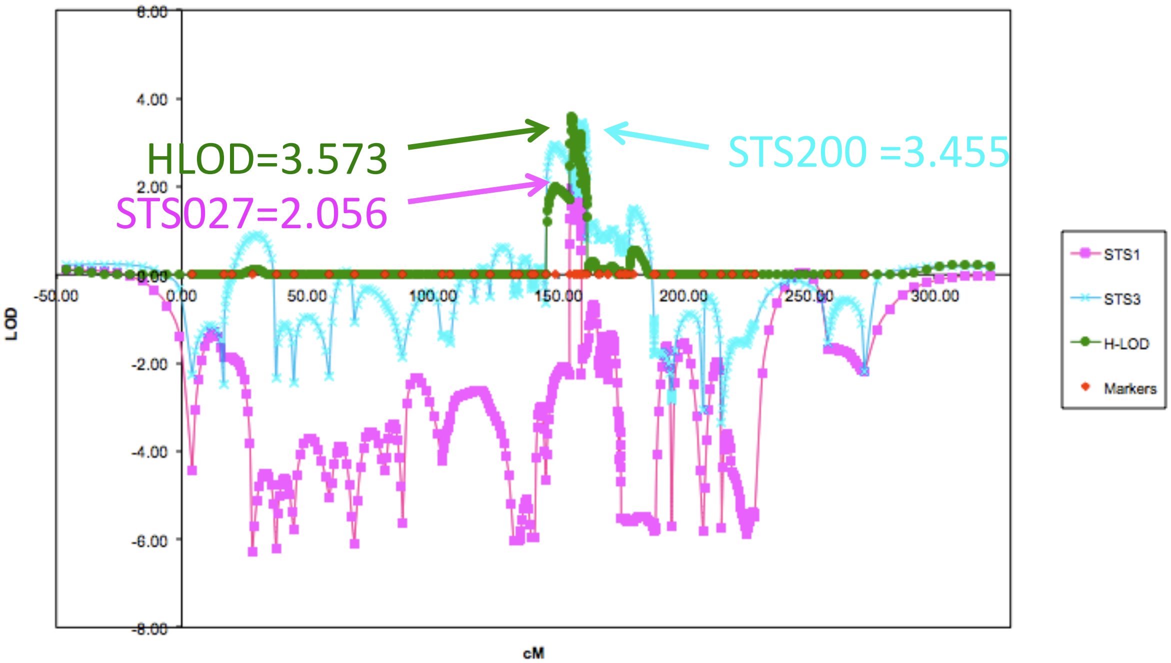


Figure 7: Linkage map showing a high heterogeneity LOD (HLOD) score with contribution from both STS200 and STS027 in the region. Used with permission from Dr. Linda Bachinski.

3 Results

3.1 Quality Control and Choosing a Sequencing Platform

3.1.1 Sanger sequencing of functional positional candidate genes

To best identify putative germline cancer predisposition genes, we first used Sanger sequencing to look for single nucleotide variants (SNVs) and splicing variants across 29 functional positional candidate genes: *AIM2, ATF6, C1orf226, CADM3, CD48, CD244, CREG1, DCAF8, DDR2, DEDD, DUSP12, DUSP23, ESR1, FCRL5, IFI16, KLDHC9, MNDA, NHLH1, NIT1, NUF2, PEA15, PRKAR1A, PYHIN1, SDHC, SH2D1B, TAGLN2, UHMK1, USP21, VANGL2*. Genes were selected relative to known function and potential relevance to cancer. Sanger sequencing was performed across four primary individuals: STS200-000, STS200-017, STS200-032, each of whom had cancer, as well as a fourth, STS200-009, a married-in founder, which served as a negative control (Figure 8).

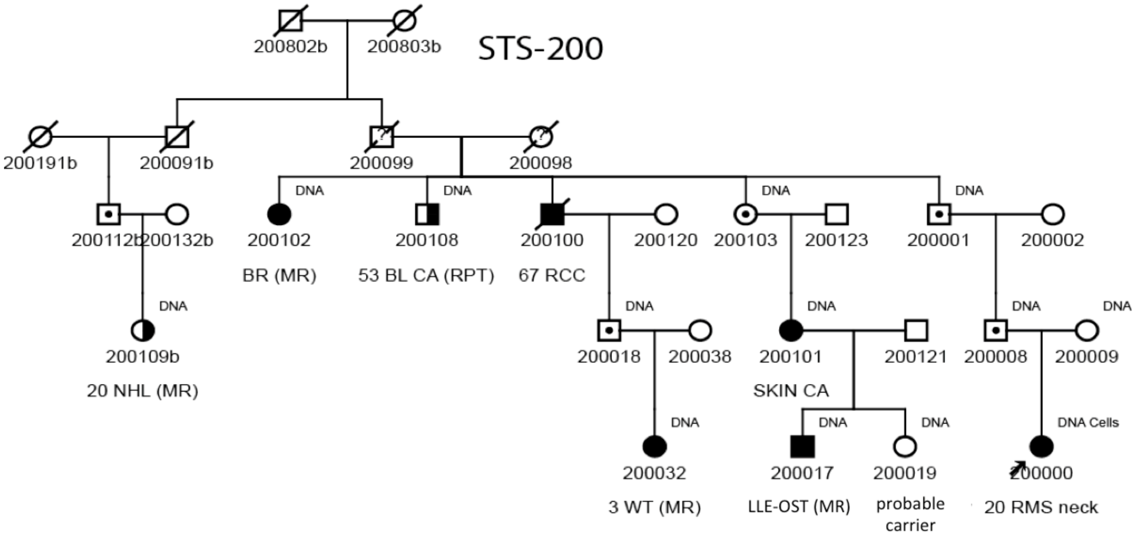


Figure 8: Non-p53 LFS pedigree with linkage in 1q23. Initial Sanger sequencing was performed on three individuals in the latest generation, all with cancer (STS200-032, STS200-017, and the proband, STS200-000), plus a married-in control (STS200-009).

To analyze Sanger data, we used a program called "Mutation Surveyor" (79), which automatically identifies variants using the chromatogram traces. To best ensure that we were not missing anything, we ran the program with both a stringent, normal set of parameters, as well as with a more relaxed set of criteria designed to limit false negatives. Appropriate criteria were discussed with SoftGenetics after we discovered some inconsistencies in their algorithms.

After manual verification using the chromatograms for clean, double peaks, we identified 61 unique heterozygous variants across the three individuals (STS200-000, STS200-017, and STS200-032). However, none of the variants appeared to co-segregate within these three individuals.

When no co-segregating mutations were observed in the Sanger data, we then used 454 sequencing data of these same four individuals to do longer, targeted sequencing of the region. Robust analysis was complicated by an intrinsic drawback of 454-related data – difficulty in interpreting homopolymers – and no strong candidates were identified.

3.1.2 Whole genome sequencing to identify putative mutations

To best identify putative germline cancer predisposition mutations, we chose to employ next generation sequencing (NGS). Given the lack of strong candidates discovered through both Sanger sequencing and 454-sequencing, we chose to leverage whole genome sequencing (WGS) over whole exome sequencing (WES), because of the ability to evaluate non-coding regions in addition to coding regions. Moreover, WGS offered an improved ability to test for copy-number changes.

3.1.3 Establishment of a Sequencing Analysis Pipeline for STS200

Initially, we sequenced two individuals, STS032-011, and STS200-017 across both the Illumina (GAIIX) and Complete Genomics (CGI, v. 2.0) pipelines to determine which sequencing technology we wanted to move forward with. Subsequently, we sequenced one individual on HS2000 (STS200-000), and five additional individuals from STS200 (STS200-001, STS200-008, STS200-009, STS200-019, and STS200-032) via Illumina (H4000).

3.1.3.1 Illumina Data

For Illumina-based data, we implemented an in-house pipeline consistent with best practices according to GATK (80).

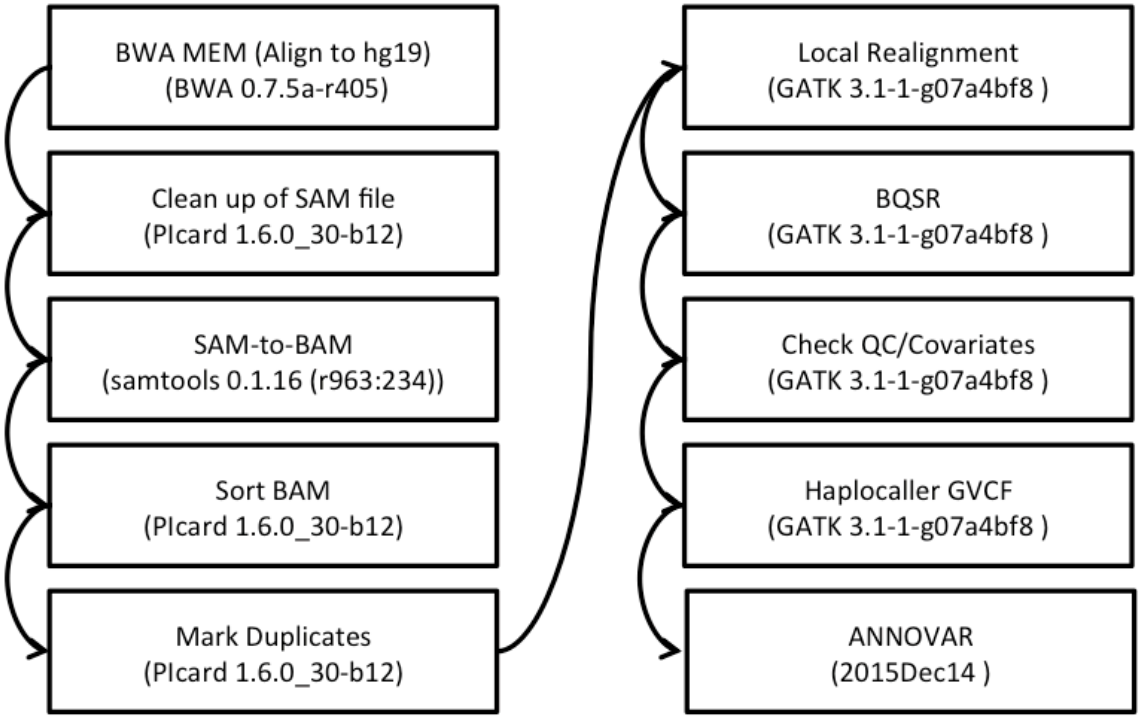


Figure 9: Final pipeline for WGS analysis, modeled after best practices according to GATK.

Briefly, for GAIIX, HS2000 data, we wrote some custom scripts to separate out FASTQ files by lane. Once separated, by machine and lane, fastq files were aligned using BWA-MEM (81). The resulting SAM files were cleaned and marked for duplicates using Picard to mitigate potential biases introduced during amplification. Samples were then realigned to adjust for potential issues near indels, followed by base-recalibration using GATK (80). For the initial comparative analysis between the two pipelines, variant calls were made via UnifiedGenotyper (82), but the most recent and relevant analysis leverages HaploCaller and joint genotyping to generate gvcf files (80).

3.1.3.2 Complete Genomics Data

For CGI data, we used their pipeline. The CGI chemistry was unique in that it used adaptor ligation technologies that introduced known 2-bp gaps into the sequence reads. Because most existing tools were designed around Illumina's more contiguous sequencing and without explicit gaps, adaptation of existing tools to CG data was problematic and would have likely yielded less accurate results. The pipeline for Complete Genomics was more of a black box, and tools to re-run the data ourselves were not provided. Therefore, we used the variant calls generated by their internal algorithms.

3.1.3.3 Illumina vs. Complete Genomics

In order to determine which platform we wanted to move forward with, we compared across a variety of metrics, including overall data quality, and ability to detect known positives, including the *p53*-mutation, and the Sanger sequencing data, where relevant. Preliminary analysis of the Illumina-based STS200-017 revealed that this data was of poorer quality and established the importance of checking QC data.

3.1.3.3.1 *Quality Control, GATK, Picard*

All samples were examined across several quality control metrics to ensure the data could be used for downstream variant calling. One sample, STS200-017 (Illumina) was of demonstrably poorer data quality than the others.

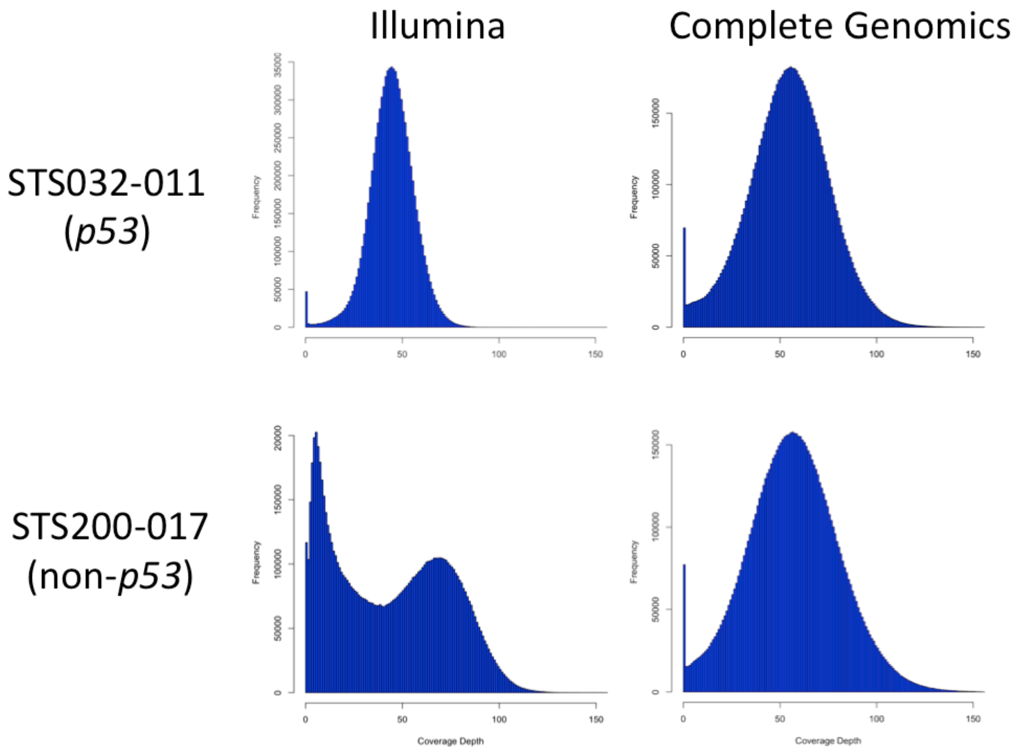


Figure 10: One Illumina sample shows an odd bimodal peak when looking at a plot of the coverage. Single modal peaks without fat tails are the best.

Perhaps most strikingly, although most coverage plots depict a single, approximately normally distributed peak, this “poorer” STS200-017 sample on Illumina produced a bimodal peak (Figure 10).

Moreover, we also saw differences when looking at histograms of the quality scores (of each base of each read). This difference was not apparent when looking only at Illumina’s original quality scores but was when using GATK-based tools to recalibrate quality scores empirically. Virtually all of the empirically determined quality scores for STS200-017 were less than 30, rendering it unsuitable for further use (Figure 11).

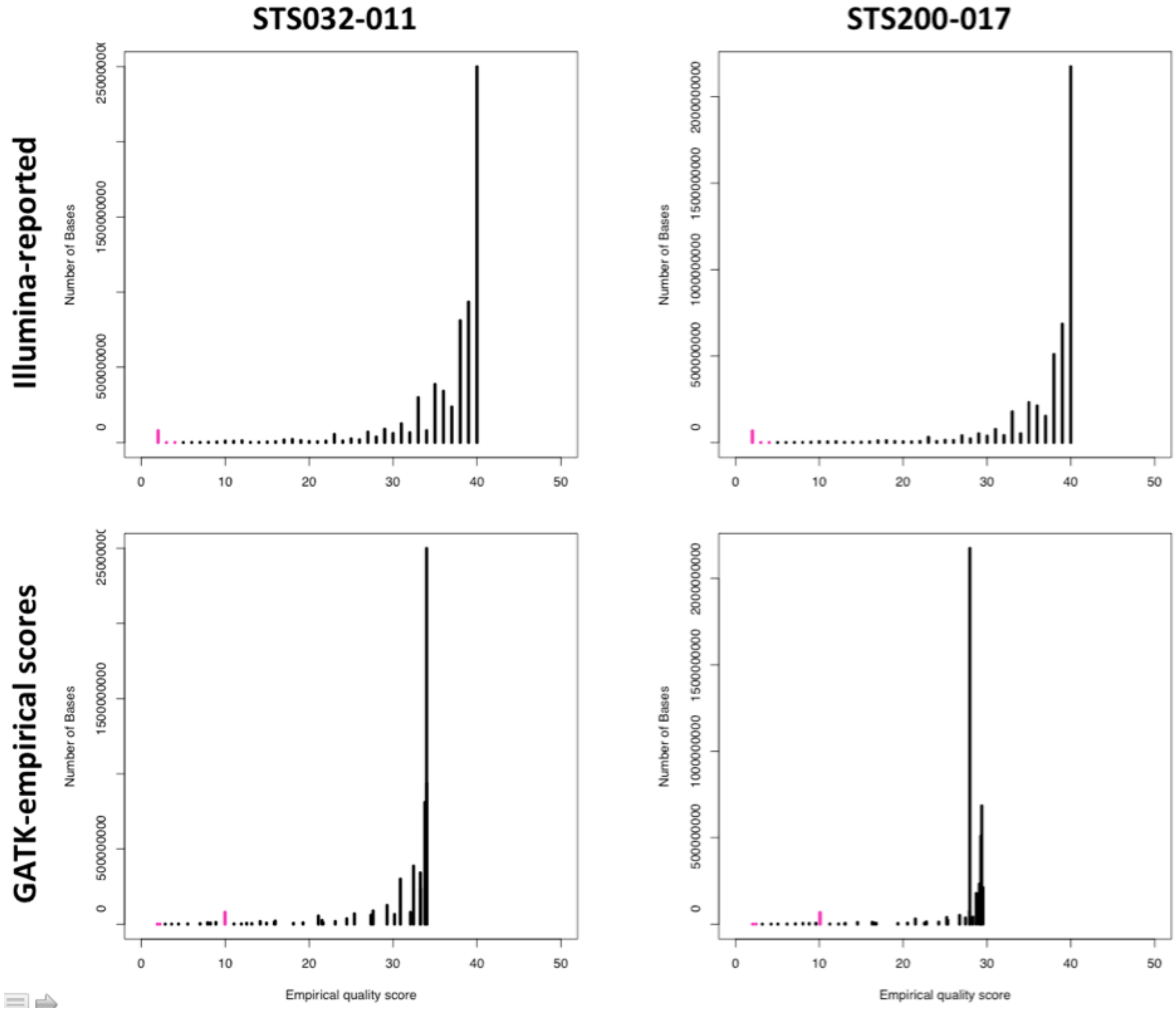


Figure 11: Histogram showing frequency of bases (y-axis), with varying base quality scores (x-axis) for two samples. Data on the top are the reported base quality scores by Illumina for the good sample (left, STS032-011), and bad sample (right, STS200-017). These two graphs are nearly identical. When adjusting empirically for accuracy (bottom), the poorer sample fares significantly worse, with the majority of data falling below a q-score of 30.

We also reviewed additional metrics such as a histogram of base by cycle, and base quality by cycle (Figure 12). Histograms of these data comparing STS032-011 to STS200-

017, suggested that STS200-017 had steep drop-offs in quality near the ends of the reads at higher cycle counts. These data may have contributed to the percentage of unmapped, or singly mapped reads found in STS200-017, which was 5X higher than for STS032-011.

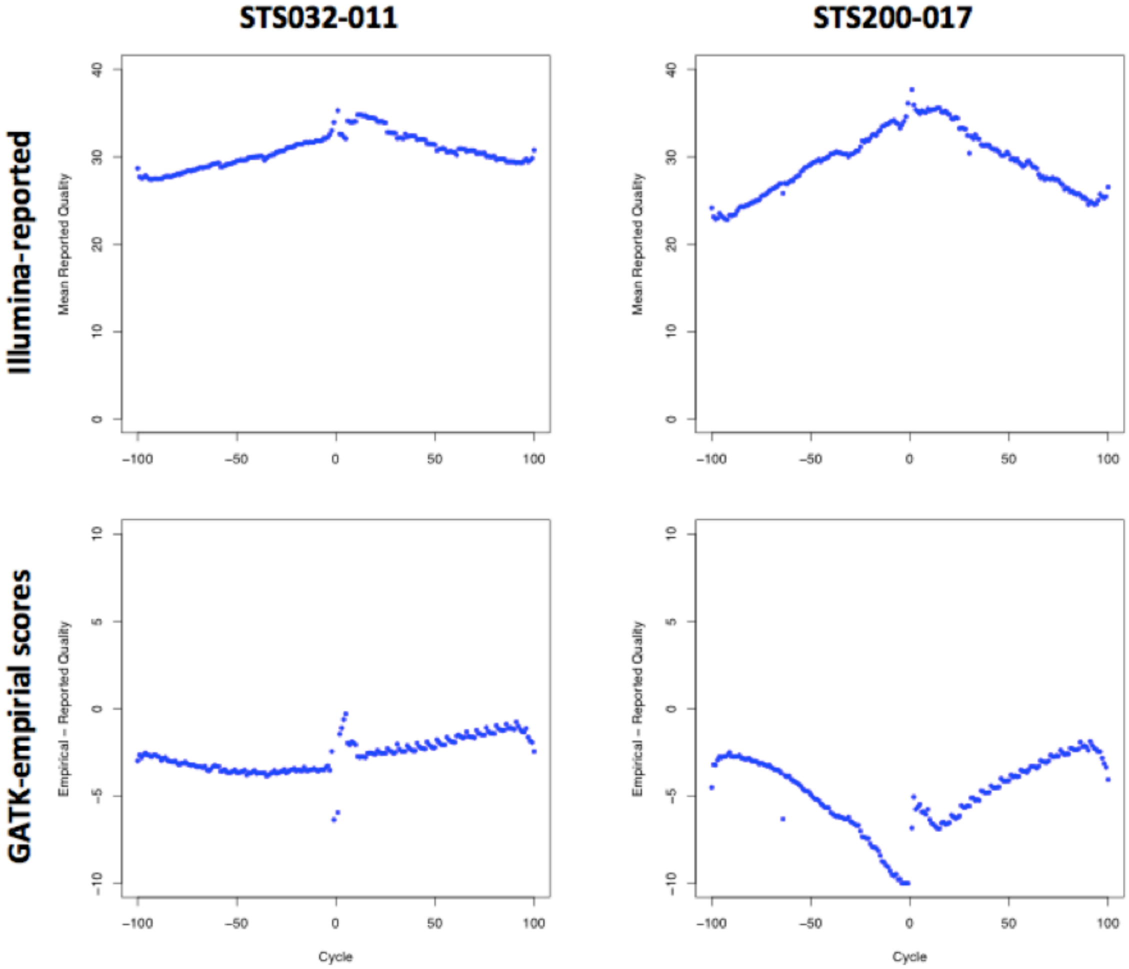


Figure 12: Quality control of base quality by cycle after recalibration shows dramatic differences between the good sample (STS032-011) and the poorer sample (STS200-017). In STS200-017 the second of the paired reads performs especially poorly, with empirical scores being up to 10 worse than reported score.

We then checked to see if the bottom-line was affected and if these apparent differences in quality impacted variant calls. Comparison of variant calling between a basic variant calling algorithm setup for Illumina (via GATK, UnifiedGenotyper (82)), and Complete Genomics (Complete Genomics internal pipeline), saw dramatic differences when comparing SNVs for STS200-017 (51%) vs. STS032-011 (85%).

Taken together, these data suggested that the STS200-017 sample for Illumina performed more poorly. Upon presentation to Illumina, they agreed to re-sequence a second sample from the same individual (now on HS2000), and this second sample passed all QC metrics and was in-line with the metrics generated for STS032-011 and other samples sequenced at the same time with Illumina. These data strongly support the value of QC in sequencing studies.

3.1.3.4 Comparison of WGS Technologies Relative to Existing Sanger Data

We next leveraged existing Sanger data for STS200-017 that arose out of the functional positional candidate screen. Forty-eight variants in the Sanger data for STS200-017 were identified using Mutation Surveyor, and were hand-validated individually by checking the chromatograms for double-peaks in both the forward and the reverse strand. Neither Illumina (44 SNVs)) nor Complete Genomics (45 SNVs) identified all 48 SNVs; collectively they were able to identify 46 total SNVs. All "missed" SNVs were homozygous for the reference by WGS. The data suggest that both Illumina and Complete Genomics are approximately equivalent; they both identify the majority of Sanger-ascertained SNVs. However, assuming Sanger sequencing as the Gold standard, the data did suggest that the WGS might have contain some false negatives. We additionally checked the other sample, STS032-011, for which we had more limited Sanger data available, because it had a known *p53* mutation, for the expected variant, and confirmed we were able to identify the *p53 R175H* mutation.

3.1.4 Choosing Illumina as a Sequencing Platform

Several factors lead us to choose Illumina over Complete Genomics. First, the majority of the Complete Genomics analysis pipeline is not well understood, and exists in a black box. It requires the use of proprietary algorithms and analyses cannot be cross-checked with existing tools like BWA (81), Samtools (83), Picard (84), and GATK (80, 85). However, at the time, we did not feel as if the variant calling for Complete Genomics was sub-par or compromised due to these differences. Moreover, although Illumina did have a slight hiccup during the initial phase of testing, subsequent resubmission of a different sample from the same individual cleaned up the data and QC metrics well.

3.1.5 Additional Sequencing of LFSL Family Members.

Once we had selected Illumina to move forward with, we sequenced the proband STS200-000 (HS2000). However, these data produced no compelling co-segregating variants between STS200-000 and STS200-017. To better ascertain co-segregating variants, we then sequenced five additional individuals (STS200-032, STS200-019, STS200-008, STS200-009, and STS200-001) at MD Anderson Cancer Center on HS4000 at 100-bp paired end and analyzed according to the previously outlined pipeline (Figure 9)

3.1.6 Establishment of a Variant Prioritization Strategy

Given the STS200 pedigree (Figure 8), we expected that a germline cancer predisposition variant could be driving the cancer and sarcoma phenotypes seen in the latest generation. However, clinical testing suggested that a *p53* mutation was not responsible, suggesting the presence of a different, LFSL gene. We hypothesized that the driver(s) behind LFS and LFSL would be etiologically similar.

Therefore, we established several criteria

(1) We expected the variant to co-segregate between affected individuals with cancer and obligate carriers because it resembles an inherited cancer syndrome

(2) The variant would be in 1q23, particularly given the relatively high LOD score

(3) The mutation would be an SNV (70%-80% of *p53* LFS is driven by missense mutations

(4) The mutation would be predicted to be damaging, either through lack of conservation, or through a big change in the amino acid properties

(5) Lastly, we expected the variant to be rare because LFS is rare.

In summary, we initially focused on identifying co-segregating, rare (MAF<1%), heterozygous, coding SNVs that were predicted to be damaging by both SIFT (<0.05), (86) and Polyphen2 (PP2, >0.453), (87) that were in the linked region in 1q23.

Unfortunately, there were no variants that met these stringent criteria. To better isolate putative mutations, we then relaxed the criteria in several ways (Table 1).

Table 1: Table of various criteria used to identify putative cancer predisposition variants in 1q23. Under the stringent criteria, no putative variants were found. We incrementally adjusted criteria to be more relaxed to better identify candidate mutations.

	Stringent	**More Relaxed**
Minor Allele Frequency (MAF, 1KG)	<1%	<5%
SIFT/PP2 predictions	Both damaging	Either damaging
Co-segregation	Absolute	FamSeq (88)
Variant type	SNV only	SNV or indels
Location (via linkage analysis)	In 1q23	Whole genome

3.2 Identification of ARHGAP30 as an LFSL Predisposition Gene

3.2.1 Penetrance and Rarity in *p53* and LFS

The presence of multiple LFS-related criteria suggests uncertainty in the field about the best and most appropriate definitions for LFS. (60, 61, 63, 64) Inevitably, less stringent definitions of LFS, such as under the Chompret criteria, where just a single individual can be sufficient for an LFS diagnosis, imply that penetrance need not be significantly high in LFS/LFSL families. Moreover, we posit that given the overall importance and prominence of *p53* in cancer, additional LFSL genes are likely to be less penetrant. In turn, this can lead to greater ambiguity in clinical ascertainment, and underreporting, such that in sum, less penetrant genes and mutations could be somewhat more prevalent in the general population than would otherwise be expected based on LFSL (non-p53 LFS) incidence. If we do consider that such genes may be less penetrant, then it may make sense to adjust the cutoff point for a MAF-filter.

3.2.1.1 Framework for Rare Variation

According to several databases, the most common LFS mutations are quite rare in the general population. We leveraged the IARC (89) database to determine the most frequently mutated codons, and then looked up the mutation frequency of variants in these codons in databases such as Exome Aggregation Consortium (ExAC) (90) and ESP (91). In Figure 13 we show a histogram of the most commonly mutated codons in families that have been clinically ascertained as classic LFS (strictest criteria). For the nine most common codons, I have included the codon number, and codon sequence. Above these, where present, are the "MAF" of a variant in that codon as indicated in the ExAC database, as well as the incidence. At the time of inquiry, ExAC contained 60,706 individuals. Despite the fact that ExAC attempts to remove individuals affected by severe pediatric disease, it is clear that *p53* variants are occurring at some relative frequency. Under the exome sequencing project, which covered 7,000 individuals, we saw exactly 1 individual with a mutation in codon 273 (0.02%), while we saw 28 with a mutation in codon 337 (0.56%).

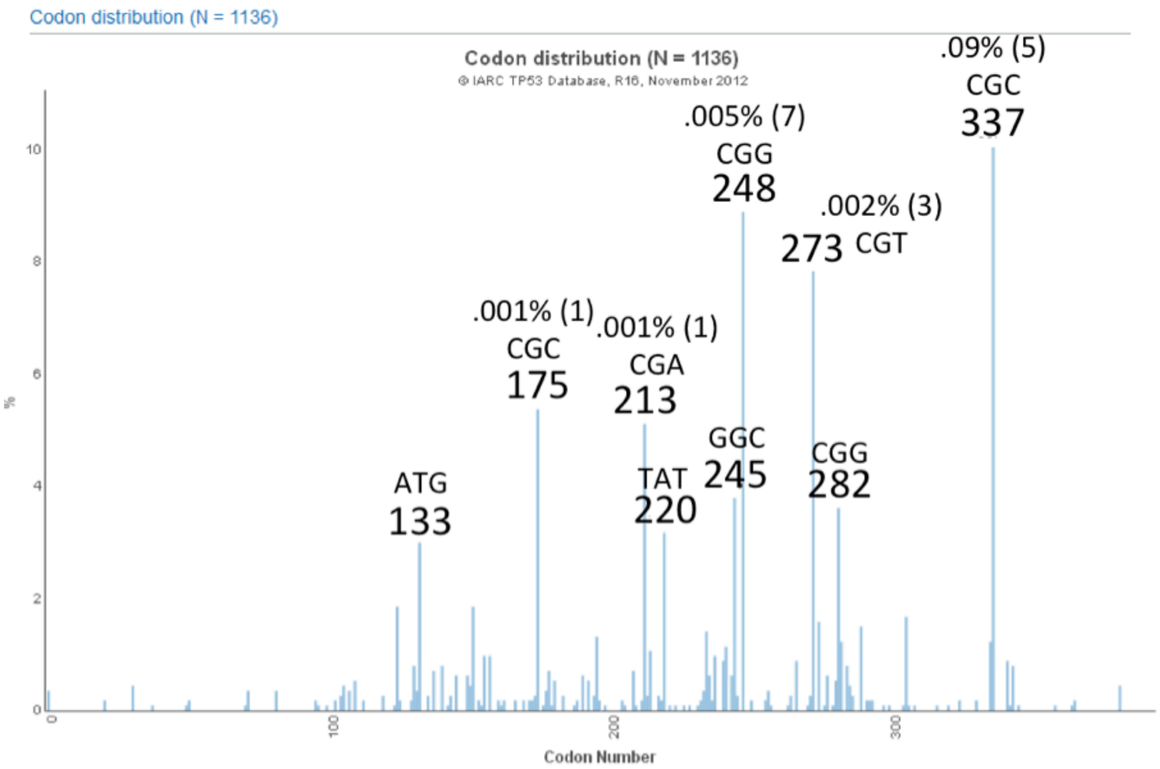


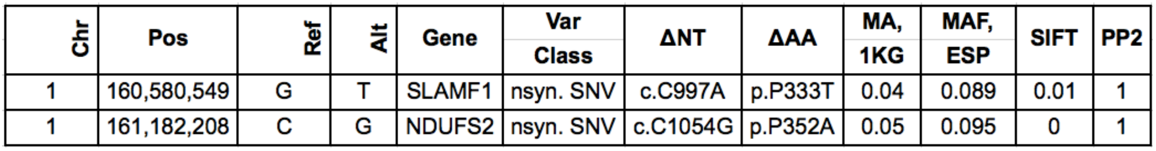
Figure 13: Histogram from IARC (89) showing hotspot mutations in *p53* at the codon level. For the nine most commonly altered codons, the codon number is listed, along with the codon sequence. Above that is a representation of the incidence in the ExAC database, either as a percentage of all samples, or as the total number of individuals with a mutation in that codon (in parentheses), indicating that these mutations are very rare in the general population.

In contrast to this, recent research supports the idea that germline *TP53* mutations may be more common than previously appreciated based on LFS incidence. A perusal of exome databases found 131 individuals (0.2% of samples) with *TP53* mutations. These mutations largely fell in the DNA binding domain (~80%), and included some that are known to cause LFS, suggesting that numerous individuals may be carrying a non-penetrant deleterious *p53* allele (92). Overall penetrance through close examination of the IARC database was found to be ~80% (tumors before the age of 70), with varying penetrance based on sex, age, and which *p53* allele was mutated (93).

3.2.2 Allowing for more common variants identifies two variants that co-segregate

Given these emerging data, particularly that *p53* variants may occur as much as 0.2% in the general population, may mean that a MAF threshold of 1% is too conservative (92). In order to cast a wider net, and account for the possibility that variable penetrance could partially explain the results, we opened our MAF filter to account for variants with a MAF less than 5%. Although this could potentially lead to the identification of a multitude of variants, in practice maintaining the absolute co-segregation criterion is highly restrictive. There are exactly two variants that both co-segregate with a MAF threshold of less than 5%, and are predicted to be damaging by both SIFT and Polyphen2 (Table 2).

Table 2: Table showing all variants in 1q23 which both co-segregate, have a MAF < 5%, and are predicted to be damaging by both SIFT and Polyphen2 (PP2).



Chr	Pos	Ref	Alt	Gene	Var Class	ΔNT	ΔAA	MA, 1KG	MAF, ESP	SIFT	PP2
1	160,580,549	G	T	SLAMF1	nsyn. SNV	c.C997A	p.P333T	0.04	0.089	0.01	1
1	161,182,208	C	G	NDUFS2	nsyn. SNV	c.C1054G	p.P352A	0.05	0.095	0	1

SLAMF1 encodes signaling lymphocytic activation molecule 1. It is involved in the activation and differentiation of immune cells and has roles in both innate and adaptive immune response. As noted in Table 2, although the SIFT and PP2 scores are both promising (i.e. predicted to be deleterious – scores are close to 0 and 1 respectively), the MAF is quite high (close to 4% in the 1000 genomes project (94), 1KG)). When considered under the ESP

(adjusted for ethnicity - the STS200 family is of Caucasian descent), this number is about twice as high, 8.9%. Such a high MAF is perhaps more common than we would prefer for what is theoretically a rare disease. SLAMF1 has not yet been associated with a disease. Nevertheless, the role of the immune system in cancer is becoming increasingly apparent; it is a new hallmark in the “Hallmarks of Cancer” v2. (2), thus making it a plausible functional candidate.

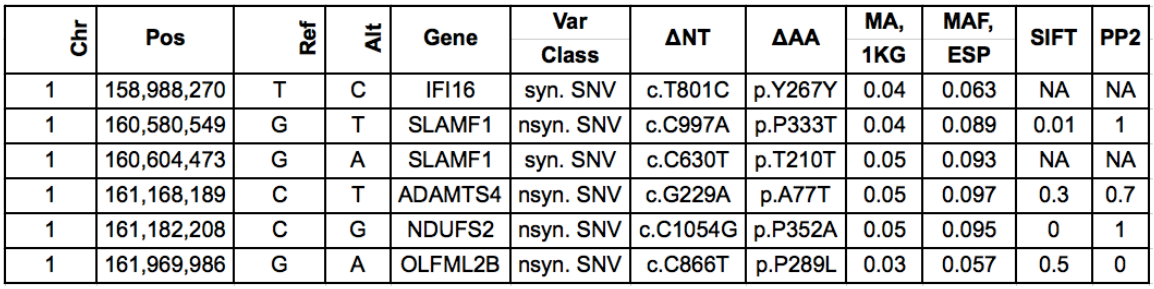
NDUFS2 encodes NADH dehydrogenase ubiquinone iron-sulfur protein 2. NDUFS2 catalyzes NADH oxidation within the mitochondria (including ubiquinone reduction and proton ejection). Much like *SLAMF1*, the MAF in both the 1KG and ESP (for European Americans) are quite high for what is understood to be a relatively rare disease. Mutations in the gene *NDUFS2* have been associated with Mitochondrial Complex I Deficiency. This mitochondrial disorder has heterogeneous presentation, including macrocephaly (large head) and myopathies. However, cancer has never been associated with this disease, suggesting that *NDUFS2* may not be a strong candidate for additional follow-up.

To best determine if these variants were viable candidates, we performed Sanger sequencing to (a) confirm the variant in the 7 individuals with WGS, and (b) to test for co-segregation in additional members of the pedigree. In *SLAMF1*, the married-in founder turned out to have the mutation of interest. In *NDUFS2*, two obligate carriers and an affected were WT for the variant. Thus, neither of these variants is of significant interest moving forward.

3.2.3 Less stringency in SIFT/PP2 finds four new variants for consideration

A second possibility is that SIFT (86) and PP2 (87) are imperfect prediction algorithms and that requiring a mutation to be “damaging” across both algorithms may lead to premature exclusion of variants. SIFT predictions are based on amino acid conservation in closely related sequences. In contrast, PP2 emphasizes the impact of the change on protein structures. In practice, SIFT and PP2 do not always agree with one another, and it may be more worthwhile to consider variants predicted to be damaging by one or more algorithm(s). Moreover, known LFS variants such as *p53 M133T* sometimes may be predicted to be relatively benign (initially the variant had a PP2 score of 0.148, but under the most recent iteration, it is now 0.858). For example, if a cancer predisposing germline mutation is human-specific; in this case, we would expect a SIFT score, which is based primarily on conservation to be relatively high (i.e. benign). To cast the widest possible net, we then changed the stringency in SIFT/PP2 to not require any thresholds, although we did still require the variant to be exonic. In combination with the higher MAF threshold, we found 6 total variants (Table 3), two of which were the same as in 3.2.2.

Table 3: Table of all variants meeting meeting the stringent criteria, except for rarity, and SIFT/PP2 scores. All 6 variants are relatively common, and two are synonymous SNVs.



Chr	Pos	Ref	Alt	Gene	Var	ΔNT	ΔAA	MA,	MAF,	SIFT	PP2
					Class			1KG	ESP		
1	158,988,270	T	C	IFI16	syn. SNV	c.T801C	p.Y267Y	0.04	0.063	NA	NA
1	160,580,549	G	T	SLAMF1	nsyn. SNV	c.C997A	p.P333T	0.04	0.089	0.01	1
1	160,604,473	G	A	SLAMF1	syn. SNV	c.C630T	p.T210T	0.05	0.093	NA	NA
1	161,168,189	C	T	ADAMTS4	nsyn. SNV	c.G229A	p.A77T	0.05	0.097	0.3	0.7
1	161,182,208	C	G	NDUFS2	nsyn. SNV	c.C1054G	p.P352A	0.05	0.095	0	1
1	161,969,986	G	A	OLFML2B	nsyn. SNV	c.C866T	p.P289L	0.03	0.057	0.5	0

Of these, there were three new genes, *IFI16, ADAMTS4* and *OLFML2B*. ADAMTS4 encodes disintegrin and metalloproteinase with thrombospondin motifs 4 and is known to degrade aggrecan in cartilage and brevican in the brain and has a thrombospondin type motif (TSR) that binds to the ECM. (95) *ADAMTS4* is thought to possibly play roles in arthritis (96) and potentially in glioma progression. (95) Similar to previous candidates, MAF for these genes are comparatively high, around 5% and 10% for 1KG and ESP respectively, and make the variant less compelling.

The variant in Interferon Gamma Inducible Protein 16 (*IFI16*) is synonymous, and like the other variants considered so far, is reasonably polymorphic, but is actually one of the candidate genes previously identified. IFI16 is a member of the p200 family, known to inhibit cell cycle progression. Moreover, loss of *IFI16* has been associated with breast and prostate cancers (97).

There have been no papers published on Olfactomedlin-like 2B (*OLFML2B)*, but olfactomedlins have been implicated in development and organization of the nervous system, and are generally thought to facilitate protein-protein interactions and cell adhesion (98). Given the lack of immediately compelling candidates, we pushed forward to consider additional variants rather than immediately jump to consider Sanger sequencing.

3.2.4 FamSeq identifies *ARHGAP30* as an LFSL candidate gene

Without any particularly strong candidates so far identified in 1q23, we next implemented FamSeq (88) to help identify additional co-segregating variants and reduce the risk of false negatives. FamSeq leverages pedigree information to make more informed decisions about variant zygosity. Effectively, FamSeq places a probability on a given variant call, including how likely it is to be heterozygous. If one or more individuals have the same weak(er) variant call, FamSeq helps to isolate that variant and flag it for additional consideration.

We then used FamSeq on the six affecteds/obligate carriers, requiring complete co-segregation of the same heterozygous mutation of interest. At this time, we did not use the married-in control as a negative filter since this could be crosschecked later, and it appeared the filtering criteria were already particularly stringent. However, under these conditions, we found no new rare, or semi-rare coding mutations in 1q23, even when considering the more lenient parameters discussed in 3.2.2 and 3.2.3.

One other possibility is that the family contains a(nother) phenocopy, or that one affected individual could be homozygous, or even hemizygous for the mutation of interest. It would be unlikely, but not impossible, for example, for both parents to have the mutation, or for there to have been a gene conversion event prior to DNA sampling. Instead then of asking FamSeq to identify all variants that had possible heterozygous mutations in 6 of 6 affected/obligate carriers, we instead asked FamSeq to identify all variants that seemed to be heterozygous in 5 of 6 individuals. Under the more relaxed criteria listed above, we generated another list of candidate variants, with one notable addition, a mutation in *ARHGAP30* (Table 4).

Table 4: Table of FamSeq-derived variants that are semi-rare, and meet strict SIFT/PP2 requirements.

Chr	Pos	Ref	Alt	Gene	Var Class	ΔNT	ΔAA	MA, 1KG	MAF, ESP	SIFT	PP2
1	160,580,549	G	T	SLAMF1	nsyn. SNV	c.C997A	p.P333T	0.04	0.089	0.01	1
1	161017761	C	T	ARHGAP30	nsyn SNV	c.G2417A	p.R806Q	0.01	0.02721	0	1
1	161,182,208	C	G	NDUFS2	nsyn. SNV	c.C1054G	p.P352A	0.05	0.095	0	1

The ARHGAP30 variant was found to be heterozygous in 5 of 6 affecteds/obligate carriers, but in the sixth sample, the variant was found to be homozygous mutant (Figure 14).

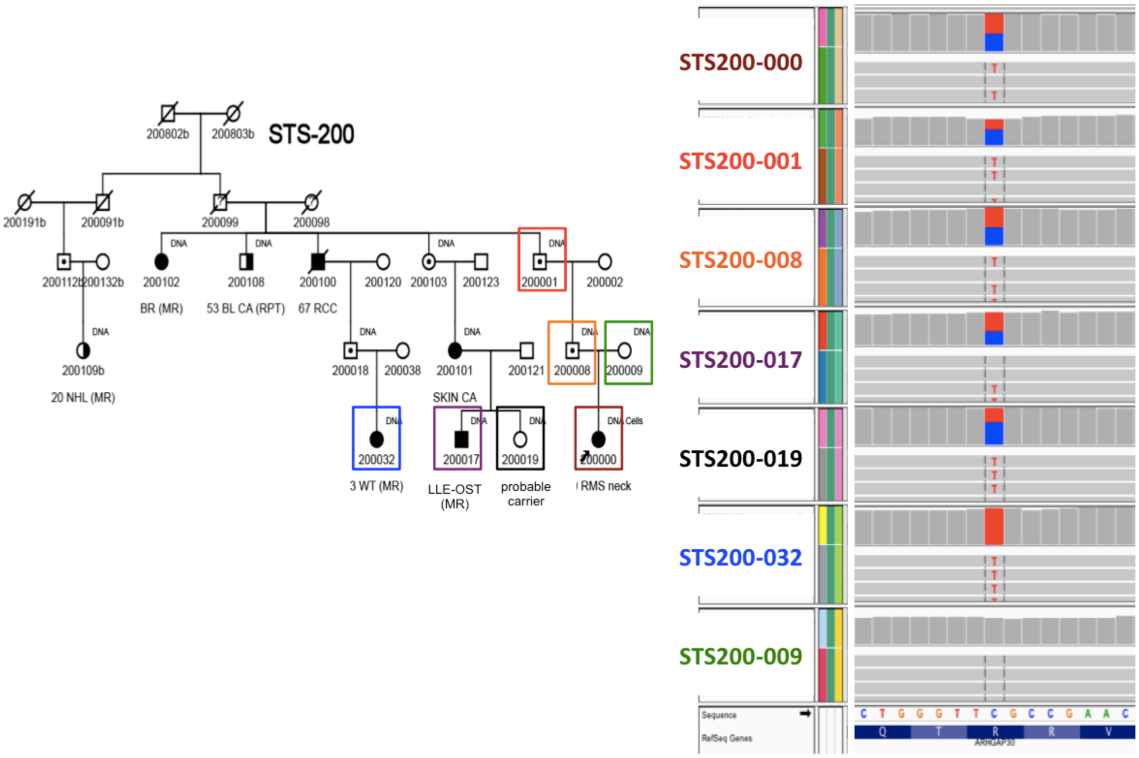


Figure 14: Depiction of STS200 and WGS data, showing co-segregation of the *ARHGAP30* mutation. One individual is homozygous (STS032-011)

ARHGAP30 encodes a RhoGTPase that has been linked to cancer and cell migration (99, 100), with research indicating it acts like a tumor suppressor gene, making it a good candidate for LFSL. Although the MAF was somewhat higher than initially desired, the fact that it was a tumor suppressor gene, and was predicted to be damaging (coding), and co-segregates implicates it as the most promising candidate identified. We then confirmed the results, including the homozygous result by Sanger sequencing for these seven individuals, indicating this mutation did co-segregate for these seven individuals. Subsequent follow-up to sequence additional members of the pedigree confirmed that co-segregation was otherwise complete. There was one seeming outlier: a bladder cancer patient (STS200-102), who was WT for the mutation. However, based on the linkage, he shared very little of the 1q23 haplotype and was expected to be a phenocopy, (i.e. WT for the mutation of interest). Thus, these data further cemented *ARHGAP30* as a candidate for future studies.

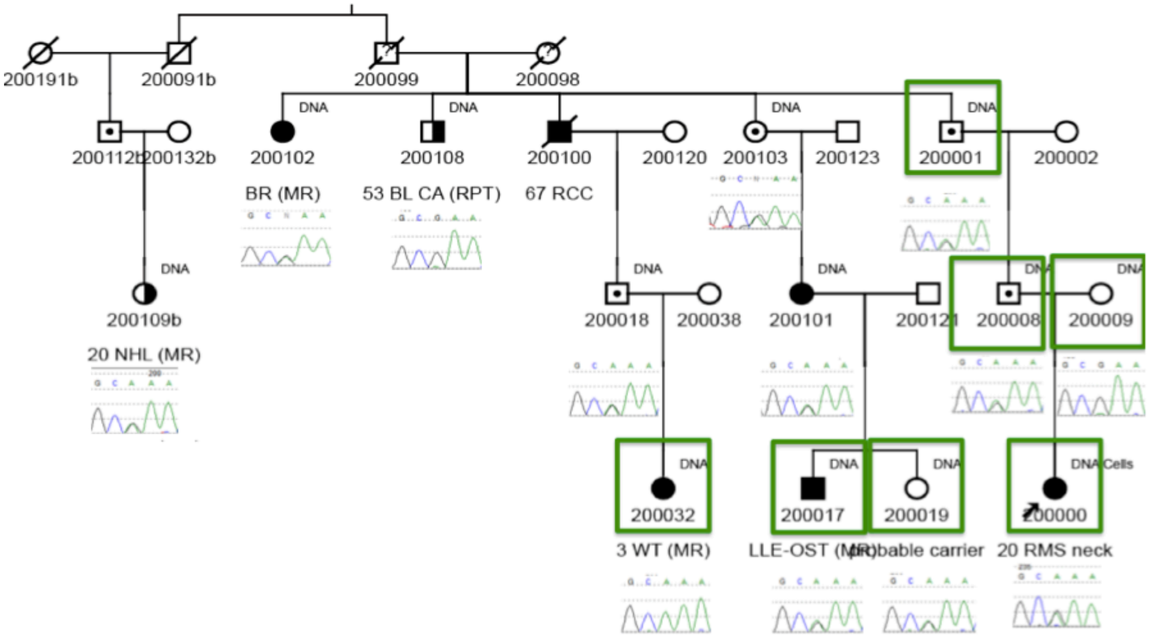


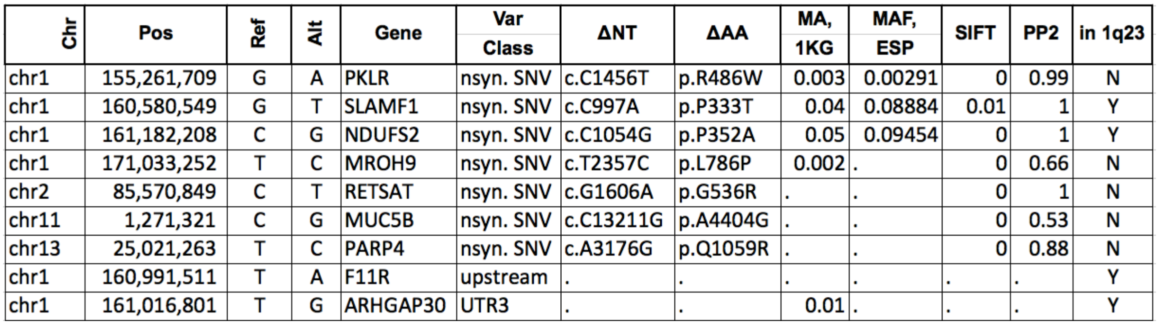
Figure 15: Sanger sequencing of additional members of the pedigree confirms co-segregation of the mutation in *ARHGAP30*, including the homozygous result for STS200-032. Based on the linkage, we expected one of STS200-108 or STS200-109 to be a phenocopy, of which STS200-108 appears to be. Individuals with WGS are marked by green boxes.

3.2.5 Co-segregation over the whole genome

In order to rule out the unlikely possibility that the linkage was incorrect, we expanded the search to include variants over the entire genome. This time, because we were not in the linked region, we changed the parameters to exclude STS200-019, because she was a probable carrier based on linkage, but has not yet developed cancer. We continued to leverage FamSeq to reduce the risk of false-positives, but decided to be more restrictive using FamSeq by not allowing any phenocopies or homozygous alternate alleles, and requiring five of five individuals to carry the putative mutation. However, we continued to use the less stringent criteria (MAF < 5%, SIFT or PP2 predicted to be damaging).

We found seven coding variants that met these criteria over the whole genome, including the aforementioned *SLAMF1* and *NDUFS2*, as well as two non-coding variants in *F11R* and *ARHGAP30* that were within 1q23.

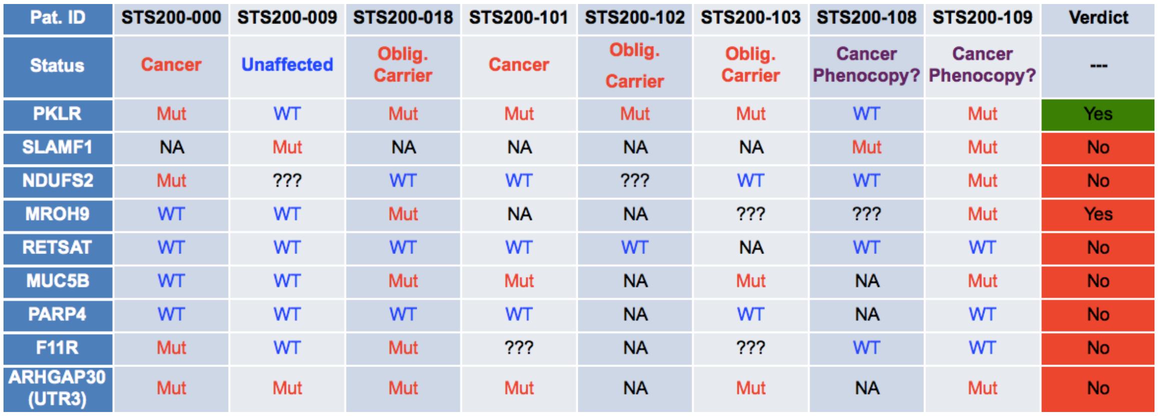
Table 5: Table of variants that appear to completely co-segregate using FamSeq over the entire genome.



Chr	Pos	Ref	Alt	Gene	Var	ΔNT	ΔAA	MA,	MAF,	SIFT	PP2	in 1q23
					Class			1KG	ESP			
chr1	155,261,709	G	A	PKLR	nsyn. SNV	c.C1456T	p.R486W	0.003	0.00291	0	0.99	N
chr1	160,580,549	G	T	SLAMF1	nsyn. SNV	c.C997A	p.P333T	0.04	0.08884	0.01	1	Y
chr1	161,182,208	C	G	NDUFS2	nsyn. SNV	c.C1054G	p.P352A	0.05	0.09454	0	1	Y
chr1	171,033,252	T	C	MROH9	nsyn. SNV	c.T2357C	p.L786P	0.002	.	0	0.66	N
chr2	85,570,849	C	T	RETSAT	nsyn. SNV	c.G1606A	p.G536R	.	.	0	1	N
chr11	1,271,321	C	G	MUC5B	nsyn. SNV	c.C13211G	p.A4404G	.	.	0	0.53	N
chr13	25,021,263	T	C	PARP4	nsyn. SNV	c.A3176G	p.Q1059R	.	.	0	0.88	N
chr1	160,991,511	T	A	F11R	upstream	.	.	.	.	.	.	Y
chr1	161,016,801	T	G	ARHGAP30	UTR3	.	.	0.01	.	.	.	Y

To definitively determine if any of these variants were also candidate germline cancer predisposition variants, we tested additional samples from the pedigree to see if they maintained co-segregation.

Table 6: Table showing that additional variants elsewhere in the genome identified by FamSeq largely do not co-segregate when tested by Sanger sequencing.



Pat. ID	STS200-000	STS200-009	STS200-018	STS200-101	STS200-102	STS200-103	STS200-108	STS200-109	Verdict
Status	Cancer	Unaffected	Oblig. Carrier	Cancer	Oblig. Carrier	Oblig. Carrier	Cancer Phenocopy?	Cancer Phenocopy?	---
PKLR	Mut	WT	Mut	Mut	Mut	Mut	WT	Mut	Yes
SLAMF1	NA	Mut	NA	NA	NA	NA	Mut	Mut	No
NDUFS2	Mut	???	WT	WT	???	WT	WT	Mut	No
MROH9	WT	WT	Mut	NA	NA	???	???	Mut	Yes
RETSAT	WT	WT	WT	WT	WT	NA	WT	WT	No
MUC5B	WT	WT	Mut	Mut	NA	Mut	NA	Mut	No
PARP4	WT	WT	WT	WT	NA	WT	NA	WT	No
F11R	Mut	WT	Mut	???	NA	???	WT	WT	No
ARHGAP30 (UTR3)	Mut	Mut	Mut	Mut	NA	Mut	NA	Mut	No

Of these, only the variant in *PKLR* was found to completely co-segregate across the rest of the STS200 pedigree. However, this exact variant is annotated in the ClinVar database (101), as being associated with pyruvate kinase deficiency (PKD). PKD is an autosomal recessive disease in which red blood cells break down too easily due to the lack of pyruvate kinase. It has never been associated with a cancer phenotype.

3.2.6 No non-coding variants are immediately compelling

Next, we considered variants that were in 1q23, but were not in coding regions, such as those in UTRs, introns, or that were upstream or downstream of the gene (with 2,500-bp), in potential regulatory regions. There were 203 such variants that were not considered to be polymorphic, the majority of which were intronic. However, we note that some of these are likely to be false positives because they existed in near regions that were homopolymers.

We then applied FunSeq2 (102), to better narrow down the list. FunSeq2 is a tool that can help prioritize non-coding regions by looking at how often variants appear in them.

The authors hypothesize that regions that almost never have a variant are somehow protected and therefore may be functionally important. However, we didn't find any variants in these so-called ultra-conservative regions.

3.3 Cross-check of *p53 confirms that p53 is not a genetic driver in STS200*

We also exhaustively reviewed *p53* to confirm it had no mutations, or copy-number loss and was definitively WT. Moreover, as discussed in 2.3.2, recent research has identified a *p53* UTR-variant that co-segregated with disease in an LFSL family where *p53* had been previously ruled out (77). To check to see if the STS200 family contained this mutation (rs78378222), we checked the WGS data, and used Sanger sequencing, finding that all tested individuals were WT for the UTR-variant.

3.4 More in-depth investigation of "known" sarcoma risk factors

Several genes have been associated with sarcomas, either through a study of individual subtypes, or through broader analysis of targeted sequencing. When they analyzed 1,162 sarcoma probands, Ballinger et al. found that, overall, probands were more likely to have multiple pathogenic variants relative to controls, and that these were best clustered in *TP53, ATM, ATR, BRCA2,* and *ERCC2* (16). To supplement these data, we generated a genome-wide list of genes that have been previously associated with sarcomas through both a literature search, and a perusal of the data put together by the WHO, across both CN data and mutational data (11).

Table 7: List of genes associated with sarcomas, either through hereditary cancer syndromes, CN changes, or as fusion transcripts.

ACP5	COL1A1	GLMN	NFATc2	SMARCA4
ACTB	COL1A2	GNAS	NFIB	SMARCA5
ACVR1	COL6A3	H19	NR4A3	SMARCB1
AIP	CREB1	HAS2	NTRK3	SOS1
AKT1	CREB3L1	HEY1	PATZ1	SP3
ALK	CREBBP	HMGA2	PBX1	SQSTM1
ANTXR1	CSF1	HRAS	PDGFB	SS18
ANTXR2	CTNNB1	IDH1	PDGFRA	SSX1
APC	CXCR7	IDH2	PIK3CA	SSX2
ASPSCR1	DDIT3	IGF1R	PLAG1	SSX4
ATF1	DICER1	IGF2	PMS2	STARD13
ATIC	DUX4	INI1	PORCN	T
BCL2	EBF1	KCNQ1OT1	POU5F1	TAF15
BCOR	ERCC2	KDR	PRKAR1A	TEK
BLM	ERG	KIT	PTCH1	TFE3
BRAF	ETV1	KRAS	PTEN	TIE1
BUB1B	ETV4	LEMD3	PTEN	TNFRSF11A
C11orf95	ETV6	LHFP	PTH1R	TNFRSF1A
CAMTA1	EWSR1	MAP2K1	PTPN11	TP53
CCNB3	EXT1	MDM2	RANBP2	TPM3
CCND1	EXT2	MEN1	RB1	TPM4
CCNE1	FEV	MID1	RECQL3	TSC1
CDH11	FGFR4	MKL2	RECQL4	TSC2
CDK4	FH	MTAP	RET	USP6
CDKN1C	FLI1	MYH9	ROR2	VGLL3
CDKN2A	FLT1	MYOCD	SDD	VHL
CDKN2B	FOSL1	NBN	SDHB	WRN
CHEK2	FOXO1	NCOA2	SDHC	WT1
CIC	FUS	NF1	SDHD	WWTR1
CLTC	GLI1	NF2	SH3BP2	

However, we did not find any other pathogenic variants in the germline in these sarcoma-related genes in any of the STS200 family members. It is possible that some of this is driven by the fact that many of the listed genes are part of fusion transcripts.

3.5 STS200 has no co-segregating indels or deletions anywhere in the genome

Lastly, LFS can be caused not only by SNVs in *p53*, but also by indels and CN changes (56). To best determine if LFSL in the STS200 family could also be caused by either of these, we leveraged both GATK and pindel (103) for calling indels, and ERDS (104) for CNVs. However, we found no recurrent indels or CN changes suggesting that neither indels of CN changes were responsible for LFSL in STS200.

3.6 Sequencing of additional LFSL pedigrees finds three other cases with the exact same variant

To next determine if this variant was more widely relevant, we used Sanger sequencing to test probands from other LFSL pedigrees (where *p53* was ruled out) to see if they had the exact same mutation in *ARHGAP30*. We tested 6 probands from Creighton University, 27 probands from MD Anderson Cancer Center (including STS200), 9 probands from Ohio State University, and 6 probands from the National Cancer Institute (NCI) for a total of 48 probands. Of these, three additional probands contained the exact same mutation in *ARHGAP30*, for a total of 4 of 48 (8.4%), suggesting that this mutation could be a potential hotspot.

Interestingly, this mutation is somewhat analogous to hotspot mutations in *p53*. In *p53*, several hotspot mutations occur in CpG sites, such as those at codons 175, 248, and 273, with deamination and G→A transitions being common (105). Our variant in *ARHGAP30,* GTTCGGC[G/A]AACCCAG, also occurs in a CpG site and may arise due to deamination. This may, at least in part, explain why the mutation could be a hotspot, and/or why the mutation is relatively common in the general population.

3.6.1 Statistical case for enrichment of *ARHGAP30* R806Q/R1017Q

To test statistically if these LFSL families were enriched for the mutation of *ARHGAP30*, we used to a Poisson approximation based on two reported allele frequencies. The dbSNP database contained the lowest reported MAF (0.86%), while the ESP (for European Americans) contained the highest MAF (2.72%). Using this information, we counted the number of cases we would expect to see on average, for 48 samples, and computed a one-sided p-value to determine the likelihood that we saw 4 or more cases. For dbSNP, this corresponds to a p-value of 8.7e-4, and for 4.4e-2 for ESP, suggesting that these LFSL pedigrees are enriched for the mutation in *ARHGAP30*.

The result of having multiple families with the same exact mutation in a tumor suppressor gene is unexpected, but not impossible. The majority of well-studied tumor suppressor genes and oncogenes follow a specific pattern. Oncogenes are typically recurrently mutated at specific amino acids, while tumor suppressor genes have flatter mutational profiles, with mutations, especially protein-truncation mutations, occurring throughout the length of the gene (54). However, *p53* is a notable exception for tumor suppressor genes where hotspots for cancer predisposition are observed (1.3.4).

3.6.2 Co-segregation in non-STS200 families

Subsequent testing of these families was somewhat limited. Of the two other LFSL families from MDACC, we had additional DNA from only one of the families (just the parents of the proband). In this family, of which the proband had a Wilms' tumor, one parent did contain the variant of interest but had not (yet) developed cancer. The other was wild-type.

3.6.3 Co-segregation is almost complete in the Creighton pedigree

The pedigree we received from Creighton University is substantial. We received both DNA and formalin-fixed paraffin-embedded (FFPE) blocks from Dr. Lynch to test for co-segregation in this family. For the DNA samples, we used Sanger sequencing, but for the FFPE blocks, we used pyrosequencing. Samples from this pedigree included FFPE blocks from before 1972, and contained highly cross-linked and fragmented DNA, making Sanger sequencing difficult. Therefore, using pyrosequencing, which amplifies a much smaller fragment (~100-bp vs. 500-800-bp) was advantageous. Even with the use of pyrosequencing, we found that we needed to use a nested PCR in a clean room to generate sufficient DNA from these samples and additional optimizations for FFPE samples as described by Doyle et al. (106), including enrichment of the polymerase (106).

These data showed that co-segregation was almost complete in this family. Three of four tested family members contained the exact same mutation, including two with cancer. The fourth is a theoretical obligate carrier. She has not yet developed cancer, but had two children with breast cancer (F, diagnosed at 35 y.o., dead at 36 y.o.) and lung cancer (M, diagnosed at 32 y.o. dead at 32 y.o.). We have requested a second sample from this individual for retesting, but have not yet received it.

We received one FFPE tumor sample from Creighton University; when tested this tumor had retention of heterozygosity (ROH). However, we note that loss of heterozygosity (LOH) may not be required for tumorigenesis. First, given the limited number of studies on ARHGAP30, it is not clear if the gene could be haploinsufficient. Secondly, in sequencing both human and mouse LFS tumors with *p53* mutations, it is clear that *p53* LOH is less common than otherwise appreciated. Therefore, we do not consider this single result showing ROH to rule out *ARHGAP30* from additional consideration.

To summarize, the same *ARHGAP30* mutation is found across multiple LFSL pedigrees, and appears to be enriched in LFSL families over the general population. For the most part co-segregation is complete in the families we have tested so far. Taken together these data suggest that this specific variant in *ARHGAP30* is a strong candidate for a putative cancer predisposition mutation, or at the very least, a modifier for cancer predisposition.

3.7 Structure of ARHGAP30

ARHGAP30 has one canonical domain, a RhoGAP domain towards the 5' end of the gene (codons 34-182), and two primary isoforms, dubbed here as a long isoform (L-ARHGAP30) and a short isoform (S-ARHGAP30). The two isoforms differ from each other only by a glutamic acid-rich repeat. The ARHGAP30 mutation is marked by an asterisk, and lies just outside this region (at R806Q in the short isoform and at R1017Q in the long-isoform).

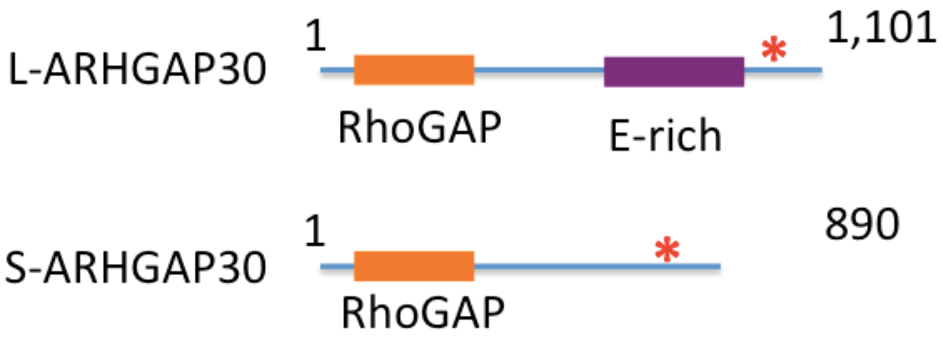


Figure 16: Depiction of important elements of ARHGAP30 in relation to the mutation. ARHGAP30 has only one known domain. The long and short isoforms differ only by a glutamic acid (E) rich element. However, the putative cancer predisposing mutation (marked by an asterisk) lies outside this element.

3.8 ARHGAP30 acts like a tumor suppressor gene

There have been exactly two papers published on ARHGAP30 to date. Both papers suggest that ARHGAP30 is a tumor suppressor gene with roles consistent with cancer (99, 100). In the first paper, the majority of figures lacked controls, and few conclusions can be drawn from it. However, data with controls indicate that when *WT ARHGAP30* (short isoform) is overexpressed, PAE/PDGFRb cells are less spread out, and more rounded, and therefore may have more migratory potential.

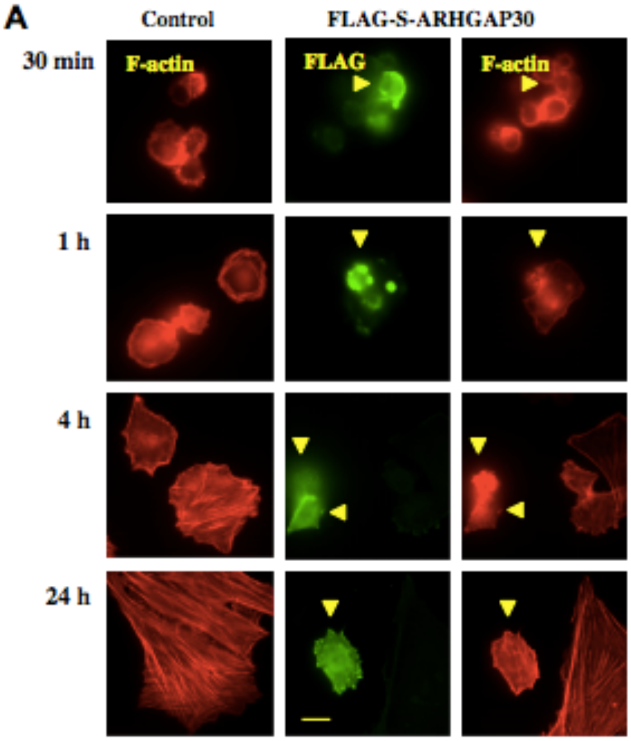


Figure 17: Figure 3a from Naji et al. 2011. In the presence of the overexpressed short isoform of WT ARHGAP30, cells appear significantly more rounded relative to untreated controls, and may have more migratory capacity. Used with permission: Naji, L., D. Pacholsky, and P. Aspenstrom. 2011. ARHGAP30 is a Wrch-1-interacting protein involved in actin dynamics and cell adhesion. *Biochemical and biophysical research communications* 409: 96-102. L/N 4294720235318.

The second paper, by Wang et al. is significantly more comprehensive (100). They showed that ARHGAP30 was significantly downregulated in colorectal cancer vs. normal tissues, and that it was consistently associated with poorer prognosis, having both larger tumors and more advanced stages of cancer on average, *ARHGAP30* was initially identified via a microarray analysis of TCGA data in colorectal cancer.

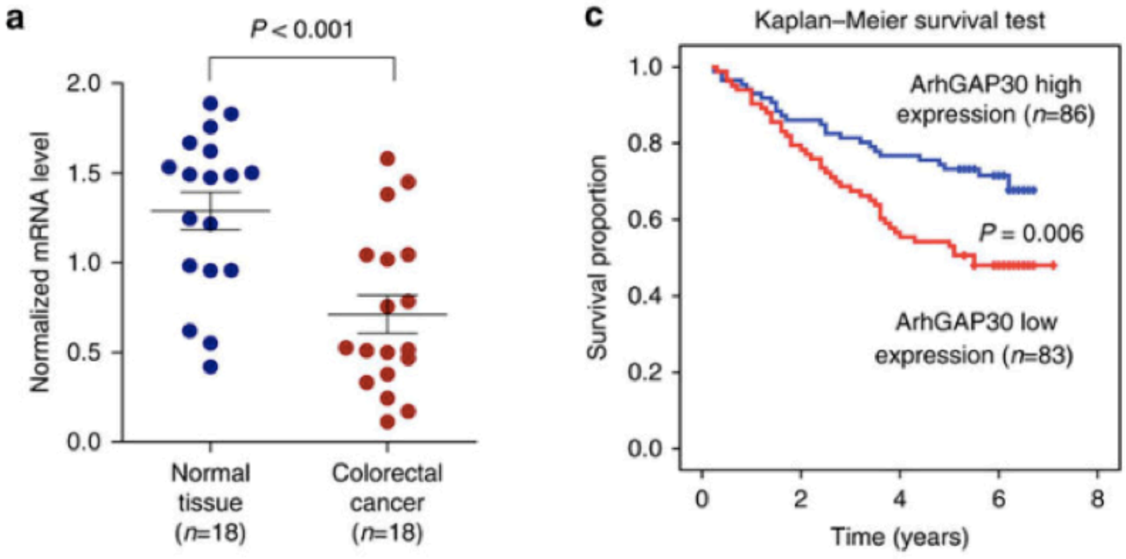


Figure 18: Figure 1a,c from Wang et al. 2014. a) qRT-PCR data showing ARHGAP30 expression was downregulated in colorectal cancer vs. normal tissues. c) Kaplan-Meier curve showing that higher ARHGAP30 expression levels was correlated with survival. Used with permission: Wang, J., J. Qian, Y. Hu, X. Kong, H. Chen, Q. Shi, L. Jiang, C. Wu, W. Zou, Y. Chen, J. Xu, and J. Y. Fang. 2014. ArhGAP30 promotes p53 acetylation and function in colorectal cancer. *Nat Commun* 5: 4735. L/N 4294720087364.

Importantly, we note that *ARHGAP30* and *p53* mutations should not be considered to be mutually exclusive. Under cBio (22), multiple tumors appear to have alterations in both genes (Figure 19), including a very small subset of samples that have truncating mutations in both *ARHGAP30* and *p53.*

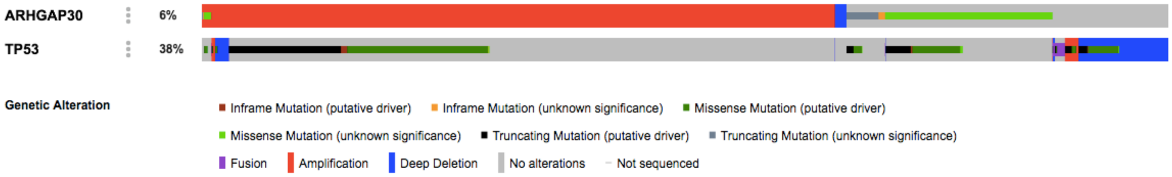


Figure 19: ARHGAP30 and p53 are not mutually exclusive according to data from cBio Oncoprint (22). Only a close up of the data is presented.

3.8.1 ARHGAP30 has a role in cell proliferation and migration

Wang et al. tested the impact of stable overexpression of ARHGAP30 across three main ideas: cell proliferation, cell migration, and apoptosis, finding that cells transfected with WT ARHGAP30 grew more slowly, moved less relative to controls, and promoted apoptosis (100).

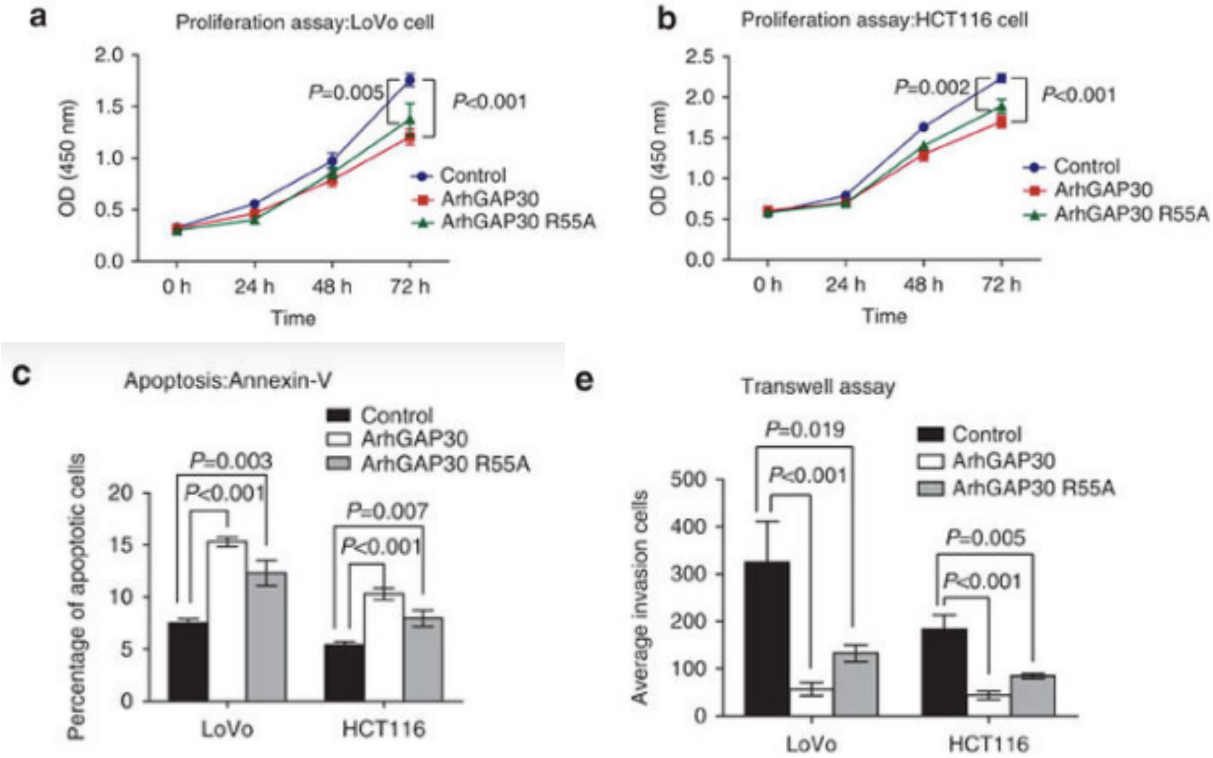


Figure 20: Figure 4a,b,c,e from Wang et al. 2014. (100) a) MTT assay showing proliferation of LoVo cells is decreased when ARHGAP30 is over-expressed. Moreover this functionality seems to be independent of the GAP domain since the effect in the R55A mutant, which has no GAP-relation functionality. b) HCT116 cells also grow more slowly in a GAP-independent manner. c) Both LoVo and HCT116 cells showed increased apoptotic activity when overexpressed, and e) decreased capacity to migrate relative to controls. Used with permission: Wang, J., J. Qian, Y. Hu, X. Kong, H. Chen, Q. Shi, L. Jiang, C. Wu, W. Zou, Y. Chen, J. Xu, and J. Y. Fang. 2014. ArhGAP30 promotes p53 acetylation and function in colorectal cancer. *Nat Commun* 5: 4735. L/N 4294720087364.

3.8.2 ARHGAP30 has a *p53*-dependent role in cell proliferation and migration

The authors also find that ARHGAP30 has *p53*-dependent effects. Ectopic expression of WT ARHGAP30 significantly upregulates *p53*-target genes like *p21, NOXA, BAX,* and *PUMA*, an effect that was also RhoGAP domain-independent in HCT116 cells. Moreover, when *p53*-null HCT116 cells were used, this effect was abrogated, and p53 target genes were not upregulated. These effects extended to proliferation (Figure 21j) and apoptosis assays (Figure 21k) with no significant effects observed, thus suggesting that ARHGAP30 has *p53*-dependent effects.

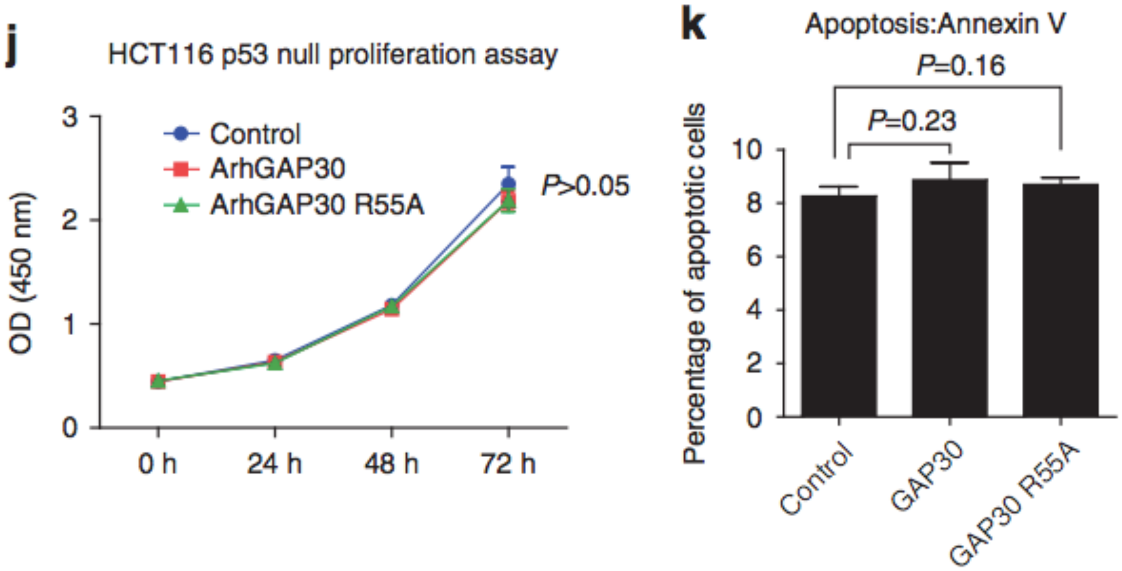


Figure 21: Figure 4.j,k from Wang et al. 2014 (100). J; In contrast to Figure 20b, in a *p53* null HCT116 line, ARHGAP30 overexpression does not suppress growth. k; similarly, they observe no changes in apoptosis in a *p53* null context. Used with permission: Wang, J., J. Qian, Y. Hu, X. Kong, H. Chen, Q. Shi, L. Jiang, C. Wu, W. Zou, Y. Chen, J. Xu, and J. Y. Fang. 2014. ArhGAP30 promotes p53 acetylation and function in colorectal cancer. *Nat Commun* 5: 4735. L/N 4294740034573.

These data persisted in the context of DNA damage. qRT-PCR data showed that cells treated with etoposide had upregulation of *p53* target genes, and that this effect was attenuated by knockdown of WT ARHGAP30. Moreover, chromatin immunoprecipitation (ChIP) analysis of cells with ARHGAP30 knockdown and etoposide treatment found reduced binding of *p53* to *p21, BAX, NOXA,* and *PUMA*. This effect seemed to be driven by acetylation of *p53* at Lys382 (K382), which is known to activate *p53*. Additional experimentation found that this effect was abrogated by knockdown of *p300*, and that under co-immunoprecipitation, ARHGAP30 can pull down both *p53* and *p300.* However, these *p53*-dependent effects were found to be long-isoform specific, rather than the short-isoform of ARHGAP30, suggesting a functional role for the glutamic acid rich element in the carboxy terminus of the gene.

3.9 *ARHGAP30 c.G161,017,761A/p.R>Q* has cancer-like functions for both the *long isoform (p.R1017Q) and the short isoform (p.R806Q)*

Given our strong genetic evidence, including co-segregation and recurrence, implicating *ARHGAP30 R1017Q/R806Q* as a putative cancer predisposition mutation, as well as the functional data presented by Wang et al. (100), we next tested if our mutation conferred any cancer-like advantages when transfected *in vitro.* In line with previous experiments, we first sought to test cell migration and cell proliferation.

We assayed several sarcoma cell lines, plus the HEK293T cell line as a transfection control. In addition, we perused the Cancer Cell Line Encyclopedia (CCLE) (107) to look for cell lines with no, or low expression of ARHGAP30 to better test the potential effects of the mutation, but weren't able to find any. Therefore, we picked the cell line with sarcoma lineage, and lowest expressing ARHGAP30*,* Hs 863.T. No CCLE cell lines had reported mutations, truncating, missense, or otherwise in *ARHGAP30*.

Table 8: Table of sarcoma cells lines used for transfection, with *p53* status and ARHGAP30 status included.

Cell Line	**Tissue**	***p53* status**	**ARHGAP30 status**
HEK293T	Kidney	*p53 R280S (108)w*	WT, Norm. Expr.
U-2 OS	Osteosarcoma	*p53 WT*	WT, Norm. Expr.
Saos-2	Osteosarcoma	*p53 deletion* homozygous*;* *c.1-1182del1182* (ATCC)	WT, Norm. Expr.
HT-1080	Fibrosarcoma	*p53 WT*	WT, High Expr.
Hs. 863T	Ewing Sarcoma	*p53 WT*	WT, Low Expr.

Given that LFSL arises out of a *p53 WT* background, we would certainly expect to see an effect in cell lines with WT *p53.* However, because *p53* mutations are so important in the context of LFS, and because ARHGAP30 has been demonstrated to interact with *p53*, we felt it valuable to concurrently consider the potential ramifications of cooperativity between the two genes. Moreover, *p53* null mice seemed to abolish the effect of ARHGAP30 across several phenotypes (100).

To generate the appropriate mutant plasmids, we acquired expression constructs for the short- and long- isoforms from Origene and used site-directed mutagenesis to introduce our putative LFSL mutation. After sequencing the WT and mutant plasmids to ensure the appropriate sequence and mutation were present, we next tested transfection across both Lipofectamine 3000 (LPF) and FuGene 6 (FG6), finding overall better results for FG6 (FG6), especially for Saos-2. Indeed for Saos-2, when staining for ARHGAP30, we observed practically no transfection when using LPF. For two other cell lines (HT-1080, 293T), we also observed that FG6 appeared to transfect at higher efficiencies. However, for the fourth cell line, U-2 OS, we observed the reverse; LPF appeared to perform better (Figure 22). For the fifth cell line, with low-expressing ARHGAP30 (Hs 863.T), the cells did not survive transfection with either FG6 or LPF and therefore could not be pursued further. Given that we did see successful transfection with FG6 for U-2 OS though, to best include Saos-2, we decided to use FG6 moving forward.

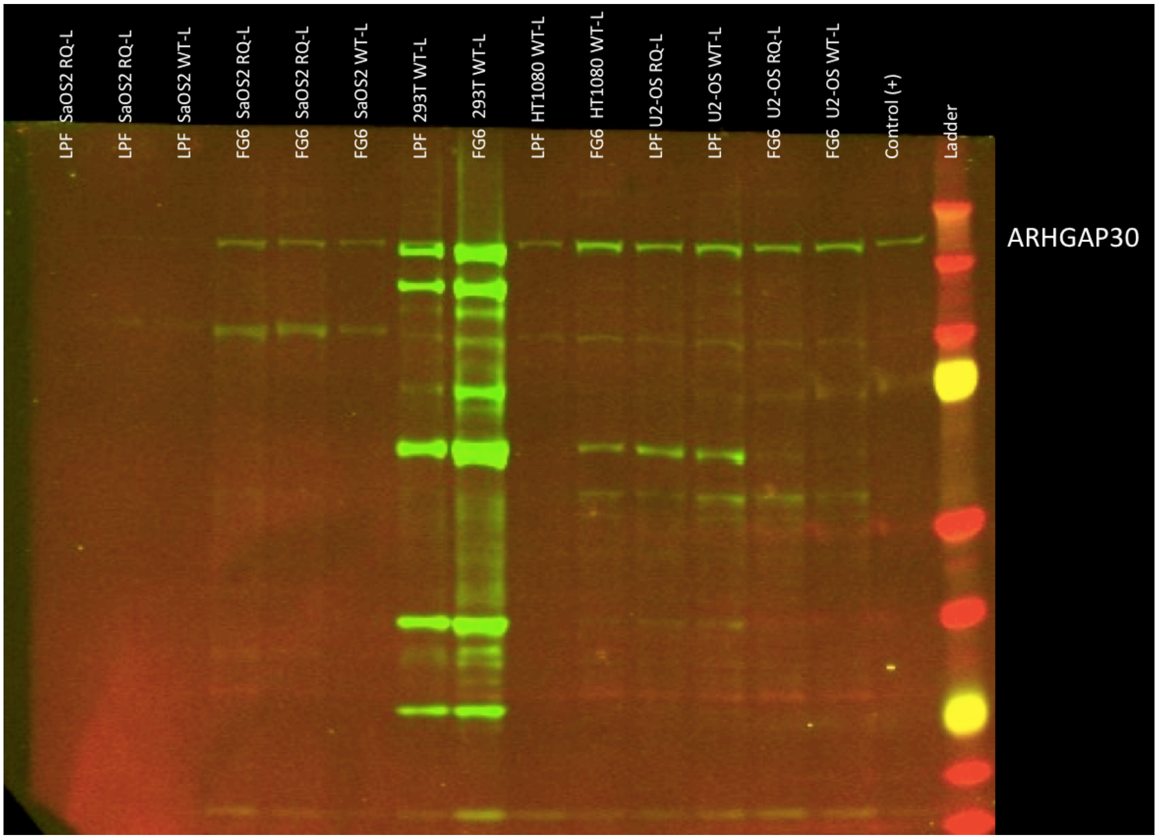


Figure 22: Western Blot performed in conjunction with Kevin Tracy (CPRIT summer student) comparing the efficacy of transfecting U-2 OS, HT-1080, HEK293T, and Saos-2 with either Lipofectamine 3000, or FuGene 6, across both wild-type and mutant ARHGAP30 (long-isoform). Stained with an ARHGAP30 antibody specific for long-isoform. Control was selected from a previous experiment with positive staining. The data suggest that overall, cells transfected with FuGene 6 appear to uptake more of the plasmid and/or produce more protein. This is especially true for Saos-2 (left edge of figure), where transfection was almost non-existent when using lipofectamine.

3.10 ARHGAP30 R806Q/R1017Q have increased migratory potential

To test migratory potential, we used a scratch assay.

3.10.1 HEK293T

In HEK293T cells, we observed that cells transfected with mutant ARHGAP30 appeared to show increased migratory capability. Due to the nature of the cell line, during scratching, cells came off in sheets, and were unable to limit the scratch to a single field-of-view.

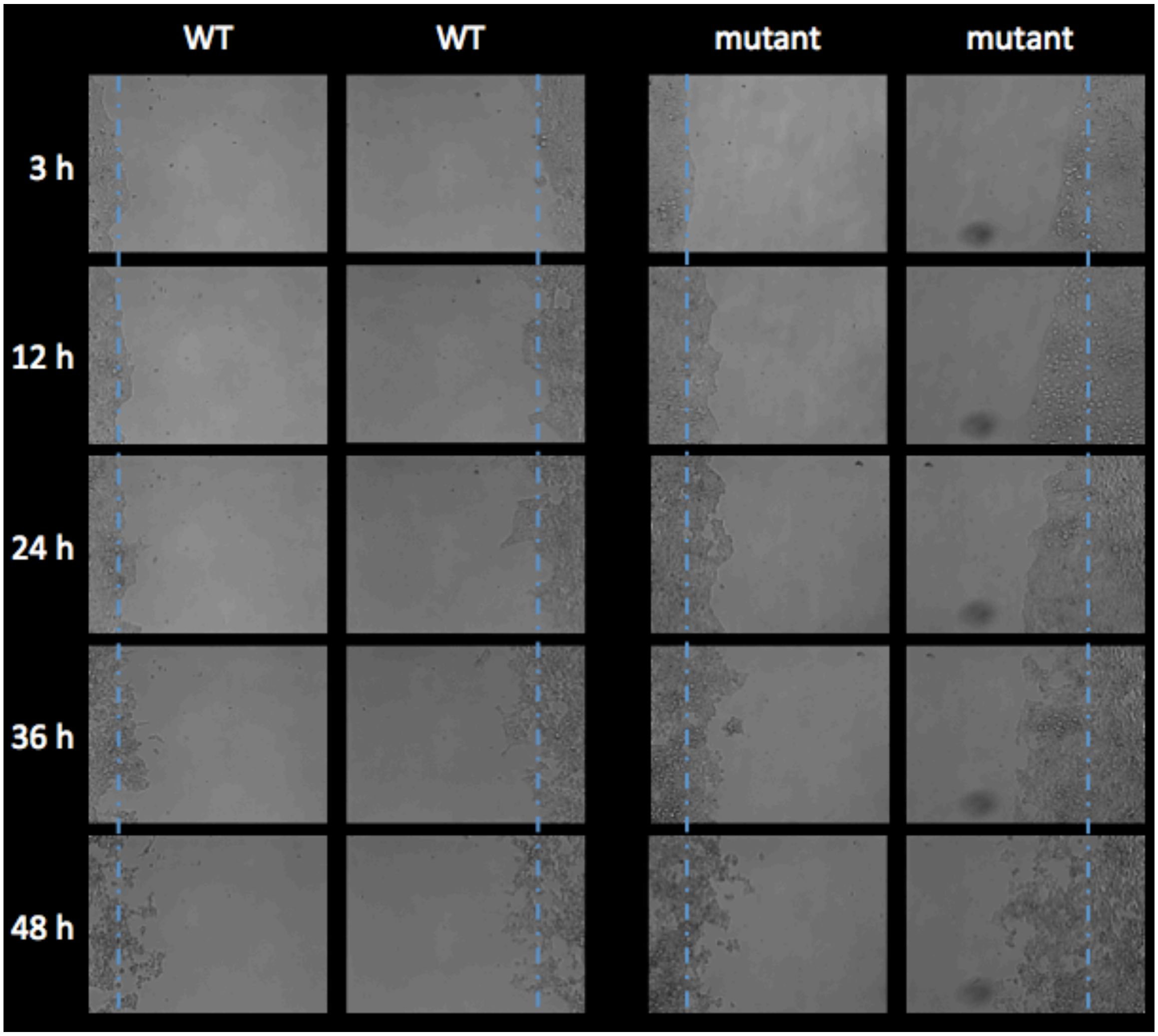


Figure 23: Scratch assay of HEK293T cells, transfected with either WT or mutant L-ARHGAP30. Cells transfected with mutant ARGHAP30 had increased migratory potential.

3.10.2 U-2 OS

U-2 OS cells are WT for *p53* and therefore fall best in line with what we would expect to see in an LFSL patient without inherited *TP53* mutations. In these cells, we see the most dramatic change of all tested cell lines when comparing the wound healing capability between transfection with WT and mutant plasmids. Cells transiently transfected with WT ARHGAP30 (both long- and short-isoforms) show limited ability to close the wound. In contrast, the mutant L-ARHGAP30 shows pronounced ability to close the gap with the shorter isoform showing more limited, but still significant ability to close the intervening space (Figure 24). Given that U-2 OS cells are WT for *p53*, these data demonstrate that the *ARHGAP30* mutation is sufficient to generate an effect even when *p53* is not disrupted, including in an LFSL patient-related context.

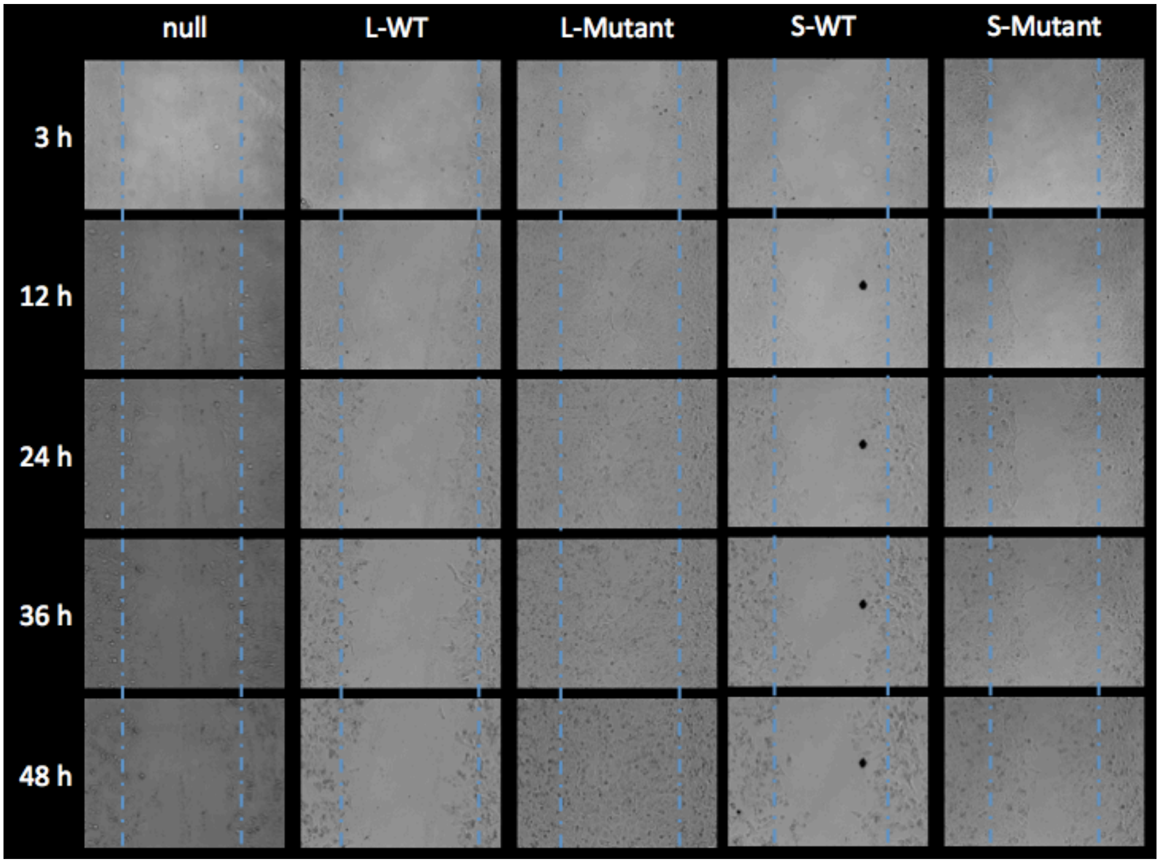


Figure 24: Scratch assay of U-2 OS cells shows dramatically improved ability to close the gap when transfected with mutant ARHGAP30, particularly for the long-isoform compared to WT-transfections.

3.10.3 Saos-2

Saos-2 cells are *p53-null*, but we again see increased ability of the cells to close the wound under transient transfection of the mutant relative to WT controls, although this effect is not as dramatic as in U-2 OS. One distinct difference here though is that the short-isoform shows increased closure relative to the long-isoform (Figure 25). Given that Saos-2 cells are null for p53*,* these data suggest that the effect is present even in the absence of p53. Although the result is not as accelerated as in U-2 OS cells, taken together these data continue to point towards a *p53*-independent mechanism.

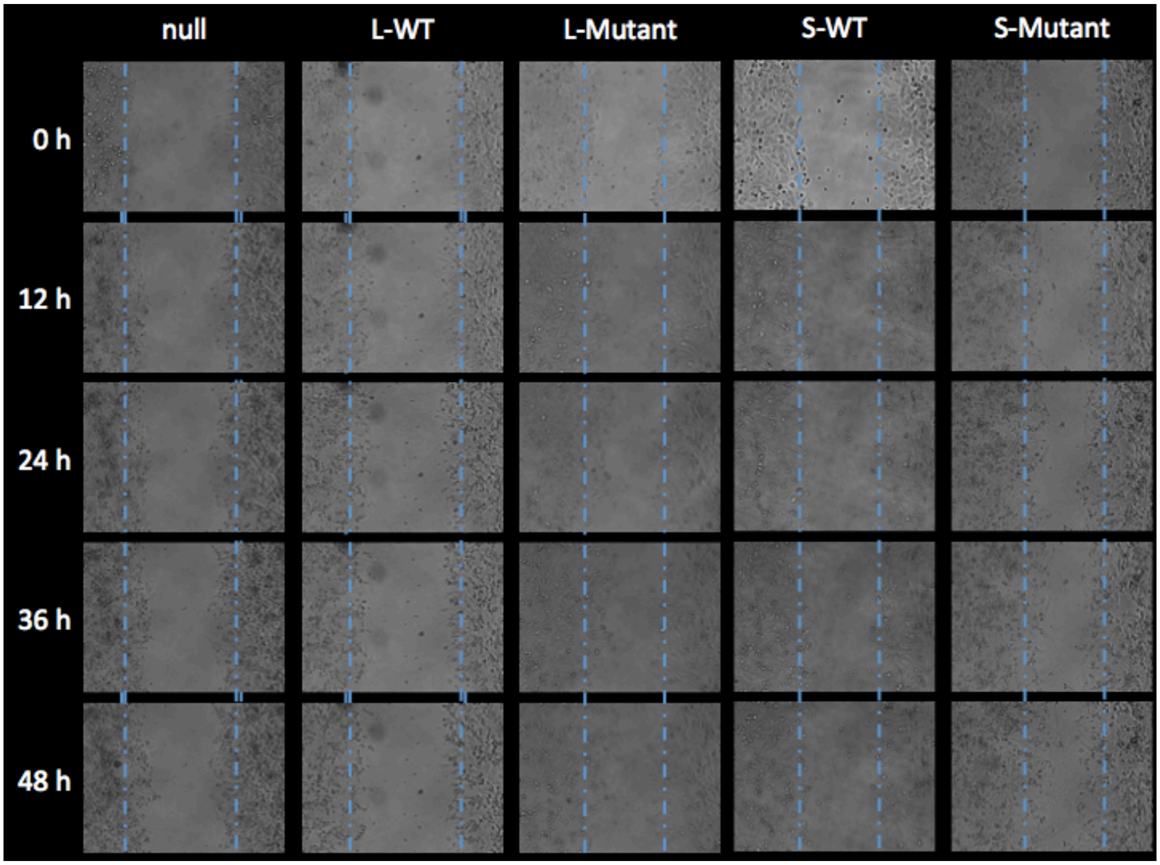


Figure 25: Scratch assay of Saos-2 cells. Mutant ARHGAP30 was not able to close the gap entirely over 48 hours, but did show increased movement relative to cells transfected with WT ARHGAP30.

3.10.4 HT-1080

Similar to U-2 OS cells, HT-1080 cells are WT for *p53* (53, 107). However, in contrast to previous data, the endogenously high ARHGAP30 HT-1080 cells closed remarkably quickly, including when the WT-ARHGAP30 plasmids are overexpressed, having closed after just 24-hours (Figure 26). Notably, even the untreated (?) control (with endogenously high ARHGAP30 levels) is able to somewhat close the wound (though not completely). Again, mutant ARHGAP30 shows increased ability to close relative to WT transfections, but this effect is somewhat muted given the rapid closing.

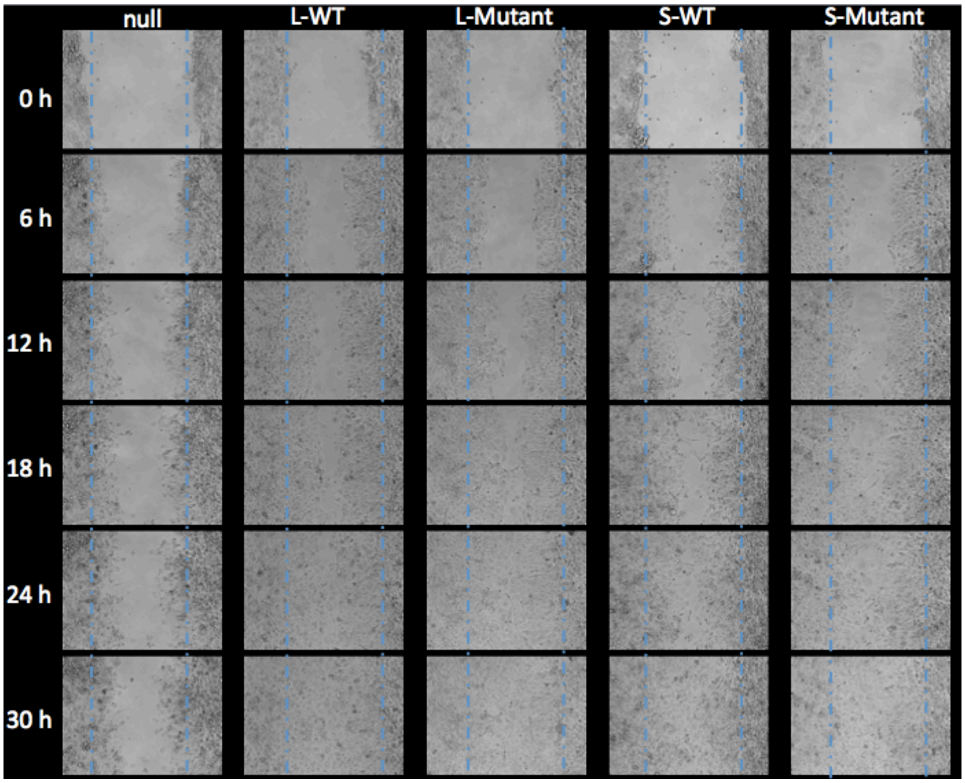
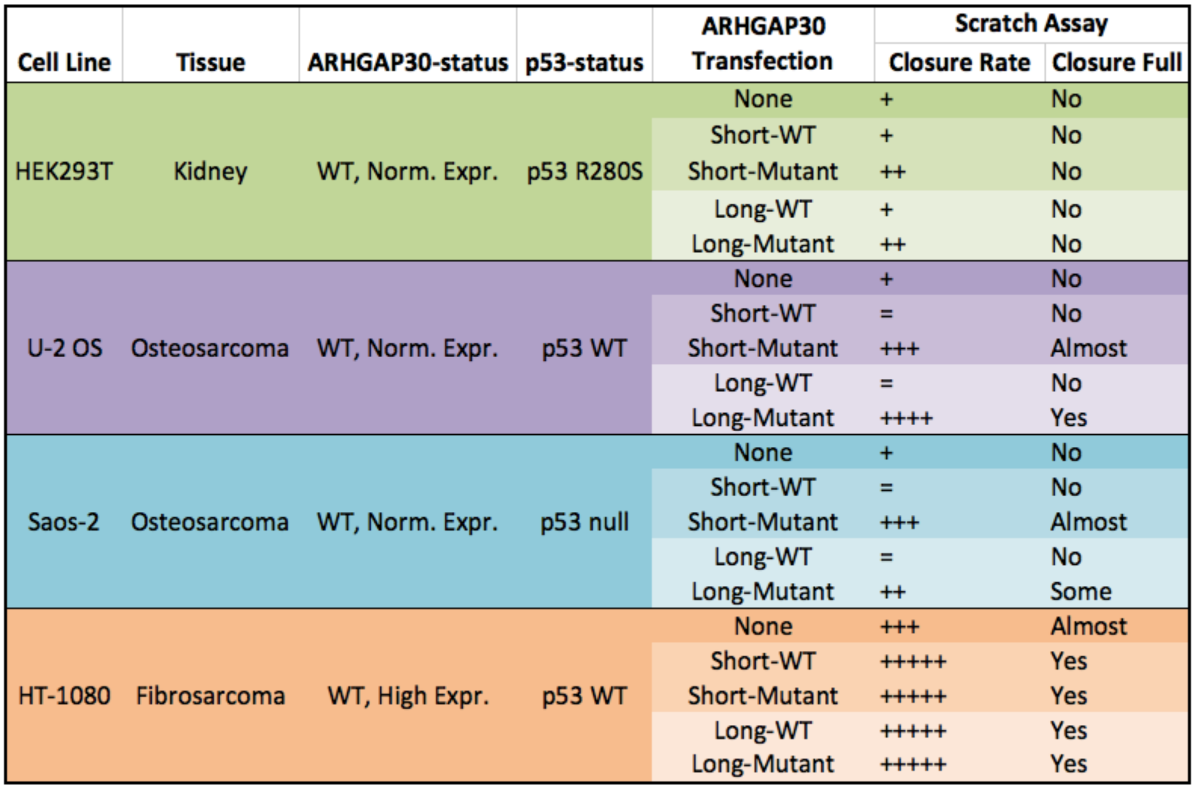


Figure 26: Scratch assay of HT-1080 data. Both WT ARHGAP30 and mutant ARHGAP30 are able to close the wound over 30 hours. Mutant ARHGAP30 does lead to marginally faster wound closure relative to WT ARHGAP30.

3.10.5 Summary of wound healing experiments

These data can be summarized either qualitatively (as shown in Figure 27) or quantitatively.



Cell Line	Tissue	ARHGAP30-status	p53-status	ARHGAP30 Transfection	Scratch Assay	
					Closure Rate	Closure Full
HEK293T	Kidney	WT, Norm. Expr.	p53 R280S	None	+	No
				Short-WT	+	No
				Short-Mutant	++	No
				Long-WT	+	No
				Long-Mutant	++	No
U-2 OS	Osteosarcoma	WT, Norm. Expr.	p53 WT	None	+	No
				Short-WT	=	No
				Short-Mutant	+++	Almost
				Long-WT	=	No
				Long-Mutant	++++	Yes
Saos-2	Osteosarcoma	WT, Norm. Expr.	p53 null	None	+	No
				Short-WT	=	No
				Short-Mutant	+++	Almost
				Long-WT	=	No
				Long-Mutant	++	Some
HT-1080	Fibrosarcoma	WT, High Expr.	p53 WT	None	+++	Almost
				Short-WT	+++++	Yes
				Short-Mutant	+++++	Yes
				Long-WT	+++++	Yes
				Long-Mutant	+++++	Yes

Figure 27: Qualitative description of the closure results across the four cell lines, and various transfection conditions. More "+", implies a faster closure rate.

Quantitative representation can be done in two ways. First, we can try to find the size of the wound by finding the leading edges of the wound and using a lasso-type tool to try to determine the most appropriate area. This approach has challenges when the leading edge is not well defined – either from being slightly out of focus, or from live cells which are migrating, but not part of a distinct edge. It also has the potential to be subjective. The second approach uses a threshold tool; since cells are often darker than the unclosed wound, this fact can be used to identify the percentage of the field of view that is covered by cells. This has the advantage of perhaps being more unbiased, and is better at picking up lone cells in the middle of the wound, or at correcting for cases where cells may be migrating more than proliferating (leaving some areas as being less dense).

We have used the first method, measuring the area every three hours, because we found it to be less error prone given the quality of pictures that we obtained, but quantification makes one simple fact clear. Despite efforts to the contrary, it is difficult to create single scratches with the same exact width – see different starting points for % Covered for various treatments (Figure 28).

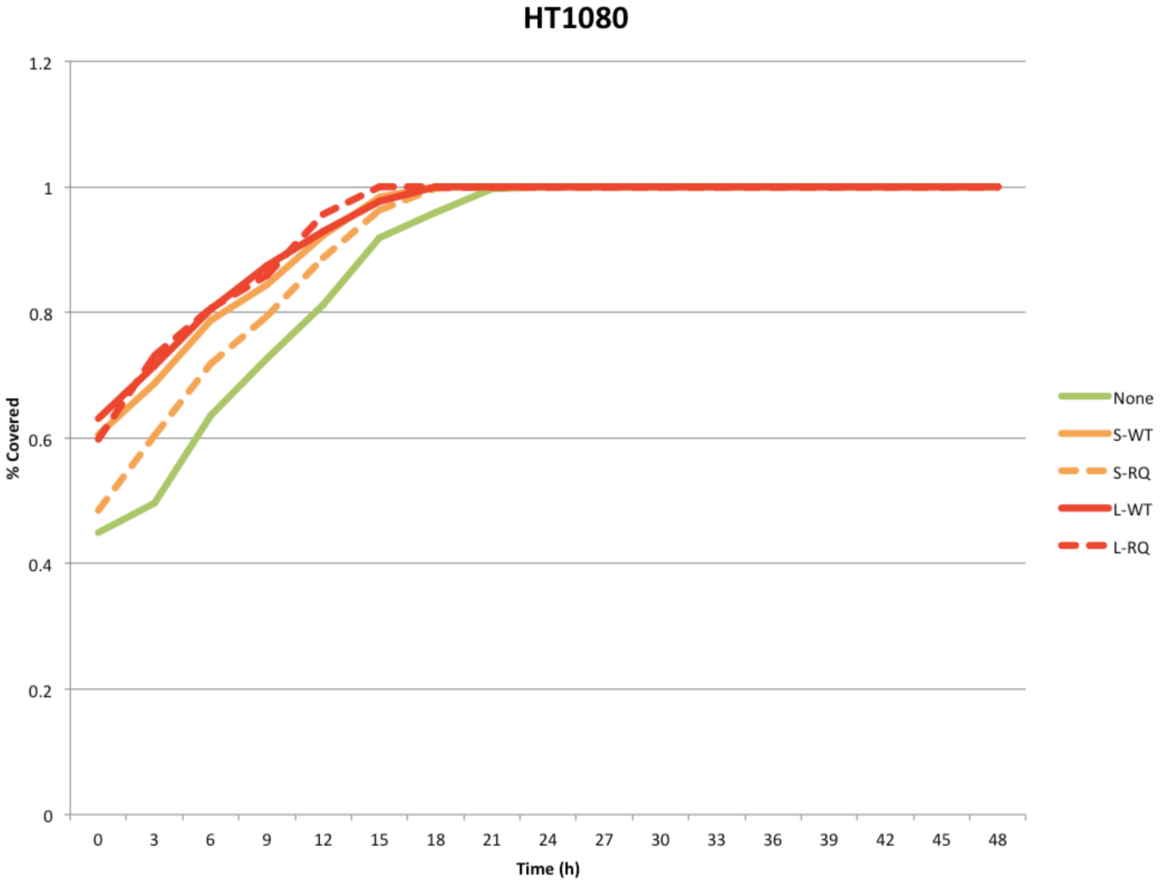


Figure 28: Quantification of wound closure in HT-1080 cells.

3.11 ARHGAP30 R806/1017Q proteins show increased proliferative capacity

3.11.1 HEK293T

Consistent with *ARHGAP30*’s status as a tumor suppressor gene, in 293T cells we observed suppression of growth when cells were treated with *WT ARHGAP30*, but when treated with mutant *ARHGAP30* we saw although they did not confer a proliferative advantage relative to a control, they did appear to abolish the slower growth rates observed when overexpressing *WT ARHGAP30*.

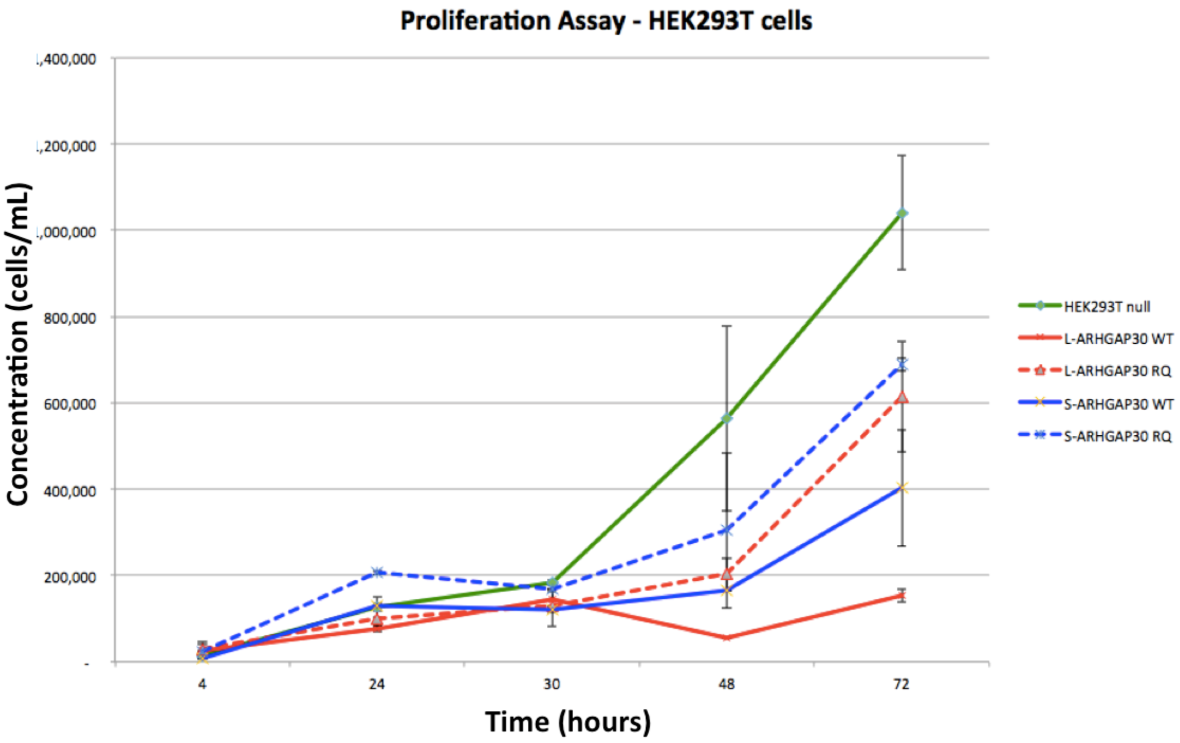


Figure 29: Proliferation assay of HEK293T cells showing reduced growth rates in conjunction with WT ARHGAP30 transfection. Cells transfected with mutant ARHGAP30 in contrast show increased proliferation.

3.11.2 U-2 OS

In contrast, in U-2 OS cells, we saw a much more muted response, although the overall phenotype is similar to 293T cells. Cells transfected with WT *ARHGAP30* did grow more slowly relative to a control, and cells transfected with mutant *ARHGAP30* appeared to abrogate some of that response bringing it nearly in line with the control, but not above the control.

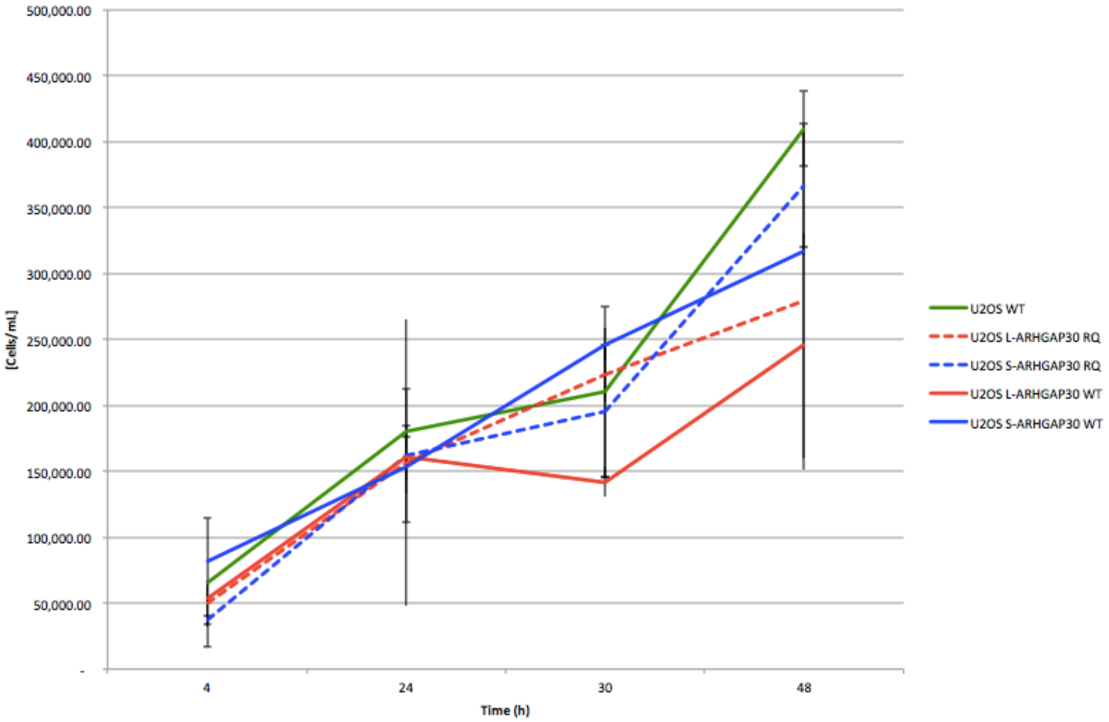


Figure 30: Proliferation assay of U-2 OS has minimal differences between the different conditions. However, there do appear to be slight differences consistent with ARHGAP30 being a tumor suppressor gene, and the mutant abrogating some tumor suppressor-like behavior via increased growth rates.

3.11.3 Saos-2

Data for Saos-2 appears to run somewhat counter to the previous data. We saw a trend for the long isoform that seems to be similar to that observed for 293T and U-2 OS cells, but not for the short isoform. Namely transfection with WT *ARHGAP30* suppresses growth, while transfection with mutant *ARHGAP30* leads to higher growth rates more in line with the control. However, we note that making this assumption assumes that cells were seeded at the same initial concentration at time of zero hours.

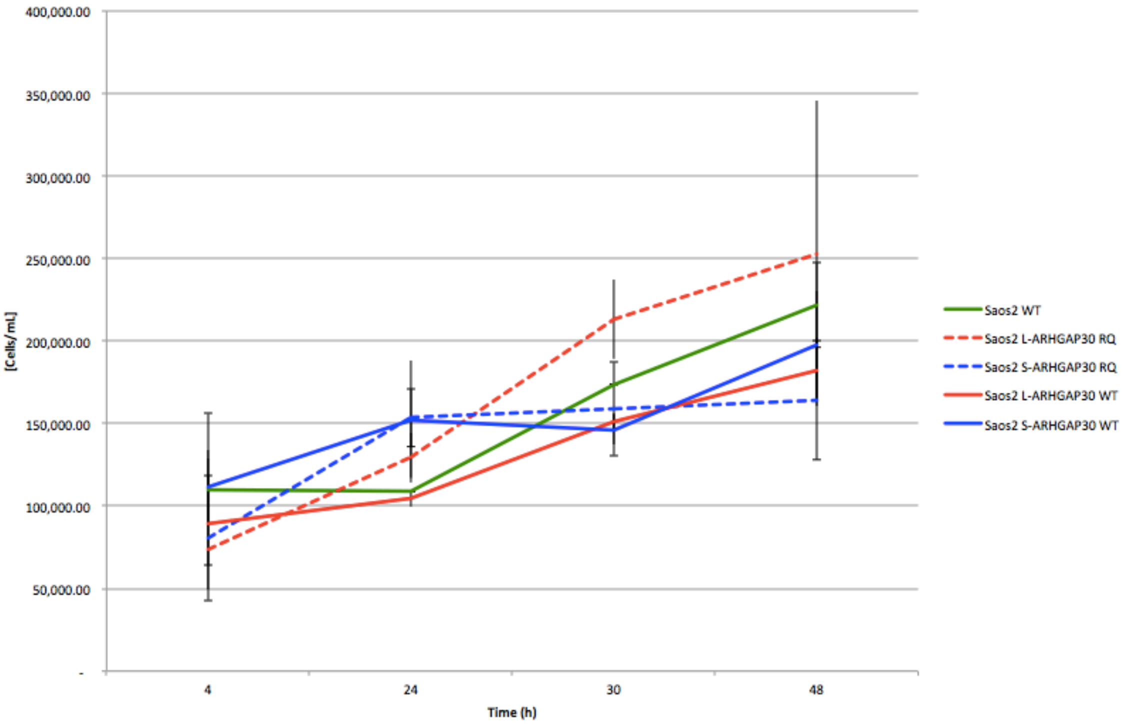


Figure 31: Proliferation assay of Saos-2 showing that mutant ARHGAP30 appears to have an effect in the long-isoform, but not the short-isoform.

3.11.4 HT-1080

Similar to the scratch assay, in HT-1080 cells, it is difficult to tease out a difference in proliferation rates between transfecting with WT or mutant *ARHGAP30* for both the short and long isoforms.

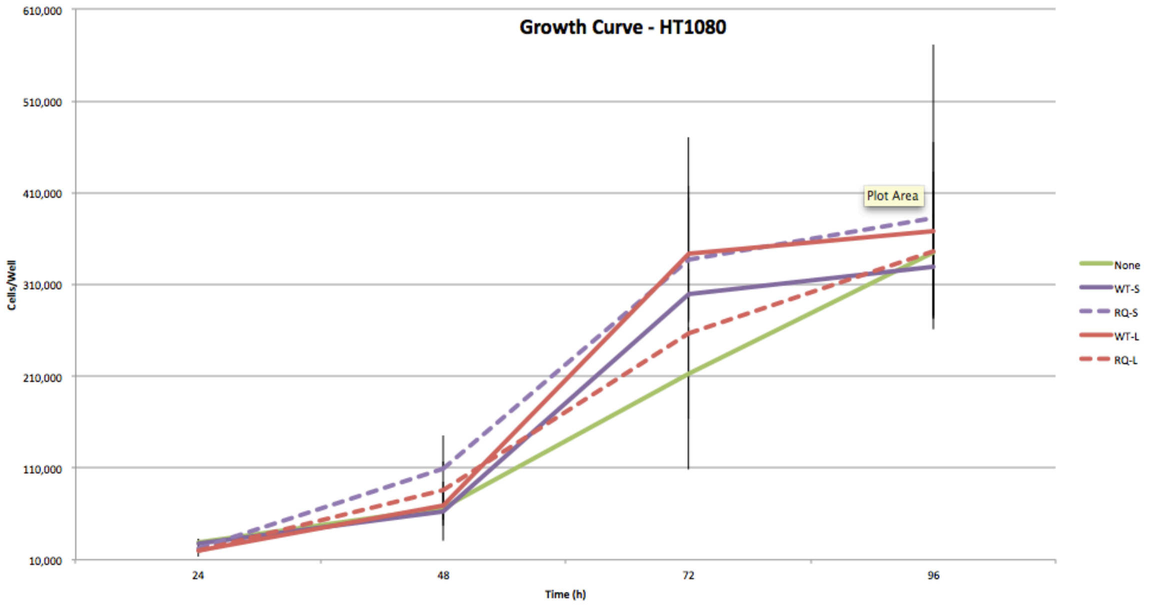


Figure 32: Proliferation assay for HT-1080s showing minimal differences between the various transfections.

4 Discussion

Co-segregation analysis was used to isolate a plausible LFSL gene. We identify a reasonably rare, co-segregating mutation in 1q23 that is recurrent across multiple LFSL pedigrees. Moreover, this gene has demonstrated significance in colorectal cancer in the literature (100), and our functional evidence shows the mutation itself appears capable of inducing cancer-like phenotypes such as proliferation and migration.

The functional data presents an interesting picture when juxtaposed against the results from Wang et al. (100). Previously, these authors showed ARHGAP30 overexpression was found to have no effect on apoptosis, proliferation, and expression of *p53* target genes in *p53* null HCT116 cells (100). This left one big arm of their functional work unexplored – what happens to migration in an ARHGAP30 overexpression/*p53* null background. If cell migration is abrogated alongside apoptosis and proliferation in p53 null HCT116 cells, it runs in direct contrast to what we observe in Saos-2 cells. In Saos-2 cells with the cancer predisposing mutation, cells are able to close the cap relative to overexpression of WT ARHGAP30. This suggests that not only does the *ARHGAP30 p.R>Q mutation* have a *p53*-independent means of impacting tumorigenesis, but also that it may be a gain-of-function mutation.

Consistent with this interpretation, we observe wound healing and proliferation phenotypes across both S-ARHGAP30 and L-ARHGAP30, where as data from Wang et al. (100) suggest that *p53* acetylation is L-ARHGAP30 specific. If our mutation were to impact phenotypes through acetylation, we would not expect to see differences in S-ARHGAP30.

Interestingly, the effect on wound healing is greatest in *WT p53* U-2 OS cells, placing further weight on a *p53*-independent mechanism for mutant ARHGAP30. However, we acknowledge that some of this difference could be due to reduced transfection efficiency (Figure 22), and/or cell properties. Gaps in wound healing are typically closed through cell proliferation and/or cell migration and of the cells cultured, Saos-2 cells grew the slowest.

Our data from HT-1080 are curious. ARHGAP30 expression is constitutively high in this cell line relative to other cell lines that we used (Figure 33), and has no known mutation in *ARHGAP30* according to the CCLE (107). Given that *ARHGAP30* is a putative tumor suppressor gene, and the previous establishment that WT ARHGAP30 can negatively impact cell motility (99, 100), we might expect these cells to have the most limited response to the wound healing assay. Instead, they have the most pronounced effect, closing within 30 hours, perhaps suggesting that these cells have acquired some sort of escape mechanism. Moreover, additional WT ARHGAP30 exacerbates the closure relative to no-treatment controls, which, on its own, implies ARHGAP30 acts more like an oncogene in this cell line. Data from the wound healing assay indicates that overexpression of the mutant construct slightly improves closure relative to WT ARHGAP30, but overall these data suggest that the contribution of ARHGAP30 to tumor etiology in HT-1080 may be considerably more complex than first expected.

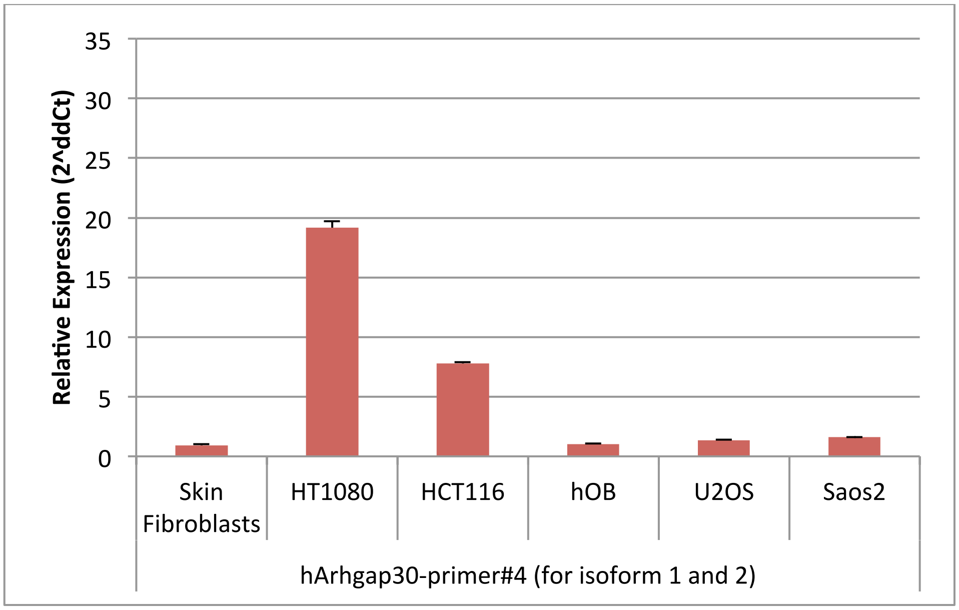


Figure 33: qPCR data for several cell lines, showing that ARHGAP30 expression in HT-1080 cells are significantly higher than healthy skin fibroblasts, a colorectal cancer cell line (HCT116) osteoblasts (hOB), or two osteosarcoma lines (U2OS, Saos-2).

In addition, we had previously begun to look at the potential effect of *ARHGAP30* mutations on downstream effectors of *p53* (Figure 34). These data were generated by Dr. Yu Deng in the Krahe Lab. She used lymphoblastoid cell lines (LBCLs) from STS200, three with the *ARHGAP30* p.R>Q mutation, and one with WT ARHGAP30 ("009"). She cultured all four lines, and then treated each with and without radiation. Focusing only on the untreated samples, we see that as expected, levels of *p53* are about the same across all individuals. However, the levels of *p21* are not the same. In patient-derived LBCLs with the mutant ARHGAP30, we saw lower levels of *p21.* Thus, it is possible that ARHGAP30, in addition to interacting with *p53*, may also be part of a *p21, p53*-independent.

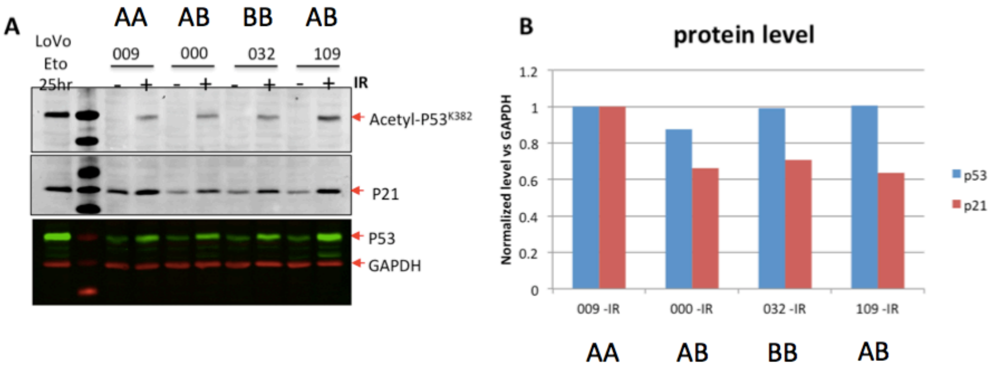
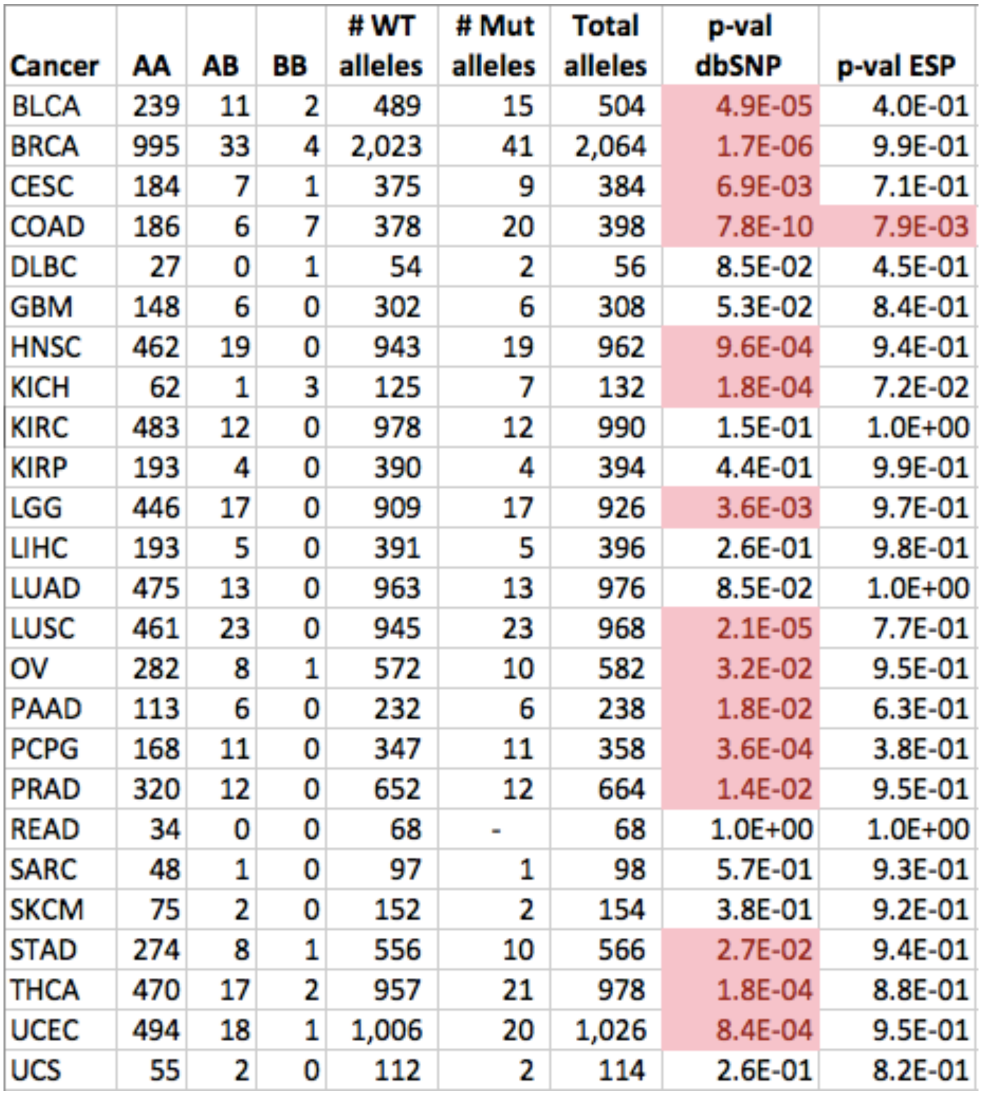


Figure 34: Western blot of p53 and p21. Data generated by Yu Deng, Ph.D. (A) Western blot showing protein levels of *p53* and *p21* in STS200-derived cells. LBCLs from STS200 were untreated (-) or irradiated (+) to check for *p53* acetylation in response to the stressor. Concomitantly, we checked for a response in a *p53* downstream target, *p21.* (B) Quantification of the protein levels of *p53* and *p21* in non-irradiated cells shows reduced *p21* levels in cancer patients. ARHGAP30 genotypes are provided above/below patient IDs to aid interpretation: AA – homozygous WT, AB – heterozygous mutant, BB – homozygous mutant.

We next considered potential relevance to sporadic sarcomas. Under cBio, ARHGAP30 is recurrently mutated in just 0.7% of all samples, and our specific germline variant has not yet been called in publicly available somatic mutation data. However, it turns out that our mutation is assayed by SNP arrays. We worked with Dr. Keith Baggerly and Dr. Ying Wang to acquire these data and to generate a list of tumors and their genotypes. We sorted by genotype, letting the A-allele stand for the WT-allele, and the B-allele for the mutant-allele. We then counted total alleles, and ran a one-sided Poisson approximation to see how likely we were to uncover at least that many alleles. We ran these data under two conditions, once for the lowest reported MAF (0.86%, dbSNP) and once for the highest (2.72%, ESP-EA). At the lower allele frequency, multiple cancers appear to be enriched for the mutation, but when using the ESP as a baseline, there is just one showing statistical significance, colorectal adenocarcinoma (COAD). We further note that for COAD tumors, there are seven tumors with a "BB" genotype, versus six with an "AB" genotype – these data imply that LOH may be a reasonably common event in these tumors.

Table 9: Table showing allele incidence of *ARHGAP30 c.G>A, p.R>Q* in TCGA SNP array data (data generated by the TCGA: "http://cancergenome.nih.gov"). p-values are calculated via a Poisson approximation to determine if we saw enrichment for the number of mutant alleles relative to expectation based on MAF (dbSNP, 0.86%; ESP, 2.72%). Data highlighted in red have p-values less than 0.05.



Cancer	AA	AB	BB	# WT alleles	# Mut alleles	Total alleles	p-val dbSNP	p-val ESP
BLCA	239	11	2	489	15	504	4.9E-05	4.0E-01
BRCA	995	33	4	2,023	41	2,064	1.7E-06	9.9E-01
CESC	184	7	1	375	9	384	6.9E-03	7.1E-01
COAD	186	6	7	378	20	398	7.8E-10	7.9E-03
DLBC	27	0	1	54	2	56	8.5E-02	4.5E-01
GBM	148	6	0	302	6	308	5.3E-02	8.4E-01
HNSC	462	19	0	943	19	962	9.6E-04	9.4E-01
KICH	62	1	3	125	7	132	1.8E-04	7.2E-02
KIRC	483	12	0	978	12	990	1.5E-01	1.0E+00
KIRP	193	4	0	390	4	394	4.4E-01	9.9E-01
LGG	446	17	0	909	17	926	3.6E-03	9.7E-01
LIHC	193	5	0	391	5	396	2.6E-01	9.8E-01
LUAD	475	13	0	963	13	976	8.5E-02	1.0E+00
LUSC	461	23	0	945	23	968	2.1E-05	7.7E-01
OV	282	8	1	572	10	582	3.2E-02	9.5E-01
PAAD	113	6	0	232	6	238	1.8E-02	6.3E-01
PCPG	168	11	0	347	11	358	3.6E-04	3.8E-01
PRAD	320	12	0	652	12	664	1.4E-02	9.5E-01
READ	34	0	0	68	-	68	1.0E+00	1.0E+00
SARC	48	1	0	97	1	98	5.7E-01	9.3E-01
SKCM	75	2	0	152	2	154	3.8E-01	9.2E-01
STAD	274	8	1	556	10	566	2.7E-02	9.4E-01
THCA	470	17	2	957	21	978	1.8E-04	8.8E-01
UCEC	494	18	1	1,006	20	1,026	8.4E-04	9.5E-01
UCS	55	2	0	112	2	114	2.6E-01	8.2E-01

Despite these data, there are a couple of mitigating questions remaining to be answered. First, the co-segregation analyses suggest that co-segregation may not be canonically complete. However, we believe there are plausible, if rare, scenarios that could explain these seemingly “outlying” individuals.

First, for the one individual in STS200 that is homozygous for the mutation of interest (STS200-032), it is unlikely, but not impossible that she received a mutant allele from both parents. Alternatively, she could have experienced an extremely early gene conversion event that resulted in an extended stretch of LOH. To address this possibility we used our WGS data and looked in the region near *ARHGAP30* and plotted whether variants were homozygous for the alternate allele (blue), or heterozygous for the alternate allele (red). If the sample was homozygous WT, we left it white/blank. In this region, it does appear as if the STS200-032 sample is on average more homozygous for the SNVs/SNPs interrogated (Figure 35), though it impossible to say for sure. A third, related possibility is uniparental disomy (UPD), in which two copies of the chromosome come from the same parent. This may also help explain some of the other heterozygous mutations in this region in STS200-032, the majority of which are rare. Unfortunately, to definitively distinguish between all possibilities is impossible in the absence of an available DNA sample from the mother.

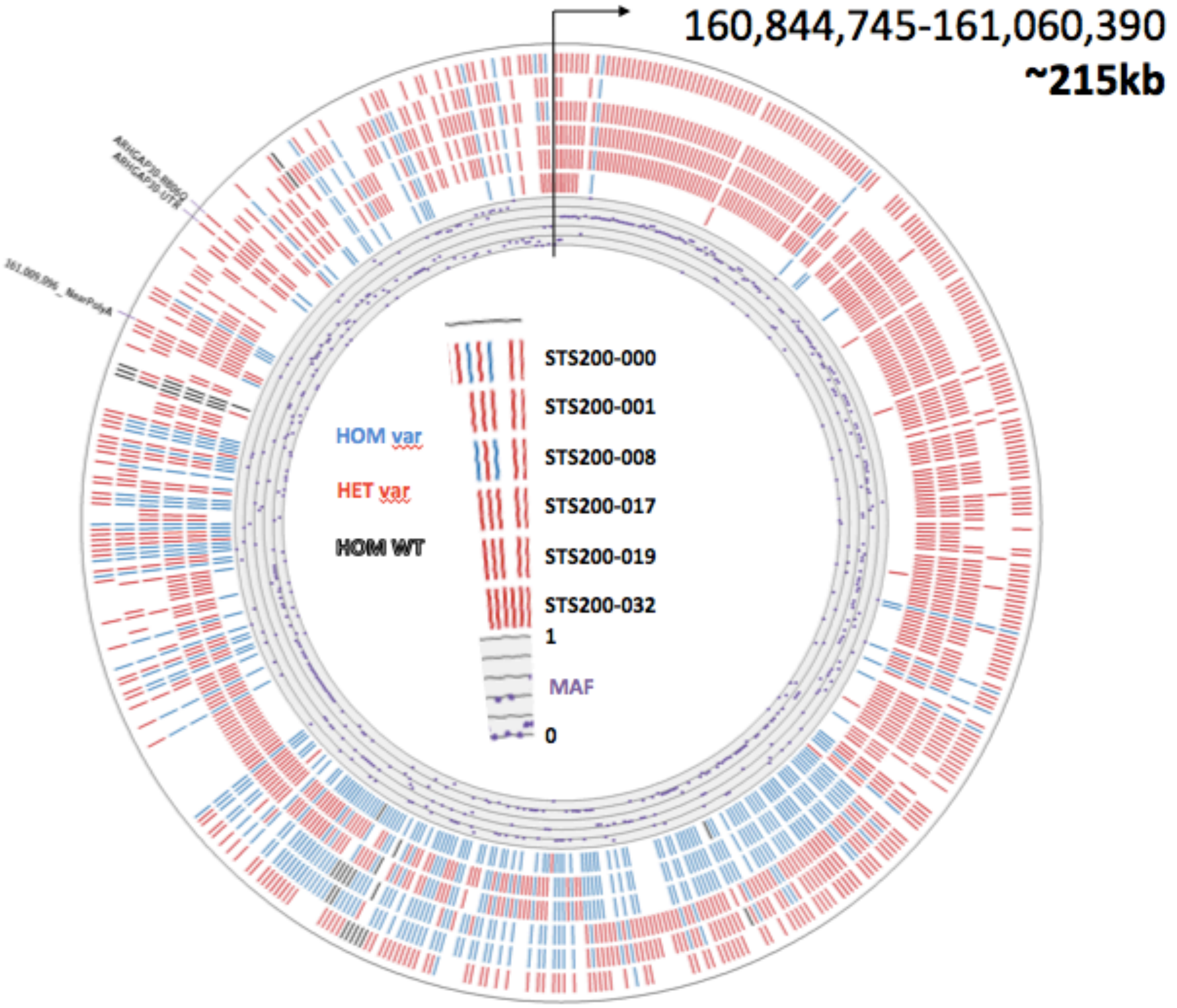


Figure 35: CIRCOS plot showing possible gene conversion event. Homozygous events are in blue and white/blank. Heterozygous events are red. The majority of samples have significant heterozygous variants in this region, with the exception of STS200-032. We feel that this data could be consistent with a gene conversion event.

In addition to STS200, one other co-segregation question remains under investigation. One theoretical obligate carrier in the Creighton pedigree did not have the *ARHGAP30* mutation. She had two children with LFSL-associated cancers (breast and lung cancer, both before the age of 40). However, she is in a more distant branch of the pedigree, and it is possible that she has a second, non-*ARGHAP30* cancer predisposition gene. Alternatively, it is possible that the father may have carried a cancer-predisposing mutation; however, no information is known about his family.

However, we do not consider the MAF a major sticking point. As previously discussed, emerging research suggests that germline *p53* variants may be more common than previously expected. Thus, it may be perfectly reasonable to have a more common, less penetrant allele (*p53* or non-*p53*), even for a rare disease. Interestingly, we do not observe any variants in known cancer genes – this implies either that ARHGAP30 has the potential to have a strong monogenic effect, or that there are additional yet to be uncovered polygenic or modifier genes that further modulate sarcomagenesis. For example *MDM2 SNP309*, with a high MAF, affects tumorigenesis (1.3.1).

5 Future Directions in ARHGAP30 R806Q/R1017Q

Our evidence strongly supports that *ARHGAP30* and the *p.R806/1017Q* mutation is a recurrent, co-segregating, cancer predisposition gene in LFSL families. However, we note that these data do have some mitigating question marks. The variant is not overly rare, putting it in line with a previous purported LFSL gene/mutation *CHEK2 1100delC,* and in the STS200 pedigree has an unexpected outlier, a single homozygous case. Moreover, in the one tumor we have been able to sequence, we were not able to detect LOH. However, we feel that there are plausible explanations for each of these scenarios and thus do not derail the underlying genetic and functional evidence presented. We further note functional experiments would best be done in backgrounds of null *ARHGAP30* to determine if this was a GOF mutation.

We propose two fundamental ways to move forward with this project: First, sequence (non-*p53*) LFSL pedigrees lacking p.R806/1017Q mutations for additional mutations elsewhere in *ARHGAP30*. If ARHGAP30 follows the pattern of other tumor suppressor genes, including *p53*, it should have mutations elsewhere in the gene that also give rise to LFSL. Identification of further families with additional mutations would greatly strengthen the argument that *ARHGAP30* is an LFSL, cancer-predisposition gene. Secondly, explore the functional implications *in vivo* through the use of a mouse model. CRISPR technology has greatly impacted the ease and feasibility of generating mice with single point mutations such as in the *ARHGAP30 RQ* variant. Following generation of appropriate mouse model and cohort with the mutation(s) of interest, we can the track and observe tumor formation in the mice. Moreover, in conjunction with this, it will be important to check for LOH in tumors, both in the mice, or in humans, where tumor samples are available, to look for characteristic LOH.

We note that there are many other possible *in vitro* experiments that could be run, to explore other potential functional implications. For example, we have not yet looked at apoptosis, in which ARHGAP30 plays a known role. (100) More to the point, we have so far focused on overexpression of ARHGAP30 in cells with endogenous ARHGAP30; therefore, it may be worth interrogating the results in an ARHGAP30 null background, or comparing results against the cell lines when ARHGAP30 has been silenced.

6 Introduction: Sarcomagenesis and Somatic Mutations

6.1 LFS is a model for identifying acquired somatic mutations in sarcomagenesis

Cancer has been attributed to a variety of factors, including germline cancer predisposition mutations, as well as somatic alterations. In the first portion of the dissertation, we leveraged LFSL to identify a novel germline cancer predisposition mutation in *ARGHAP30*. In the second portion, we leverage tumors from LFS patients, with known *p53* mutations, to address the potential contributions of somatic alterations. Despite the ubiquity of *p53* mutations in human cancers, some individuals in LFS pedigrees (with *p53* mutations) do not develop cancer (56, 93), or develop cancer very late, suggesting that the *p53* mutation alone is insufficient for tumorigenesis and that additional, acquired somatic changes are necessary. Thus, we propose to use *p53*-LFS as a model for identifying these changes, particularly those involved in sarcomagenesis.

6.2 Sarcomas are incredibly diverse

The rarity, heterogeneous composition, and numerous subtypes of sarcomas have complicated elucidation of genetic risk factors and driver mutations. A recent study implicates *ATM* and *ATR* as risk factors (16), but lacked the statistical power to definitively determine one way or the other. Moreover, sequencing done by the TCGA has shown that different sarcoma subtypes tend to have distinct mutational profiles across the genome, methylome, transcriptome, and proteome (10), implying sample selection is critical to achieve statistical power. Thus, researchers are faced with a conundrum – sequence many more disparate sarcomas, and try to work around confounding from multiple subtypes, or attempt to acquire additional sarcomas of more similar subtypes, of which samples may be limited, even in large consortiums.

6.3 Benefits and Disadvantages of working with Human-LFS sarcomas

Under LFS, sarcomas are a fairly common cancer, representing about 25% of all tumors (13% STS and 10% OS, Figure 6) (53), suggesting perhaps that sarcomagenesis is largely driven by *p53.* However, *p53* alterations are present in only about half of all sporadic sarcomas (10), and data from cBio seems to suggest that even this 50% number may be somewhat of an overestimate (depending on ascertainment) (Figure 36).

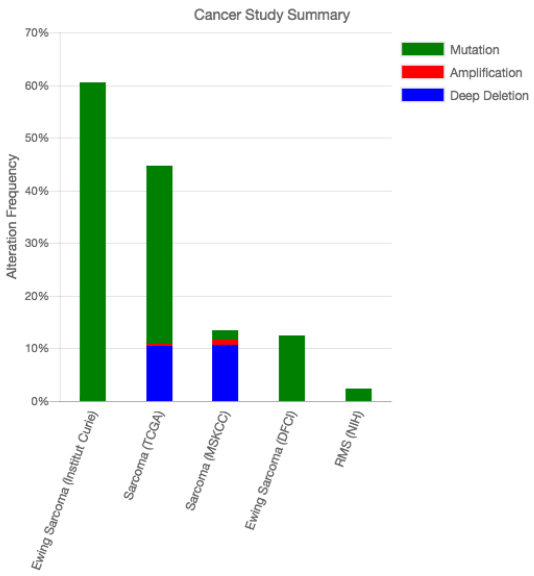


Figure 36: Five sarcoma studies have been recorded in cBio, covering bone sarcomas (Ewing), and soft tissue sarcomas (mostly liposarcomas and leiomyosarcomas). From left to right, these comprise 249, 265, 207, 107, and 43 samples. These data suggest that alterations in *p53* are not particularly common in sarcomas, topping out at ~60%.

We also leveraged cBio to check for tumors with alterations in any of *MDM2*, *MDM4*, and *TP53* with about 45% of sarcomas having an alteration in at least one of these genes, including a high percentage with *MDM2* amplification (Figure 37). Taken together, and considering the prevalence of sarcomas in LFS-patients/pedigrees, these data seem to indicate a role for *p53* in sarcomagenesis.

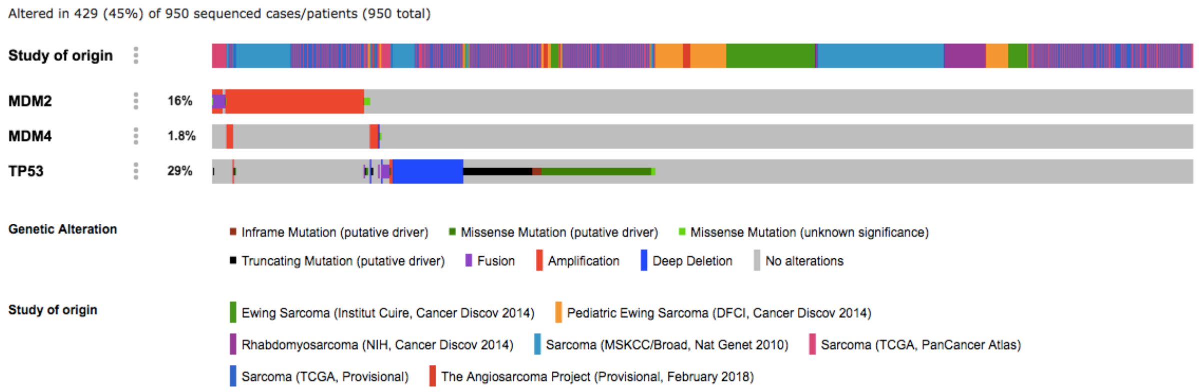


Figure 37: cBio Oncoprint plot showing distribution of MDM2/MDM4/TP53 alterations in sarcomas, covering 45% of tumors.

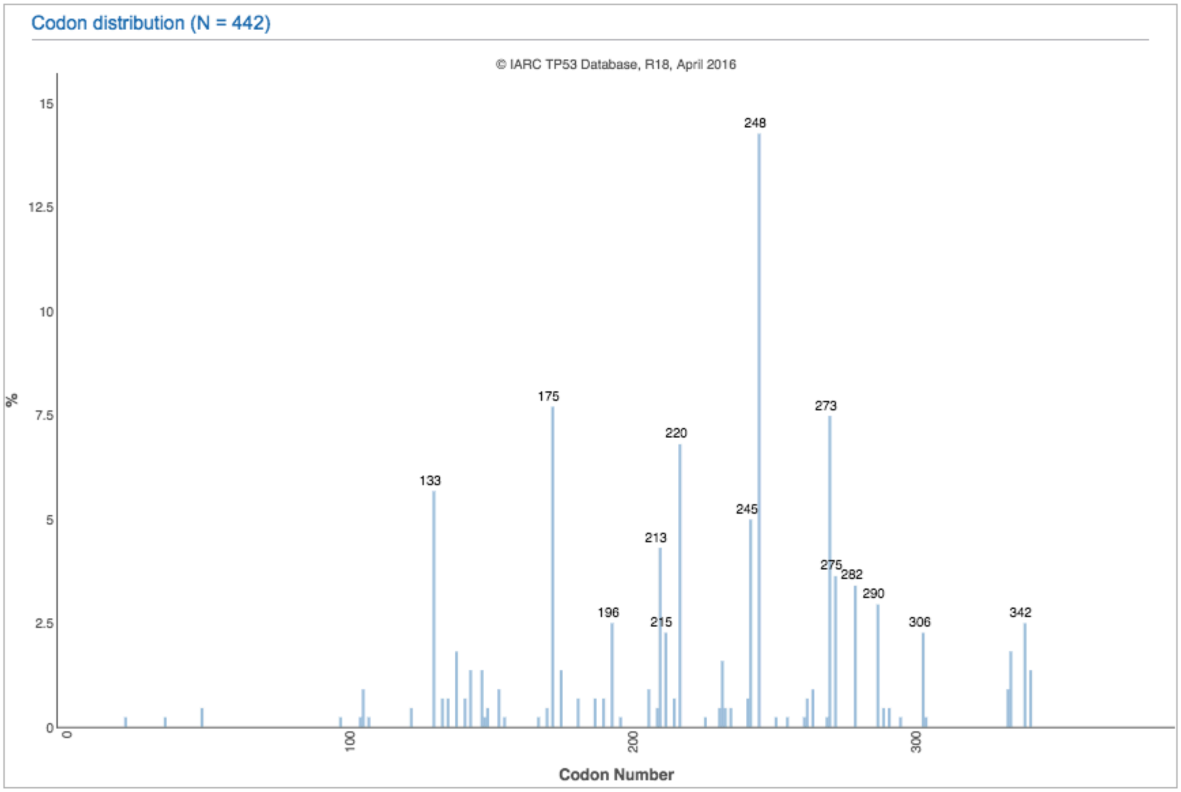


Figure 38: IARC data. (89) **Histogram of *p53* mutations that have reported families meeting the classical criteria for LFS (n=442). Of these, mutations in codon 248 (~14%) are the most common, followed by mutations in codon 175 (~8%), analogous to codon 172 in the mouse.**

Working with LFS-sarcomas may harbor one distinct advantage over that of sporadic sarcomas alone; they all involve disruption of the tumor suppressor gene *p53* increasing the likelihood that these sarcomas arise out of a *p53*-dependent manner. However, aside from rarity, it is not clear if all *p53* alterations should be treated equivalently – indeed different *p53* alleles have different penetrance (93). It remains to be seen if tumor etiology is similar for mutations in *p53* that are LOF (either by deletion, or point mutation) vs. GOF *p53* mutations. Despite these potential issues, it is our long-term hope that identifying genes associated with sarcomagenesis in LFS patients, will also have relevance to sporadic sarcomas, and other cancer types.

6.4 A Mouse Model of LFS May Help Identify Important Sarcomagenesis Genes

To overcome some of the limitations associated with sequencing only human tumors, we elected to pursue and take advantage of a mouse model of LFS. Several mouse models exist, including *p53*$^{-/-}$ mice in which cancer penetrance is 100%, with most cancers occurring before 6 months. These mice had tumor profiles similar to that of LFS-patients (109). However, since the majority of LFS patients have heterozygous point mutations, we instead elected to pursue a different mouse model. This model, from Dr. Lozano, contains the *Trp53 R172H* mutation that is analogous to the *TP53 R175H* hotspot mutation in humans ((51) Of the *p53* mutations associated with classical LFS, the *TP53 R175H* mutation is the second-most common (Figure 38), and is known to be a GOF mutation). This mouse model was also known to develop sarcomas (53%), particularly osteosarcomas (28%, most frequent tumor type reported) (51).

The overall relevance of modeling the human *TP53 R175H* mutation can be further emphasized with the use of a downloaded copy of the IARC database (53) for additional, more granular analysis. These data include a "topography" column, which contains the site of the original neoplasm, as well as the type and/or codon of the *p53* mutation. We can generate similar histograms as to Figure 38, except selected for tumors arising out of "soft tissue" and "bone" – the majority of which are sarcomas, to see if germline R175H mutations are prominent in these tumors. Although the overall mutation profile is different than compared to all tumors (Figure 38), the histograms in Figure 39 indicate that alterations at codon 175 are still prominent in both osteosarcomas (Figure 39A) and soft tissue sarcomas (Figure 39B).

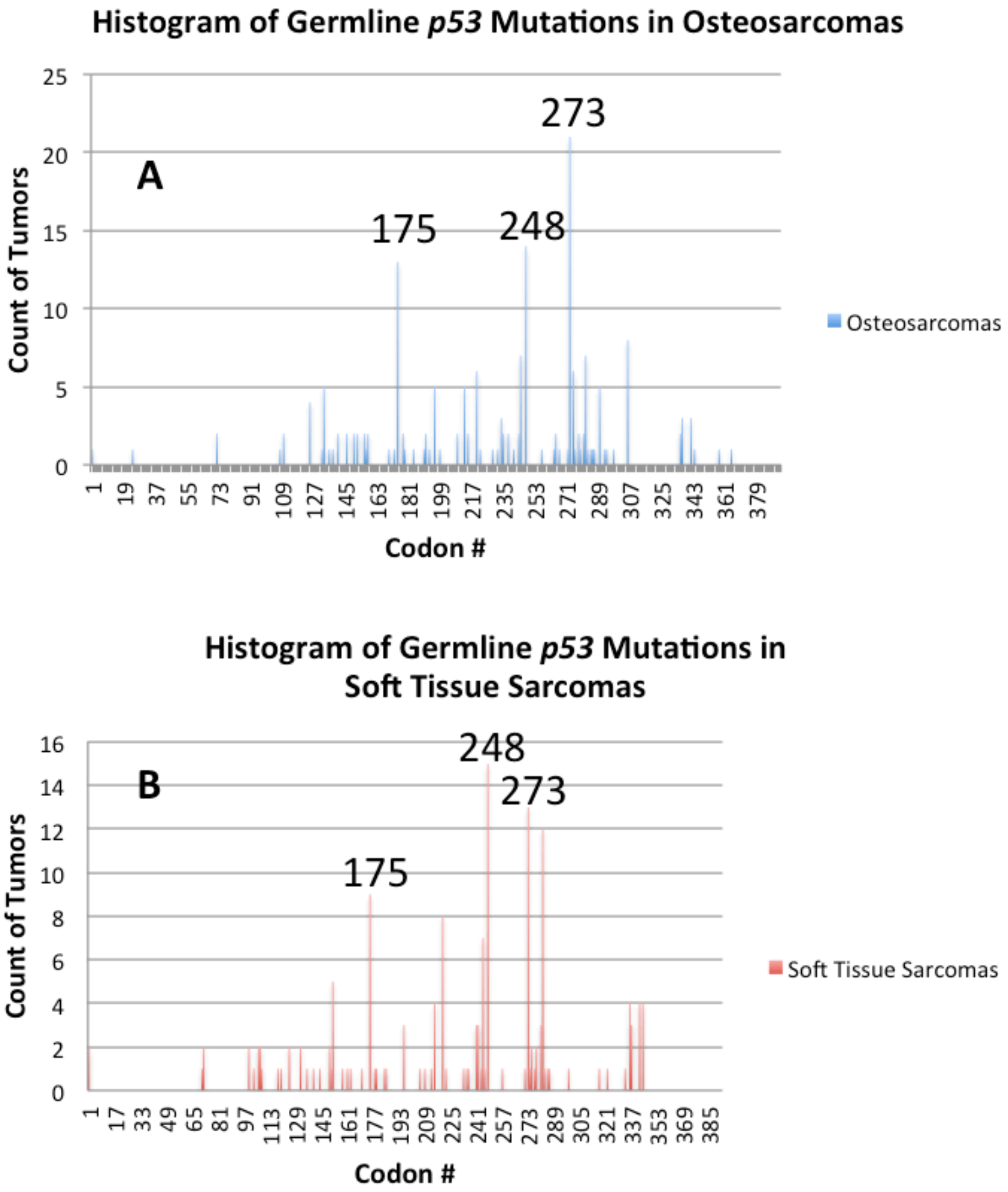


Figure 39: Histogram of *p53* mutations based on IARC data (53) and split by tumor pathology (soft tissue and bone).

6.4.1 Comparative approaches to identifying sarcoma drivers

One accepted approach to identifying drivers is to look for recurrent changes, either through gains or losses, due to gross amplifications or deletions, or SNVs in critical genes. These ideas underlie several efforts, including COSMIC, and TCGA to catalog somatic variation across a wide-range of tumors (5, 110(111).

However, recurrence alone may be a poor metric. Some genes or genomic regions may be, or appear to be, hypermutable, but may have little functional consequence. For example, olfactory genes and large muscle genes (e.g. *TTN*) often appear to be represented in lists of significantly recurrently mutated genes in cancer (112, 113). Respectively, these gene families are either not expressed in tumors, or due to their size are likely to have acquired somatic changes by chance alone, suggesting that recurrence alone is insufficient or that it may misclassify some genes. Thus, separating out drivers from passengers remains a key challenge.

Therefore, to better assist in separating passengers from drivers, we used a comparative genomics approach, leveraging a mouse model of LFS (6.4). We hypothesized that recurrent alterations in key genes and pathways, across both humans and mice, were more likely to be functionally important drivers of sarcomagenesis.

Using the mouse model confers several advantages. First, it provides an avenue to more readily generate LFS-associated sarcomas. Secondly, the mice will all have been raised in the same, shared environment, the mouse facility. Thirdly, the mice will come from reasonably similar backgrounds, and have the same exact cancer predisposition mutation, potentially improving the ability to identify recurrent changes. Fourthly, by using humans and mice, it provides additional context for recurrent variation such as CNVs that arise out of species-specific genomic context. Moreover, this comparative approach was previously successful in acute promyelocytic leukemia, where authors also used a mouse model to help implicate *JAK1 V657F* as being important in disease progression (114).

6.5 Complete tumor profiling as a goal

Although many sequencing studies have focused on looking at genomic information via WGS or WES, it is clear that other changes, such as those in the transcriptome or methylome may significantly contribute to cancer (5, 6, 115). In fact, these data can support each other when done in concert, to provide a more layered, nuanced picture of the tumor by testing, such as by testing to see if mutations are being expressed (114). In addition, fusion proteins have been implicated in several types of sarcomas, such as Ewing sarcomas (116), suggesting that sequencing DNA alone may miss key drivers. Transcriptomic sequencing may be particularly key for identifying these fusion proteins.

6.6 Open-ended Expectations for Sequencing Sarcomas

For most cancers the number of driver mutations is thought to be between 1-10 (19, 54), occurring in a subset of reasonably predictable genes (e.g. *TP53*, *NF1, RB1, PTEN, BRCA1, BRAF,* and so on). However, there is considerable ambiguity regarding the expected mutational profile of LFS sarcomas. Many LFS-associated soft sarcomas appear to exhibit chromothripsis and CNAs, particularly soft tissue sarcomas (10, 15), and exome sequencing of osteosarcomas also reveals CNAs to be common (117, 118). On the other hand, several genes with SNVs that may impact osteosarcomagenesis have been identified (117, 118).

Other papers have identified point mutations and deletions in genes such as *NF1, ERCC2, and PTEN* that should be associated with sarcomas (119, 120), further cementing the need for a multi-omics approach.

However, in many *p53*-associated non-sarcoma tumors, we observe a mutator phenotype, related to *p53*'s function as the guardian of the genome, in which these tumors appear to acquire a plethora of variants, some of which

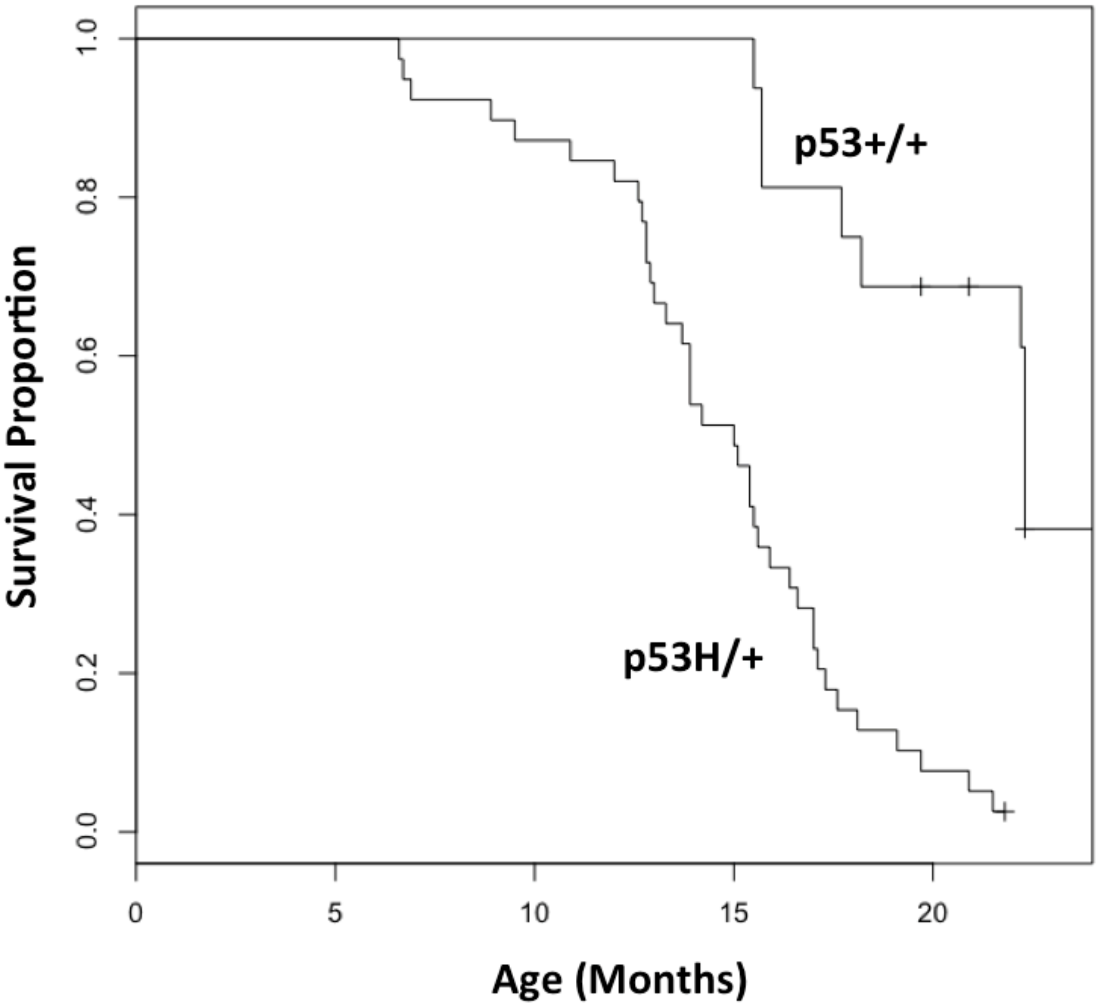


Figure 40: Tumor-free survival curve of *p53+/H (n=50) and p53+/+* (n=20) mice from our cohort, showing that that *p53H/+* mice were more prone to develop tumors (p-value: 6.37-e7). Some *p53+/+* mice developed lymphomas, particularly in the thymus.

may only be passengers (121, 122). Thus it may be unclear what we should expect when looking at *p53*-LFS sarcomas, chromothripsis, characteristic of sarcomas, or a mutator phenotype consistent with *p53* mutations, or some mixture of the two.

7 Results

7.1 Description of Mouse Cohort Results

Our original cohort consisted of 50 mice with the *p53 R172H* mutation, and 20 mice that were *p53 WT*. These mice were followed for a period of up to two years, with euthanasia performed for tumors, or other health conditions as needed. We collected multiple tissues from every mouse for pathology (including vastus lateralis, duodenum, pectoral and stomach muscles, diaphragm, kidney, spleen, heart, lung, liver, femur, and spine), thus allowing us to interrogate tumors that may not have been obvious to gross observation.

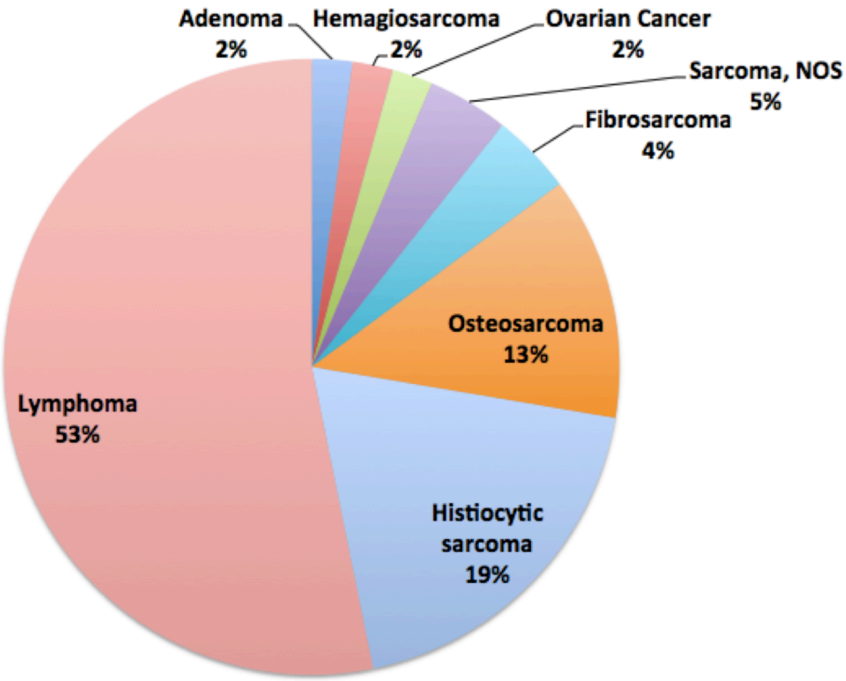


Figure 41: Pie graph of tumor distribution in the *p53 R172H* cohort (n=50) shows a high prevalence of lymphoma, with a variety of sarcoma types.

Overall, mice with the LFS cancer-predisposition did develop tumors and had reduced survival relative to control mice. However, we did observe several *p53 WT* mice that had to be sacrificed due to hyperpnea. In all such cases, mice were visibly hunched and had lymphoma in the thymus upon dissection, suggesting that the *p53 WT* C57BL/6 mice might have a predilection towards lymphomas, consistent with published results by Donehower et al. (123).

7.1.1 Lymphomas and sarcomas were the most common tumor types

We sent all our tissues to Dr. Elizabeth Whitley (Pathogenesis, LCC) for full pathology. The majority of tumors that we observed in our mice were lymphoma (including in some *p53*-WT mice), or sarcoma, with the majority of these being osteosarcomas and histiocytic sarcomas. Consistent with the notion that sarcoma types could have different etiologies, we attempted to minimize the types of sarcomas we observed during sample selection.

7.1.2 Choice of sarcoma type for sequencing

The most common sarcoma type we observed in the mice was histiocytic sarcomas, a neoplasia composed of hematopoietic cells, that has historically been classified as lymphoma, and is typically ascertained only after extensive immunophenotypic characterization (124), making it a less compelling candidate. In addition, histiocytic sarcomas are not considered to be among the most common tumor types associated with LFS in humans; in fact, in the IARC *TP53* database, there are no germline *p53* variants associated with histiocytic sarcomas. Therefore, we instead chose to sequence osteosarcomas and fibrosarcomas, leveraging additional samples that were previously generated in the Krahe Lab with a *p53 R172H* mutation. In total, these comprised two additional osteosarcomas, and four additional fibrosarcomas, for eight osteosarcomas, and six fibrosarcomas in total for omics analysis.

7.1.3 List of mouse sarcomas for sequencing

The table below (Table 10) contains a list of samples, their ages, gender, types of tissue sequenced, and whether or not the mouse was also diagnosed with lymphoma.

Table 10: Table of Mouse Sarcomas for NGS. VL – vastus lateralis, FR – femur, FS – fibrosarcoma, OS – osteosarcoma. Several mice were found to also have lymphoma in addition to sarcomas – these are noted in the table.

Sample	Age (mo)	Sex	Normal Type	Tumor Type	Lymphoma
F8-23	13.0	M	VL	FS	Yes
F1-13	11.8	M	VL	FS	No
F2-3	9.1	M	VL	FS	No
F8-17	14.2	M	VL	FS	No
F3-8	15.5	F	VL	FS	No
F2-4	16.1	M	VL	FS	Yes
F8-70	10.9	F	FR	OS	No
F8-51	13.4	F	FR	OS	Yes
F3-20	14.3	F	FR	OS	No
F7-12	17.3	F	FR	OS	Yes
F3-98	16.5	F	FR	OS	No
F8-49	13.0	M	FR	OS	Yes
F7-2	19.1	F	FR	OS	Yes
F3-35	13.5	F	FR	OS	No

These data show one remarkable trend. The majority of fibrosarcomas were found in male mice (5:1 male to female), while the majority of osteosarcomas were found in female mice (7:1 female to male). We have no reason to believe that this is anything more than a statistical anomaly, but may be worth continuing to monitor in future cohorts. We otherwise found no distinct trends for age and/or lymphomas

7.2 Mouse-specific variation necessitates the use of N/T pairs

Initially, we had hoped to use a pool of normal mice to serve as control tissues, under the assumption that the mice were isogenic. However, during initial analysis of the normal pool, we found these mice had pockets of variation that were typically unique to a given mouse, and could be due to being on a mixed background. These patterns were visible grossly, at the chromosome level when looking at variant density plots, and by a background check of the mouse.

When generating mouse models, there is risk for introducing additional variation during the process, particularly when using embryonic stem cells (ESC) for targeting (125). In this case, especially in early generations, the flanking regions around the mutation (or gene) often come from the ESC, rather than the original mouse strain. This flanking region can be reduced through backcrossing, but even after 10 successive backcrosses, as much as 1 cM on either side of the target (~40 genes on average) is likely to retain donor sequence(125) (125), leading to the presence of passenger mutations around the region of interest, particularly since variation is generally called against the background of the original mouse strain. Thus, it is possible, if unlikely, that germline variation in these flanking regions may contribute to the outcome, and strictly considering only somatic variation may be somewhat of an oversimplification when trying to uncover genetic contributions towards sarcomagenesis.

In fact, in sequencing normal controls from our *p53* mice, we seemed to observe this phenomenon. When we looked at overall variant density profiles for some mice in our cohort (called against the reference, a C57BL/6 mouse), we found variation in the flanking regions on chr11 (Figure 42) that was present in mice with a *p53*$^{+/H}$ background (red, blue), but not WT-controls with wild-type *p53* (orange, yellow, purple).

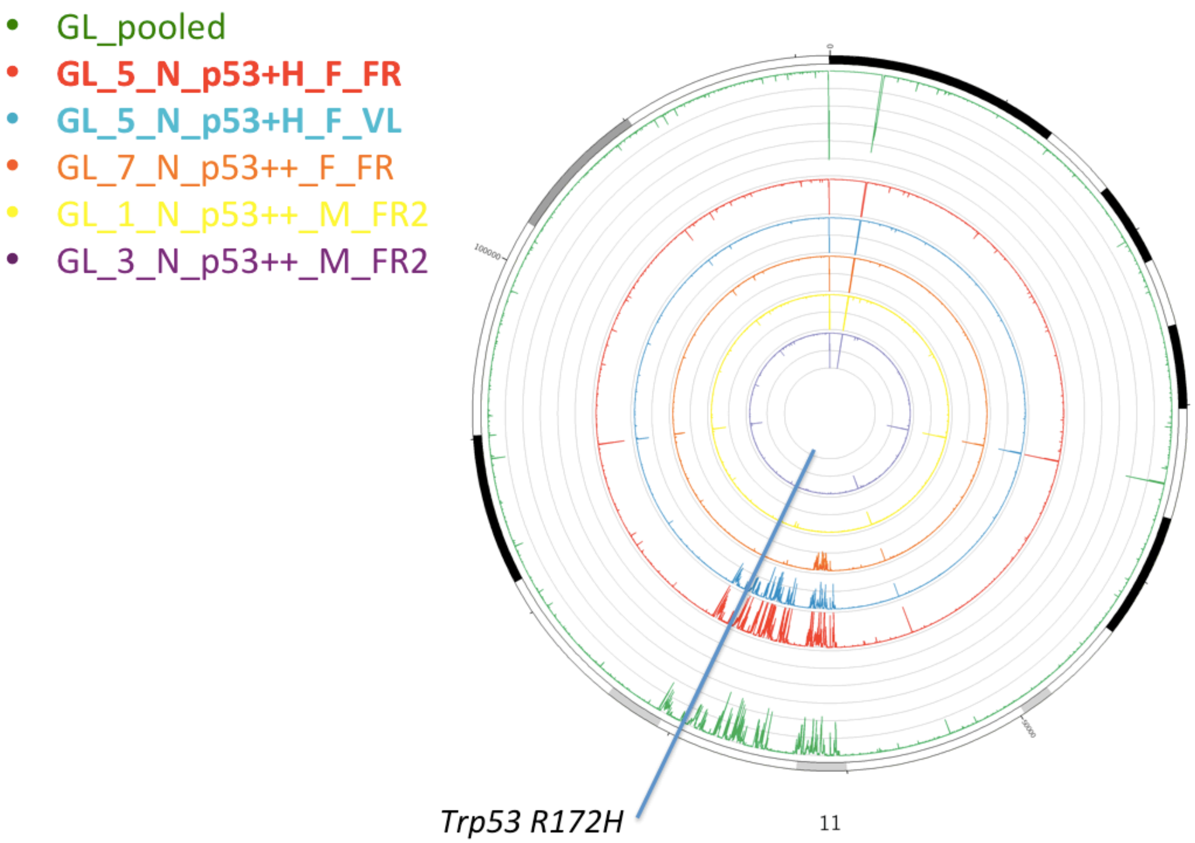
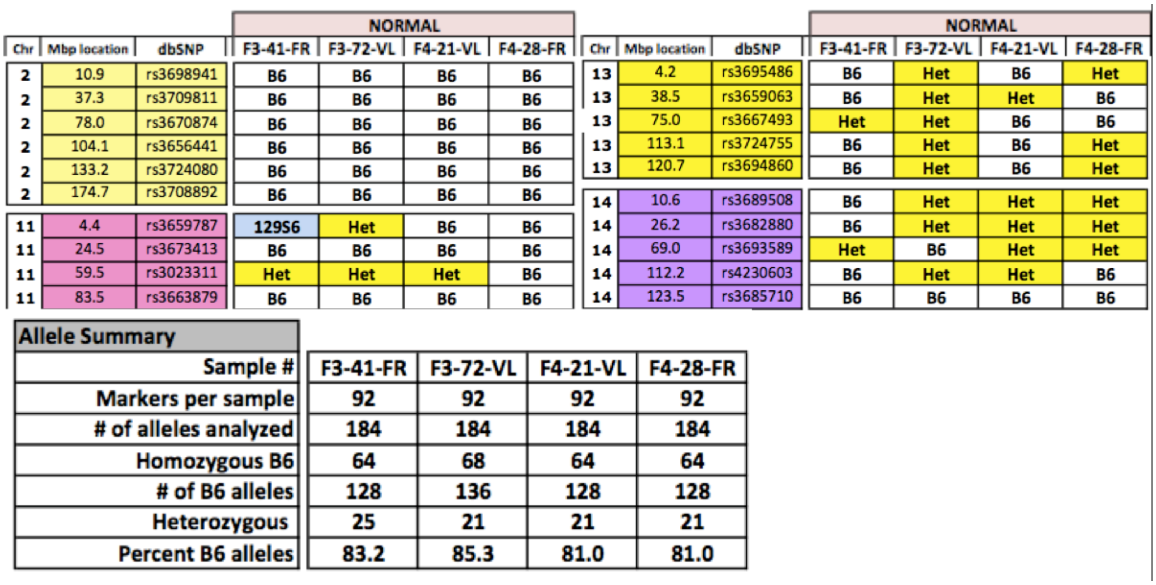


Figure 42: Variant density plot (# of variants/100kb) from a pilot study designed to determine if using a normal pool was reasonable under the hypothesis that the mice were sufficiently isogenic. We sequenced two tissues from a *p53*$^{+/H}$ mouse, and three femurs from three other mice with *WT p53*. These data show distinctly different variant profiles for the *p53*$^{H/+}$ mice vs. the *p53 WT* mice, and even hints that the *p53 WT* mice may not be isogenic with each other

7.2.1 Multiple chromosomes appear to have non-C57BL6 background

Moreover, we found potentially confounding background from the 129-mouse in other chromosomes as well. We worked with Dr. Benavides (MDACC – Smithville, TX; Research Animal Support Facility – Smithville), to characterize the mouse background using strain-specific SNPs. A subset of chromosomes in the mice we tested had no 129-alleles, but a subset of chromosomes, such as 13, and 14 had fairly extensive contamination. Overall, the data seemed to indicate that just over 80% of the alleles were from the C57BL/6 background.



			NORMAL			
Chr	Mbp location	dbSNP	F3-41-FR	F3-72-VL	F4-21-VL	F4-28-FR
2	10.9	rs3698941	B6	B6	B6	B6
2	37.3	rs3709811	B6	B6	B6	B6
2	78.0	rs3670874	B6	B6	B6	B6
2	104.1	rs3656441	B6	B6	B6	B6
2	133.2	rs3724080	B6	B6	B6	B6
2	174.7	rs3708892	B6	B6	B6	B6
11	4.4	rs3659787	129S6	Het	B6	B6
11	24.5	rs3673413	B6	B6	B6	B6
11	59.5	rs3023311	Het	Het	Het	B6
11	83.5	rs3663879	B6	B6	B6	B6

			NORMAL			
Chr	Mbp location	dbSNP	F3-41-FR	F3-72-VL	F4-21-VL	F4-28-FR
13	4.2	rs3695486	B6	Het	B6	Het
13	38.5	rs3659063	B6	Het	Het	B6
13	75.0	rs3667493	Het	Het	B6	B6
13	113.1	rs3724755	B6	Het	B6	Het
13	120.7	rs3694860	B6	Het	B6	Het
14	10.6	rs3689508	B6	Het	Het	Het
14	26.2	rs3682880	B6	Het	Het	Het
14	69.0	rs3693589	Het	B6	Het	Het
14	112.2	rs4230603	B6	Het	Het	B6
14	123.5	rs3685710	B6	B6	B6	B6

Allele Summary				
Sample #	F3-41-FR	F3-72-VL	F4-21-VL	F4-28-FR
Markers per sample	92	92	92	92
# of alleles analyzed	184	184	184	184
Homozygous B6	64	68	64	64
# of B6 alleles	128	136	128	128
Heterozygous	25	21	21	21
Percent B6 alleles	83.2	85.3	81.0	81.0

Figure 43: SNP data checking the background of the *p53 R172H* mice showing that although the majority of the alleles were from the C57BL/6 background (>80%), there were some chromosomes that had extensive heterozygous alleles. The data also strongly suggest that these mice should not be considered isogenic.

Consistent with these data, we also looked at variant density plots (line graph of variants/100 kb) across other chromosomes across five tissues from different mice (Figure 44), finding that mice appeared to be less isogenic than originally expected. Several mice had unique variation/variant profiles, such as “GL_1_N…” (grey) on chromosomes 1, 7, 11, and 13, or “GL_7_N…” (orange) on chromosome 11.

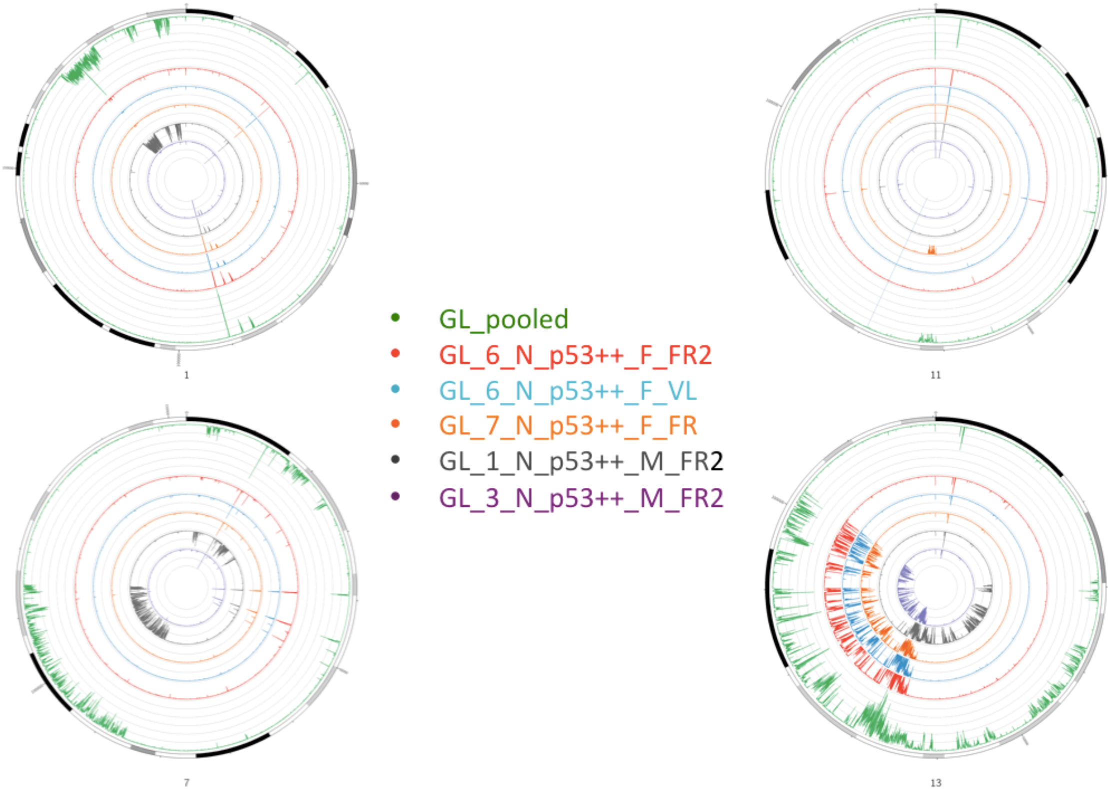


Figure 44: Variant density plots (# variants/100 kb windows), continue to support that these mice are should not be considered isogenic.

These data support the previous evidence from the background check that these mice were not isogenic, and necessitated the switch from a normal pool to only sequencing matched normal/tumor pairs to better ascertain somatic risk factors.

For fibrosarcomas, we elected to use the vastus lateralis (VL, muscle) and for osteosarcomas (OS), we used a femur (FR, bone) as a normal control. We expected that such considerations matter less for the genomic analysis, but more for transcriptomic or epigenetic analyses where tissue specificity matters more (126). In all cases, we chose to sequence normal tissue that was distal to the tumor to better delineate acquired somatic changes.

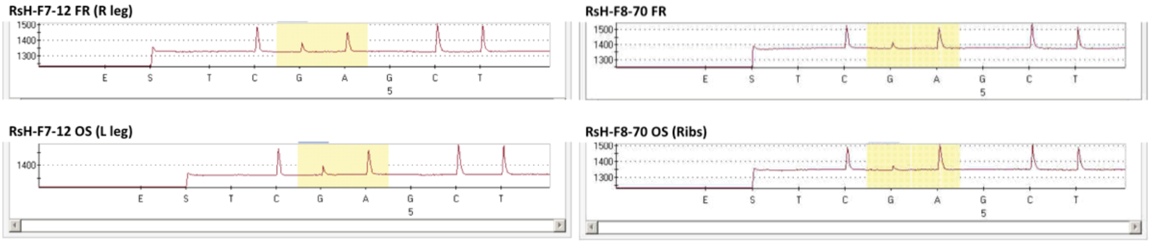
7.3 LOH and Pyrosequencing

7.3.1 Initial WGS data suggests not all sarcomas have LOH

The general consensus in the field is that the majority of *p53*-tumors have eventual LOH (127). However, in our pilot study, in which we sequenced 2 human tumors, and 2 mouse tumors by WGS, we observed one tumor with LOH and one with ROH for each organism. Alerted to the possibility that LOH might not be as common as expected, we used pyrosequencing to quantitatively determine the extent of LOH in our mouse tumors and found two initially unexpected results. First, we saw LOH in some normal tissues. Secondly, in contrast to the general consensus described in Rivlin et al. (127) of the majority of tumors having eventual, we saw LOH in *p53* in mouse sarcomas only about half the time. However, we acknowledge that previously published LOH data by Lang et al. is also around 50% (51).

7.3.2 Some normal tissues appear to have LOH

In addition to pyrosequencing tumor tissues, we also sequenced the matching constitutive tissues from the same mouse as controls, finding two outliers. We found two femurs that appeared to be normal at gross observation, but had relatively advanced LOH by pyrosequencing (Figure 45). In both cases, LOH was more pronounced in the tumor, relative to the supposedly normal femur. Moreover, these femurs were not adjacent to the osteosarcoma. One osteosarcoma was located in the ribs, while the other osteosarcoma was in the left leg (with the femur being from the right leg).



	Average Pyrogram Peak Heights (%)	
	G	A
N	32.8	67.2
T	27.9	72.1

	Average Pyrogram Peak Heights (%)	
	G	A
N	25.6	74.4
T	14.7	85.3

Figure 45: Pyrograms with peak heights. These data suggest that we had LOH in what appeared to be normal femurs upon gross observation. Moreover, these femurs were distal from the primary tumor site, suggesting that LOH may not be sufficient for tumorigenesis.

To rule out the possibility that all constitutive tissues had this same behavior, we pyrosequenced additional tissues from the same mouse, including the pectoralis muscle. These other tissues all had ROH, suggesting that the LOH was relatively unique to the femur (Figure 46).

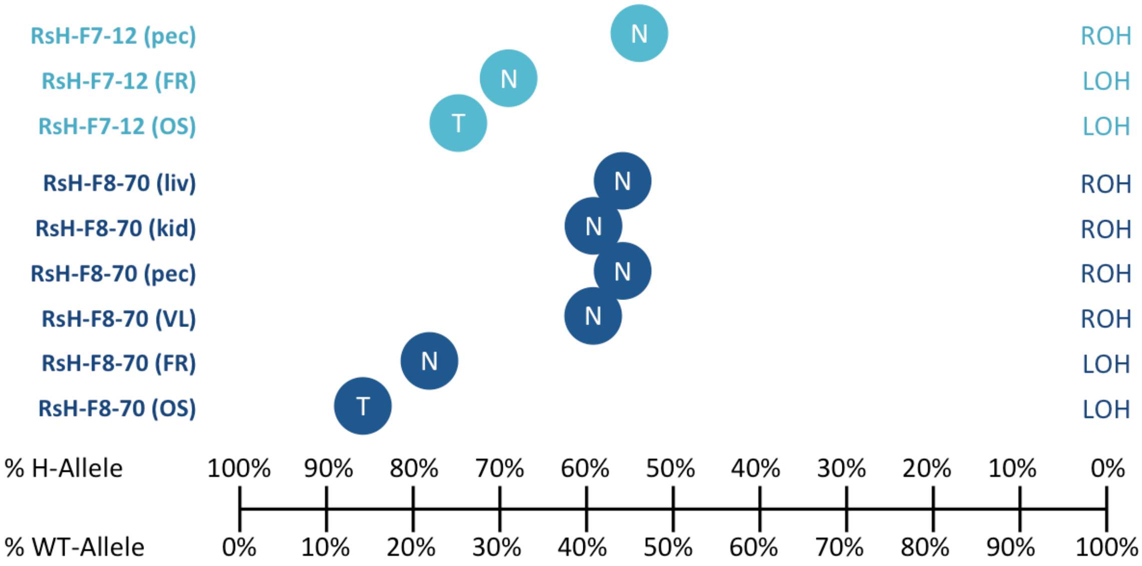


Figure 46: Graphical depiction of pyrosequencing data of additional constitutive tissues in mice with LOH in the femur. On the x-axis is a number-scale representing the %H-allele, with circles trending towards the left side of the graph having greater LOH. Circles near 50% have ROH (i.e. have retained the WT-allele). These data suggest the majority of normal (N) samples have ROH, with the exception of these two FR samples, which show relatively advanced LOH, though not exceeding the tumors.

These data are consistent with the idea that *p53 LOH* at the *Trp53 R172H* locus is not sufficient for tumorigenesis. However, because we typically send one femur for pathology, and reserve the other for experiments, such DNA extractions, it is possible that the sequenced femur had cancerous or pre-cancerous lesions that were not visible at gross observation. We did not observe such a pattern to be the case for any VL/FS pairs; no VL samples had LOH.

These data are juxtaposed against *p53*$^{-/-}$ and *p53*$^{+/-}$ mice. In *p53*$^{-/-}$ null mice, the penetrance of tumors is 100%, suggesting that LOH alone would be sufficient for tumorigenesis. In *p53*$^{+/-}$ mice, the data indicate *p53* should be considered to be haploinsufficient. Only half of all tumors in mice under the age of 18 months had LOH, and this number dropped to just 15% in tumors over 18 months. (128) However, no such comprehensive analysis of LOH has been done in *p53*$^{H/+}$ mice and it is not clear if, or how the GOF mutation may impact such behavior.

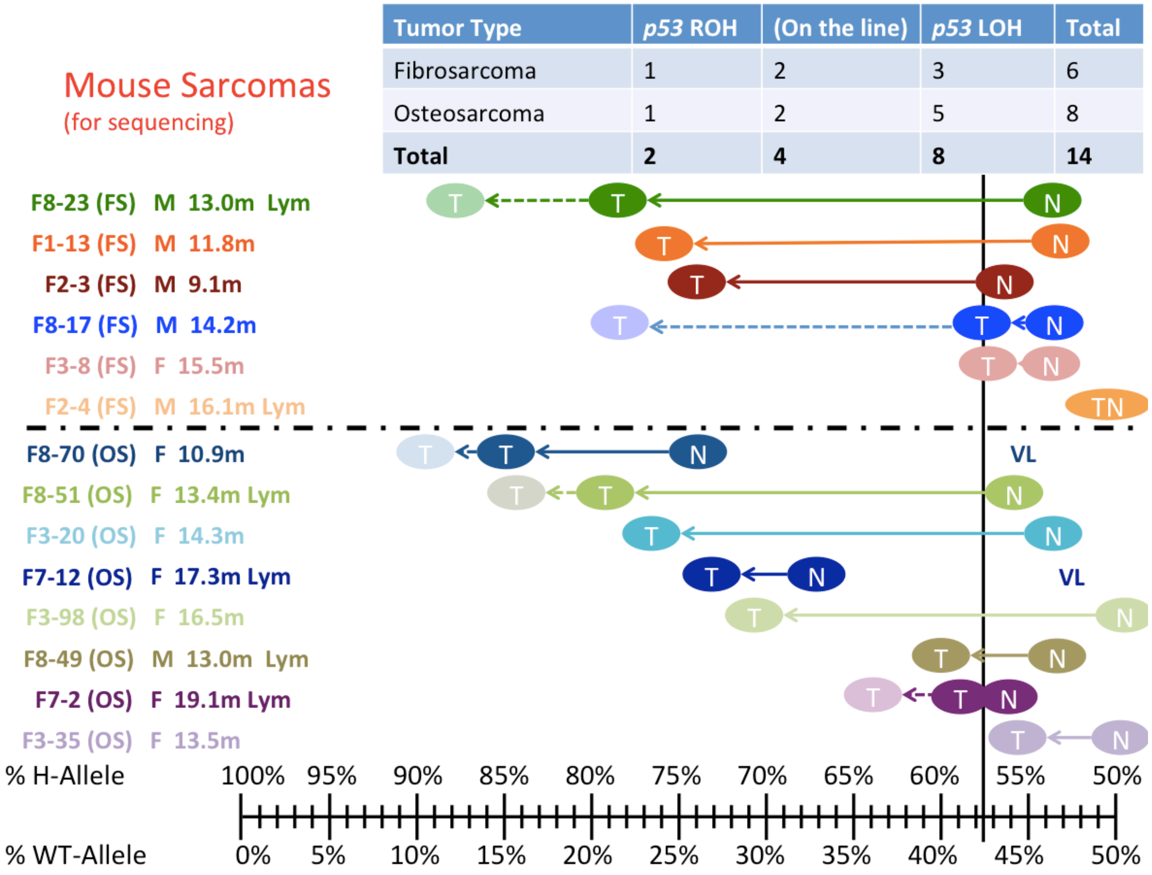
7.3.3 Use of toe tissue as normal controls

Due to observations that some tissues, which on gross observation appeared to be normal, had LOH, we wanted to find better control tissues to best determine a baseline that represented ROH. We elected to use toe tissue, originally collected for genotyping, between days 7-10 when it is still cartilaginous. Theoretically, these tissues should have had minimal time to acquire additional somatic changes and therefore serve as a more appropriate benchmark than constitutive tissues collected at time of sacrifice.

Data from toe tissue collected across our mouse tumors suggests that the totality of the experiment and analysis has a slight bias for the mutant A-allele. Whether this occurs during the PCR amplification, or as part of the pyrosequencing process, which is known to favor the A-allele is unknowable. Using the toe data as a whole, we set up our own classification scheme. We averaged the allele % from all toes, and then calculated a standard deviation. Anything within two standard deviations from the average was considered ROH.

7.3.4 Pyrosequencing shows that LOH occurs in only about 50% of mouse LFS-sarcomas

Pyrosequencing of all tumors and matched normals suggested that LOH was present only in about half of all sarcomas (Figure 47), indicating the LOH (at the R172H locus) is not required for sarcomagenesis. These data are also somewhat consistent with previous data suggesting that LOH is less common in older tumors. Using an arbitrary cutoff point of 15 months, this provides a clear delineation in fibrosarcomas. For osteosarcomas the data is less clear. There are two osteosarcomas (F8-49 and F3-35) at about 13 months that did not appear to have LOH, and two older ones with more pronounced LOH (F7-12 and F3-98). We also did not observe any trends for LOH as it relates to lymphoma status.



Tumor Type	*p53* ROH	(On the line)	*p53* LOH	Total
Fibrosarcoma	1	2	3	6
Osteosarcoma	1	2	5	8
Total	**2**	**4**	**8**	**14**

Figure 47: Depiction of pyrosequencing results plotted against % of mutant and WT alleles. Mouse name, tumor type (fibrosarcoma – FS, osteosarcoma – OS) sex, age at sacrifice, and lymphoma status are listed at left-hand column. Data indicate where the matching normal is (always femur, or vastus lateralis for corresponding tumors), with a solid line indicating the extent of LOH in the tumor. Additional dashed lines account for tumor purity as indicated by the pathologist where available. The black, solid vertical line indicates ROH. Black dot-dash line bisecting graph in two separates out fibrosarcomas from osteosarcomas. Table at top provides summary based on tumor type and ROH/LOH status.

8 Results: Sequencing of Mouse Samples

We then looked at several different kinds of sequencing. For DNA, we looked at WES (at 40X/80X N/T coverage, Nimblegen 2.0, capture region 54.3 Mb) and low-coverage WGS. For RNA, we were unable to isolate high-quality RNA from fresh frozen tissues. Despite our best attempts to snap freeze, RNA appeared degraded after isolation. For methylation data, we used reduce representation bisulfite sequencing (Illumina).

8.1 WES variant calling pipeline

We did all WES and somatic variant calling in-house according to best practices as outlined by GATK. (80) We additionally used five different somatic variant callers, MuTect, (129) MuTect2, (80) VarScan2, (130) Somatic Sniper (131) and MuSE, (132) again using the default recommended settings. Historically, somatic variant callers have generally showed poor agreement with one another, and are continuing to undergo refinement. (133) (134) (135)

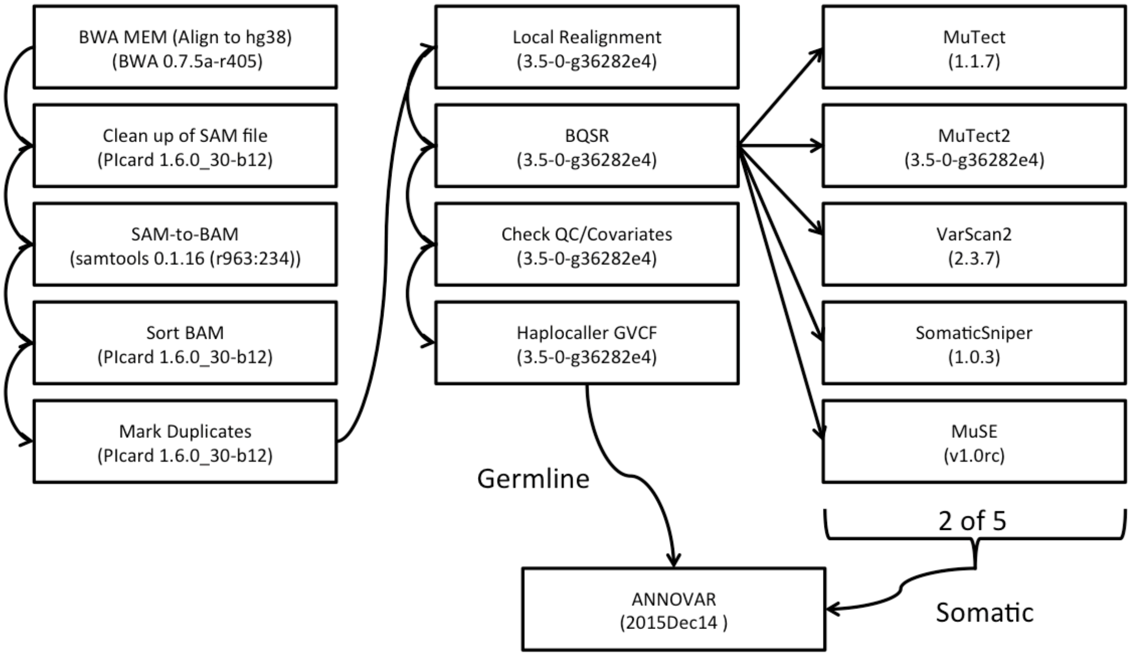


Figure 48: Krahe lab pipeline for identification of somatic variation in WES data. We found somatic variant callers to produce widely disparate calls – to compensate for this, we looked at all somatic variants that were found by at least two algorithms to best balance false positives and false negatives.

8.1.1 Filtering for high-quality somatic variants

In our pipeline, using the recommended best practices, and selecting only for variants each considered to be high confidence, we saw poor agreement between the samples when looking at somatic SNVs (Figure 49). A significant proportion of the lack of agreement seems to be due to the high number of variants found by both mutect1 and Varscan2, which call the most variants. Of all the variant callers, MuSE shows the greatest overlap with other algorithms.

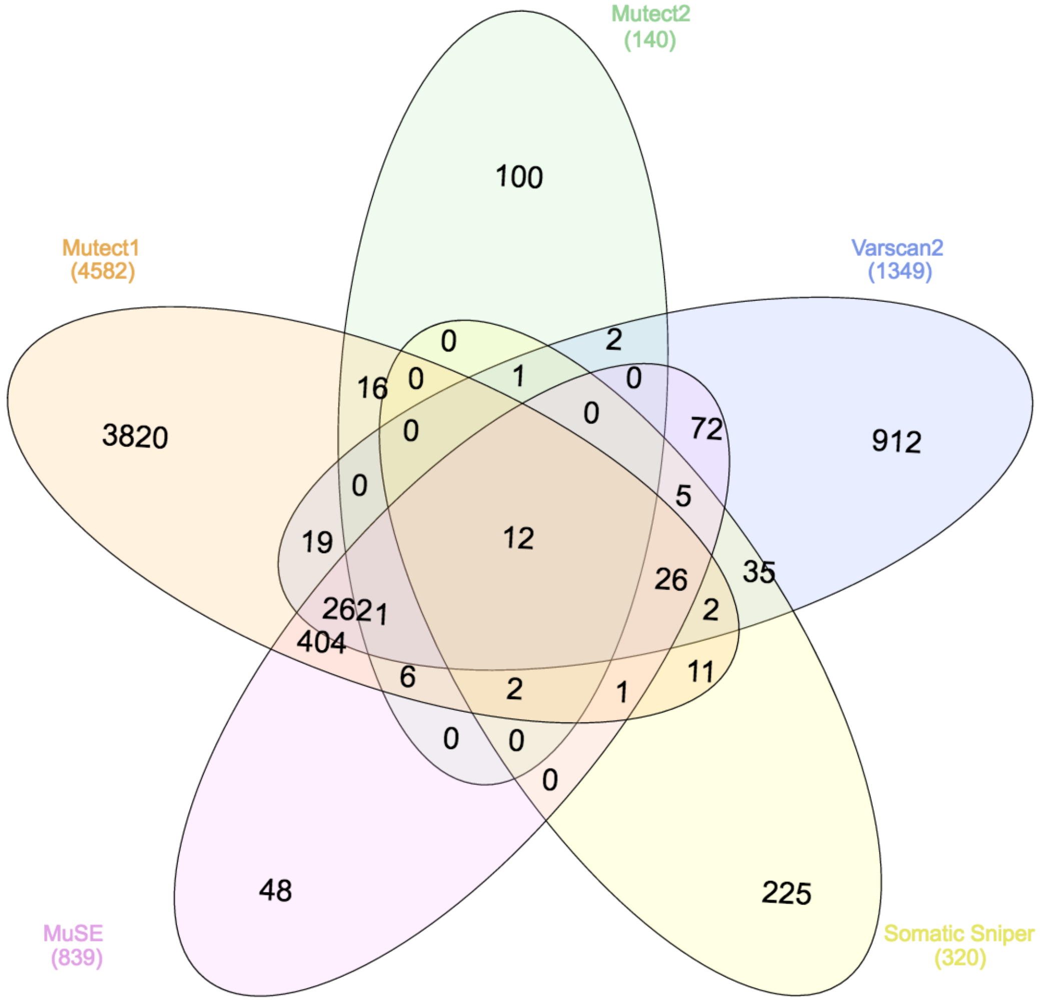
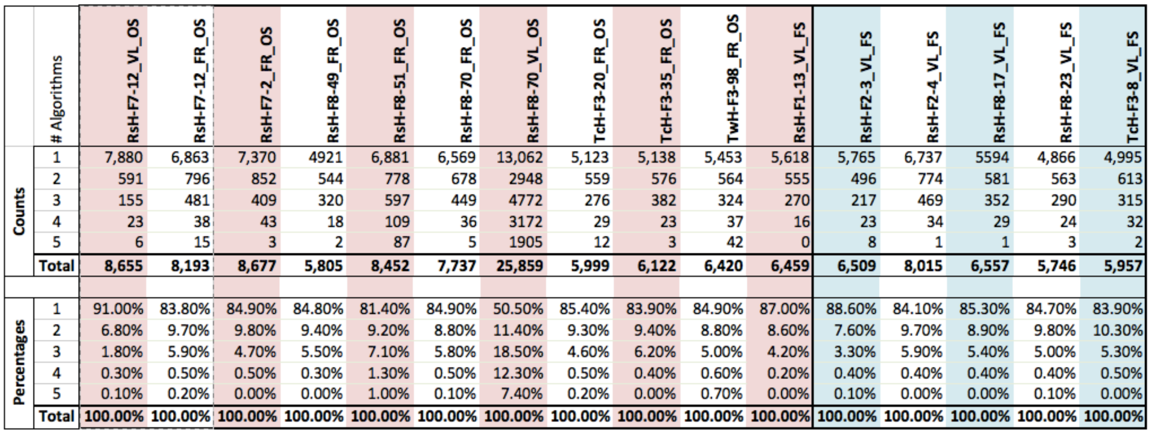


Figure 49: Representative Venn diagram of agreement between five different somatic variant callers in our mouse samples. All variants were called according to best practices as outlined in the documentation and were considered to be high confidence by the individual variant caller. Overall, the key data trends that we observe are that Mutect1 and Varscan2 call the most variants, and that MuSE shows the most agreement with other somatic variant callers.

The lack of agreement underlies a key question: which somatic variant calls are true positives? How should researchers balance retaining false positives, with true positives occurring at lower allele frequencies (e.g. arising by tissue heterogeneity). Consideration of variants found by *all* algorithms is almost certainly too restrictive – they represent, on average, just 0.1% of all somatic variant calls (by any algorithm). In contrast, consideration of variants found in *any one* algorithm may lead to spurious association. Brief visual inspection indicates that these one-algorithm mutations contain a mix of plausibly true and likely to be false variants.

To get a sense for an appropriate cutoff, we counted the number of variants found by exactly one algorithm, by exactly two algorithms (any pair was fine), by any three algorithms, and so on. On average, about 5% of total somatic variants are called are found by 3+ algorithms, and about 10-15% of total somatic variants are found by 2+ algorithms (Table 11).

Table 11: Table showing overlap of five different somatic variant callers. The top half counts the number found by exactly 1, 2, 3, 4, or 5 algorithms for 16 N/T pairs across 14 mice. The second half translates these into percentages.



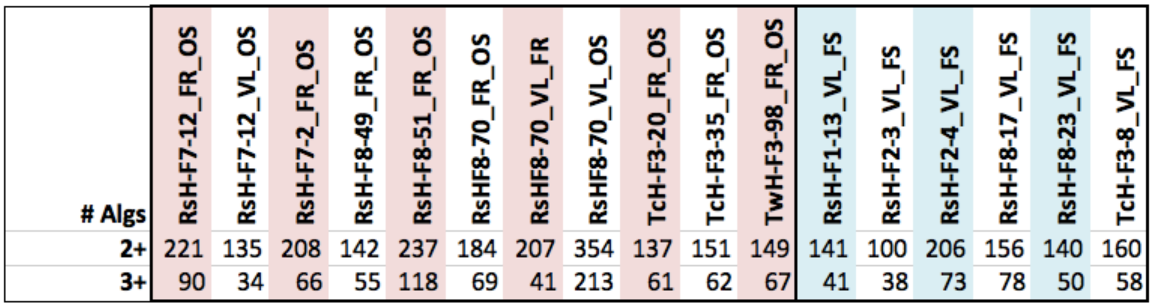
	# Algorithms	RsH-F7-12_VL_OS	RsH-F7-12_FR_OS	RsH-F7-2_FR_OS	RsH-F8-49_FR_OS	RsH-F8-51_FR_OS	RsH-F8-70_FR_OS	RsH-F8-70_VL_OS	TcH-F3-20_FR_OS	TcH-F3-35_FR_OS	TwH-F3-98_FR_OS	RsH-F1-13_VL_FS	RsH-F2-3_VL_FS	RsH-F2-4_VL_FS	RsH-F8-17_VL_FS	RsH-F8-23_VL_FS	TcH-F3-8_VL_FS
Counts	1	7,880	6,863	7,370	4921	6,881	6,569	13,062	5,123	5,138	5,453	5,618	5,765	6,737	5594	4,866	4,995
	2	591	796	852	544	778	678	2948	559	576	564	555	496	774	581	563	613
	3	155	481	409	320	597	449	4772	276	382	324	270	217	469	352	290	315
	4	23	38	43	18	109	36	3172	29	23	37	16	23	34	29	24	32
	5	6	15	3	2	87	5	1905	12	3	42	0	8	1	1	3	2
	Total	**8,655**	**8,193**	**8,677**	**5,805**	**8,452**	**7,737**	**25,859**	**5,999**	**6,122**	**6,420**	**6,459**	**6,509**	**8,015**	**6,557**	**5,746**	**5,957**
Percentages	1	91.00%	83.80%	84.90%	84.80%	81.40%	84.90%	50.50%	85.40%	83.90%	84.90%	87.00%	88.60%	84.10%	85.30%	84.70%	83.90%
	2	6.80%	9.70%	9.80%	9.40%	9.20%	8.80%	11.40%	9.30%	9.40%	8.80%	8.60%	7.60%	9.70%	8.90%	9.80%	10.30%
	3	1.80%	5.90%	4.70%	5.50%	7.10%	5.80%	18.50%	4.60%	6.20%	5.00%	4.20%	3.30%	5.90%	5.40%	5.00%	5.30%
	4	0.30%	0.50%	0.50%	0.30%	1.30%	0.50%	12.30%	0.50%	0.40%	0.60%	0.20%	0.40%	0.40%	0.40%	0.40%	0.50%
	5	0.10%	0.20%	0.00%	0.00%	1.00%	0.10%	7.40%	0.20%	0.00%	0.70%	0.00%	0.10%	0.00%	0.00%	0.10%	0.00%
	Total	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**

8.1.2 Somatic mutation burden in sporadic sarcomas

The most recent reported overall somatic mutation burden for human adult soft tissue sarcomas is, on average, 1.06 per Mb, with the highest reported being 33.5 mutations per Mb, where mutations are considered to be non-synonymous SNVs (10). Back referencing the data against the cBio portal reveals that the samples with the three highest mutation loads (>350 total mutations) have somatically acquired likely loss-of-function mutations in *p53*: TCGA-DX-AB2E-01 (R342*, age 53, myxofibrosarcoma), TCGA-3B-A9HT-01(*W91**, age 53, leiomyosarcoma), and TCGA-DX-AB32-01 (L252del, age 62, undifferentiated pleomorphic sarcoma), consistent with the possibility that *p53* LOF alterations can increase mutational load. Four patients had somatic mutations at *p53 R175H*, analogous to the mouse model; these had 36, 59, 62, and 101 mutations, at ages 44, 73, 64, and 62 respectively *(10, 22).*

We then took variants that had been identified by two or more algorithms, and annotated with ANNOVAR (136), SIFT, (86) and PROVEAN (137) to further select for non-synonymous variants to see if our mice were in line with these data (Table 12).

Table 12: Table showing the number of non-synonymous SNVs found in each mouse by two or more algorithms (2+), or three or more algorithms (3+). Results are roughly consistent with published data on mutation rates (by non-synonymous SNVs) in soft tissue sarcomas.



# Algs	RsH-F7-12_FR_OS	RsH-F7-12_VL_OS	RsH-F7-2_FR_OS	RsH-F8-49_FR_OS	RsH-F8-51_FR_OS	RsHF8-70_FR_OS	RsHF8-70_VL_FR	RsHF8-70_VL_OS	TcH-F3-20_FR_OS	TcH-F3-35_FR_OS	TwH-F3-98_FR_OS	RsH-F1-13_VL_FS	RsH-F2-3_VL_FS	RsH-F2-4_VL_FS	RsH-F8-17_VL_FS	RsH-F8-23_VL_FS	TcH-F3-8_VL_FS
2+	221	135	208	142	237	184	207	354	137	151	149	141	100	206	156	140	160
3+	90	34	66	55	118	69	41	213	61	62	67	41	38	73	78	50	58

At a capture region of 54.3 Mb in the mouse, if the same mutation rates hold in the mice, we would expect to see an average of 58 mutations. To estimate an upper bound, we took the highest mutational burden in an adult soft tissue sarcoma from the TCGA paper (10), 33.5 mutations per Mb. Applying this rate to the mouse exome (33.5 mutations per Mb * 54.3 Mb), gives an upper bound of 1,819 somatic mutations. Choosing either 2+ or 3+ is therefore roughly within the same order of magnitude.

Because we were unsure of how a germline GOF *p53* mutation in the mouse could contribute to somatic mutation rates, we decided that it was worthwhile to at least consider somatic variants identified by 2+ algorithms, especially when using recurrence as an endpoint.

8.2 Whole Exome Sequencing Results

8.2.1 Recurrence in Somatic SNV Data

We initially looked at recurrence of SNVs across all mouse tumors, finding one recurrent SNV, in *Mroh2a*. This variant (c.C4088T, p.T1363M) was predicted to be damaging (PROVEAN=-4.02, threshold < -2.5; SIFT=0.046, threshold < 0.05) but may be polymorphic; it has an rsID but no reported allele frequencies. Secondly, we looked to see if there were any genes that were recurrently mutated (using only variants predicted to be deleterious). There were nine genes that recurred across three or more tumors (Table 13).

Table 13: Table of recurrent somatically mutated genes in mouse tumors (three or more tumors). For a gene to be considered, it had to have a mutation predicted to have functional consequence. Shading indicates where two comparisons for the same tumor were made, one to a femur (FR), and one to a muscle (VL).

Gene	RsH-F7-12_FR_OS	RsH-F7-12_VL_OS	RsH-F7-2_FR_OS	RsH-F8-49_FR_OS	RsH-F8-51_FR_OS	RsHF8-70_FR_OS	RsHF8-70_VL_OS	TcH-F3-20_FR_OS	TcH-F3-35_FR_OS	TwH-F3-98_FR_OS	RsH-F1-13_VL_FS	RsH-F2-3_VL_FS	RsH-F2-4_VL_FS	RsH-F8-17_VL_FS	RsH-F8-23_VL_FS	TcH-F3-8_VL_FS	Grand total	Total tumors
Mroh2a	1	1					2	2			1				1	1	9	8
Ttn	1	1			1				1							2	6	4
Ahdc1				1		1	1						1			1	5	4
Arid1a	1	1				1	1							1			5	3
Syne2	1	1	1			1	1										5	3
Celsr3	1	1											1	1			4	3
Cherp			1	1								1				1	4	4
Phka1	1	1											1	1			4	3
Sspo			2										1	1			4	3

With the exception of *Arid1A*, these recurrently somatically mutated genes do not appear be well-known players in cancer. They are not found in any of our human-based gene lists (covering cancer, the p53 network, and sarcomas). Therefore, these genes may represent either good novel candidates to pursue further, or false-positives, arising perhaps out of multiple testing of the entire exome.

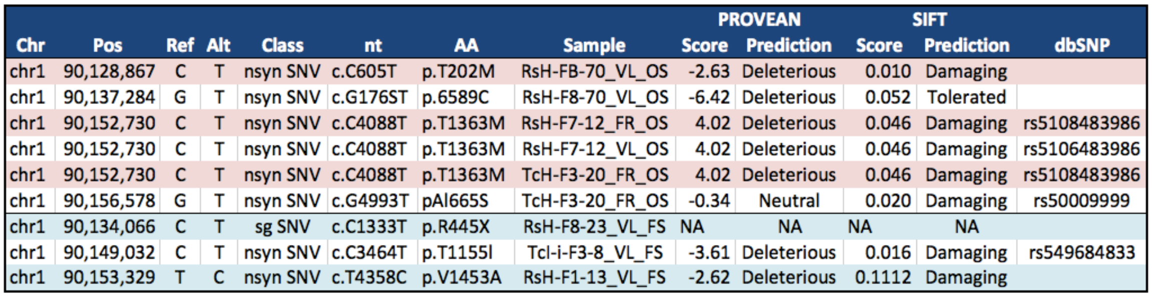
8.2.1.1 Arid1a

ARID1A (in humans) is known to be a haploinsufficient tumor suppressor and is a SWI/SNF chromatin remodeling gene that has been linked to several cancers, including ovarian, gastric, and breast tumors, but not including sarcomas. In fact, in one paper by Wu et al. the authors specifically claim they are not aware of any *ARID1A* mutations being detected in human sarcomas (138). Moreover, the majority of cancer-causing mutations in *ARID1A* that have so far been identified are stopgain or frameshift mutations (138), as opposed to the non-synonymous SNVs we identified in *Arid1a* in the mouse. Taken together, these data raise some doubt as to whether or not these mutations could impact sarcomagenesis, but do allow that *ARID1A* may be an intriguing, if not immediately compelling, candidate for follow-up.

8.2.1.2 Mroh2a

Mroh2a is the most common somatically mutated gene across our mouse tumors. These comprise seven non-synonymous variants, and one stop gain variant, some that have been flagged as potentially being polymorphic.

Table 14: Table of somatic variants in *Mroh2a* across all mice. Mice not listed here do not have somatic alterations predicted to be damaging.



								PROVEAN		SIFT		
Chr	Pos	Ref	Alt	Class	nt	AA	Sample	Score	Prediction	Score	Prediction	dbSNP
chr1	90,128,867	C	T	nsyn SNV	c.C605T	p.T202M	RsH-FB-70_VL_OS	-2.63	Deleterious	0.010	Damaging	
chr1	90,137,284	G	T	nsyn SNV	c.G176ST	p.6589C	RsH-F8-70_VL_OS	-6.42	Deleterious	0.052	Tolerated	
chr1	90,152,730	C	T	nsyn SNV	c.C4088T	p.T1363M	RsH-F7-12_FR_OS	4.02	Deleterious	0.046	Damaging	rs5108483986
chr1	90,152,730	C	T	nsyn SNV	c.C4088T	p.T1363M	RsH-F7-12_VL_OS	4.02	Deleterious	0.046	Damaging	rs5106483986
chr1	90,152,730	C	T	nsyn SNV	c.C4088T	p.T1363M	TcH-F3-20_FR_OS	4.02	Deleterious	0.046	Damaging	rs5108483986
chr1	90,156,578	G	T	nsyn SNV	c.G4993T	pAI665S	TcH-F3-20_FR_OS	-0.34	Neutral	0.020	Damaging	rs50009999
chr1	90,134,066	C	T	sg SNV	c.C1333T	p.R445X	RsH-F8-23_VL_FS	NA	NA	NA	NA	
chr1	90,149,032	C	T	nsyn SNV	c.C3464T	p.T1155I	TcI-i-F3-8_VL_FS	-3.61	Deleterious	0.016	Damaging	rs549684833
chr1	90,153,329	T	C	nsyn SNV	c.T4358C	p.V1453A	RsH-F1-13_VL_FS	-2.62	Deleterious	0.1112	Damaging	

The function of *MROH2A* in humans has not been characterized and there have been no published papers on the gene (https://www.ncbi.nlm.nih.gov/gene/339766). By cBio, *MROH2A* has a somatic mutation frequency of just 0.1%, suggesting it's not a superb candidate for follow-up. But, examination of recently added TCGA sarcoma data finds that *MROH2A* has a deep deletion in more than 5% of sarcomas (22).

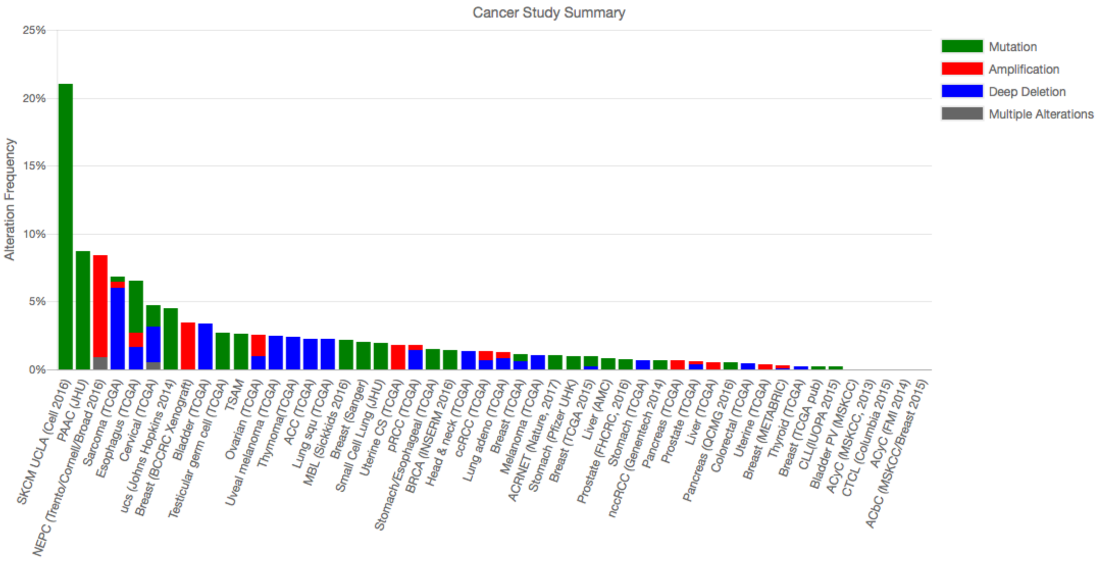


Figure 50: cBio data for MROH2A (22). Most notably, sarcomas are the fourth most common tumor with *MROH2A* alterations, with the majority of these being deletions.

We also note that *MROH2*A is a relatively large gene, spanning almost 60 kb, and 1,800 codons. Therefore, it may not be unexpected to have acquired somatic variants by chance alone. Despite this, the combination of it being the most recurrently mutated gene, and recent CN data as seen in cBio make it an interesting candidate for additional functional follow-up.

8.2.2 Inspection of multiple hits in the p53 pathway may be meaningful

Disruption of key pathways is often considered to be a hallmark of cancer (2, 54, 139). However, pathways often have built-in redundancies, and a single alteration may be insufficient for tumorigenesis (140). For example, in renal carcinoma, loss of VHL or PBRM alone are not sufficient for tumorigenesis, but co-occurring *VHL*, *PBRM1*, and *SETD2* mutations are observed, the latter two of which are both involved in chromatin remodeling via the SWI/SNF complex (141, 142).

This runs counter to the classic idea that mutations in the same pathway are redundant and therefore unlikely, and the concept of mutual exclusivity for some gene pairs (139). For example, in glioblastoma, *TP53* and *MDM2* were found to be mutually exclusive (143).

Emerging data from melanoma presents a more nuanced picture. Classically, *NRAS* hot-spot mutations are known to be mutually exclusive from *BRAF* hot-spot mutations at codons 600 and 601. That is, each is strong enough on its own to lead to tumorigenesis. However, some hotspot *NRAS/KRAS/HRAS* mutations have been shown to appear concomitantly with recurring mutations in *BRAF* (not at codons 600/601), suggesting that two alterations in the same pathway are possible or even required (144). These observations indicate that two variants, with possibly weaker effects, could combine to create a similar effect as one strong mutation.

Moreover, in malignant peripheral nerve sheath tumors (MPNST), a soft tissue sarcoma, evidence suggests that activation of the Ras pathway takes multiple hits. In addition to a germline mutation in *NF1* (coupled with somatic loss), multiple MPNSTs had additional somatic variants, predicted to be pathogenic (e.g. *PIK3CA, KIT, PTPN11,* and *FGFR1*, among others) (145). However, no single gene has emerged as being highly recurrent in conjunction with mutant *NF1*. (145). Taken together, these data suggest that looking for multiple hits in the *p53* pathway (or any cancer pathway) has intrinsic value, although this would best be done with a high number of samples to allow for potentially weaker effects to be more evident.

8.2.3 Notable somatic variants in individual tumors

To better ascertain variants more likely to be true somatic mutations, we increased the threshold for inclusion to require three or more algorithms. We took the following lists:

1. Krahe Lab p53 network genes (n = 131 genes)
2. Krahe Lab Sarcoma gene list (Table 7) (n = 147 genes)
3. Cancer Gene Census (146) (n = 609 genes)

to prioritize our search for known additional genetic risk factors. These lists are not redundant with each other (Figure 51), with at least some of the differences due to how the lists were curated. For example, some genes on the sarcoma list are based on only a few case reports.

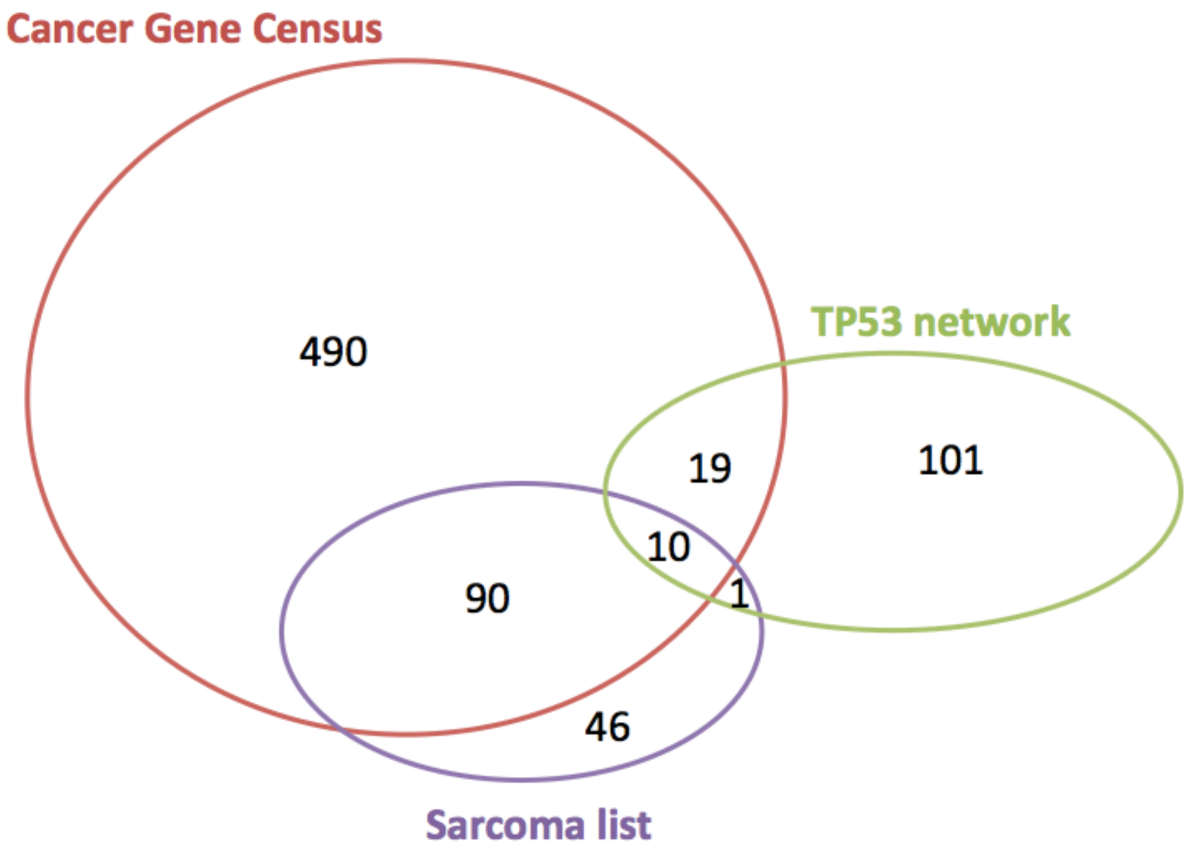
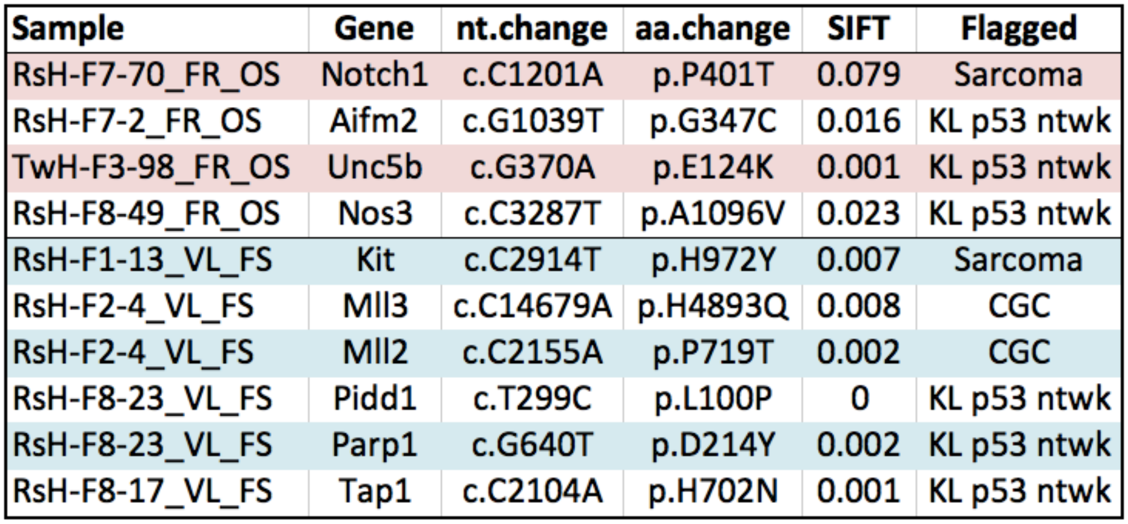


Figure 51: Venn diagram of overlap of three gene lists: Cancer Gene Census, Krahe Lab p53 network, and Krahe lab sarcoma list. Figure not to scale.

Several mouse sarcomas had somatic mutations in known sarcoma- or cancer-risk genes (Table 15). Investigation of *p53* network genes found potentially deleterious alleles in about half of tumors (Table 15), suggesting that the *p53* network does not need to be compromised in multiple places in the context of the *Trp53*^{R172H} mutation.

Table 15: Partial table of hits in known sarcoma, cancer or *p53* network genes.

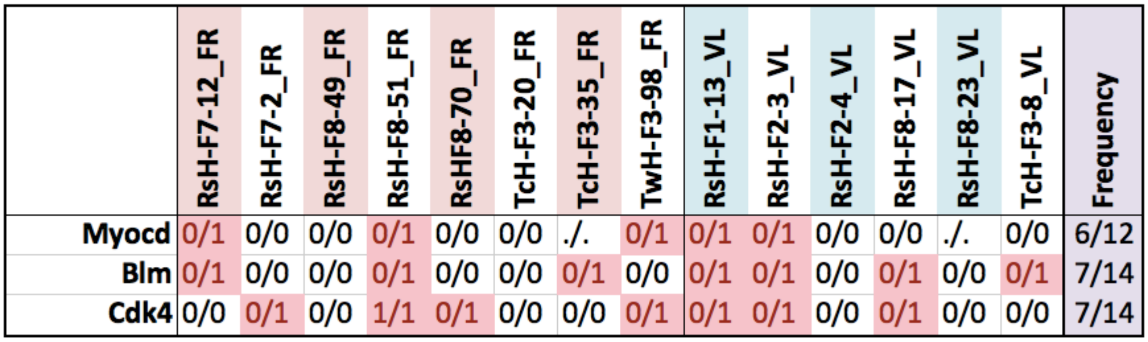


Sample	Gene	nt.change	aa.change	SIFT	Flagged
RsH-F7-70_FR_OS	Notch1	c.C1201A	p.P401T	0.079	Sarcoma
RsH-F7-2_FR_OS	Aifm2	c.G1039T	p.G347C	0.016	KL p53 ntwk
TwH-F3-98_FR_OS	Unc5b	c.G370A	p.E124K	0.001	KL p53 ntwk
RsH-F8-49_FR_OS	Nos3	c.C3287T	p.A1096V	0.023	KL p53 ntwk
RsH-F1-13_VL_FS	Kit	c.C2914T	p.H972Y	0.007	Sarcoma
RsH-F2-4_VL_FS	Mll3	c.C14679A	p.H4893Q	0.008	CGC
RsH-F2-4_VL_FS	Mll2	c.C2155A	p.P719T	0.002	CGC
RsH-F8-23_VL_FS	Pidd1	c.T299C	p.L100P	0	KL p53 ntwk
RsH-F8-23_VL_FS	Parp1	c.G640T	p.D214Y	0.002	KL p53 ntwk
RsH-F8-17_VL_FS	Tap1	c.C2104A	p.H702N	0.001	KL p53 ntwk

8.2.4 Notable germline variation in sarcoma-related genes

Given previous reports that some LFS patients may have multiple cancer-predisposing mutations or that differing *p53* alleles have different penetrance (6.3), we looked at the germline mutation profiles for each of the mice using variants called by HaploCaller. We took the lists (146) previously outlined in 8.2.3 to prioritize looking for known additional risk factors.

The majority of mice contain germline variants (non-synonymous, non-frameshift, or frameshift) across three genes associated with sarcomas, *Cdk4, Blm,* and, *Myocd* the first two of which are associated with osteosarcomas, the last of which is associated with LMS (11).

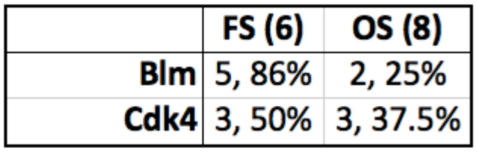


	RsH-F7-12_FR	RsH-F7-2_FR	RsH-F8-49_FR	RsH-F8-51_FR	RsHF8-70_FR	TcH-F3-20_FR	TcH-F3-35_FR	TwH-F3-98_FR	RsH-F1-13_VL	RsH-F2-3_VL	RsH-F2-4_VL	RsH-F8-17_VL	RsH-F8-23_VL	TcH-F3-8_VL	Frequency
Myocd	0/1	0/0	0/0	0/1	0/0	0/0	./.	0/1	0/1	0/1	0/0	0/0	./.	0/0	6/12
Blm	0/1	0/0	0/0	0/1	0/0	0/0	0/1	0/0	0/1	0/1	0/0	0/1	0/0	0/1	7/14
Cdk4	0/0	0/1	0/0	1/1	0/1	0/0	0/0	0/1	0/1	0/1	0/0	0/1	0/0	0/0	7/14

Figure 52: Germline variants in sarcoma predisposition genes in mice. These data are compressed; all variants indicated are predicted to be damaging, but may not necessarily be the same between mice. These data raise the possibility that they could contribute to tumorigenesis. However, we observe no LOH of these variants in the tumor, suggesting they may be haploinsufficient (in conjunction with mutated *p53*, or passengers. Key – "0/1" indicates a heterozygous variant, with the reference allele ("0") and the most common minor allele ("1"). "./." indicates insufficient data to make a genotype call, often due to insufficient coverage. "0/0" indicates a homozygous reference/WT genotype.

The overall relevance of these mutations is difficult to determine. It is possible that such germline variants contribute to the osteosarcomas observed in the R172H mouse (51), but we do not see LOH at these positions in the tumor, suggesting that these mutations may be haploinsufficient (in conjunction with mutant *p53*), or that they could be merely passengers. We also do not observe any strong predilection for mice in our cohort with germline variants in *Blm* and *Cdk4* to develop osteosarcomas over fibrosarcomas, suggesting at the very least they are not driving the sarcoma spectrum towards osteosarcomas.

Table 16: Distribution of germline variants in *Blm* and *Cdk4* by tumor type. Both genes have been previously associated with osteosarcomas, but in our mice, show no predilection towards OS over FS.



	FS (6)	OS (8)
Blm	5, 86%	2, 25%
Cdk4	3, 50%	3, 37.5%

Moreover, it is not known if these variants may be polymorphisms (though they do not carry dbSNP IDs).

We also examined our mouse cohort for non-recurrent variation, looking at each mouse individually for mutations in key genes. Three mice from our cohort have germline variants in sarcoma predisposition genes, and LOH in the tumor. RsH-F7-2 (OS) has a nonsynonymous SNV in *Chek2* that is predicted to be damaging (p.G263C, SIFT=0.04). Both RsH-F7-12 (OS) and RsH-F8-70 (OS) have the same germline mutation in *Bub1b,* (p.L726I, SIFT=0.033), a gene associated with embryonal rhabdomyosarcoma and aneuploidy in humans (147), and LOH in the tumor. Interestingly, in humans, both *CHEK2* and *BUB1B* have been associated with chromosomal instability (147-150). Although alterations in the two genes are not considered sufficient for chromosomal instability (CIN) on their own (151), these mice all have the germline mutation in *p53* (and LOH in the tumor). Thus, it is probable that some germline variants in these genes contributed to tumorigenesis in these mice, especially given their LOH in the tumors. While an accurate estimation of the penetrance of these germline mutations in the context of an underlying *p53* mutation is preliminary, given the small sample size, no other mice had these exact same mutations (including as a germline mutation with ROH in the tumor), suggesting 100% penetrance for these specific variants.

8.3 RNA-seq

Unfortunately, RNA extractions from fresh frozen tissues in the mouse produced sub-par RNA for the majority of tumors and could not be subjected to RNA-seq transcriptome profiling.

8.4 RRBS/Methylation data

RRBS data was analyzed in conjunction with Dr. Yue Lu. Data for RRBS was largely high quality, but two samples did not pass QC (RsH-F2-4 FS, RsH-F1-13 VL). Therefore, these were eliminated from further analyses.

8.4.1 Principal Components Analysis

Principal components analysis (PCA) of top 1% most variable sites, across all samples, showed roughly three distinct groups (Figure 53). There was good separation between normal femur (FR) and muscle (VL) samples, consistent with tissue-specific methylation. Meanwhile, the majority of tumors (FS, OS) seem to coalesce into a third group (towards the bottom of the plot). The OS that are not part of this third group do group with their matched tissues, normal femurs.

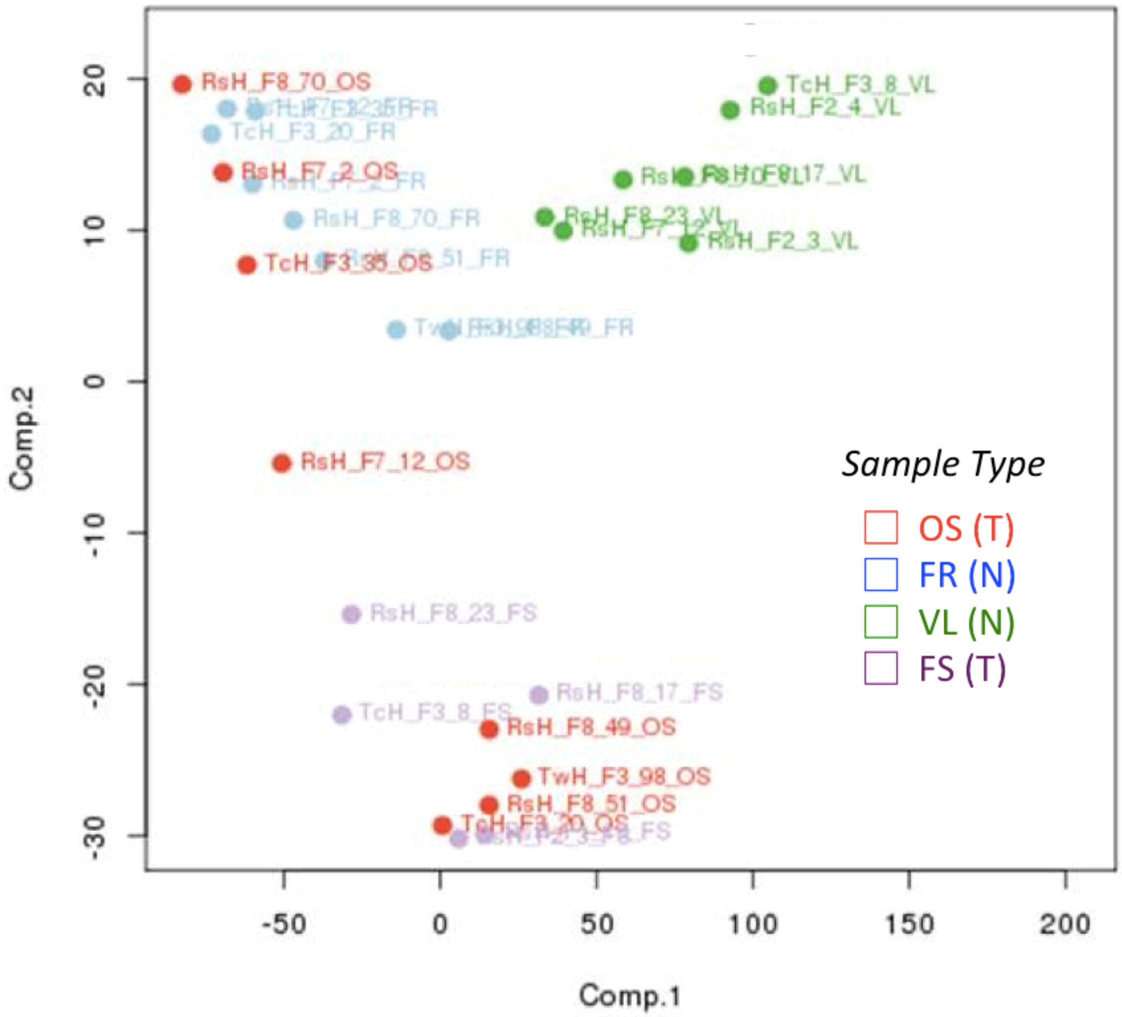


Figure 53: PCA of RRBS methylation data. Samples cluster into three distinct groups, generally comprised of FRs, VLs, and tumors (OS and FS). A couple of OS do group with the FR samples.

When we further plotted this PCA data against *p53 LOH* status from the pyrosequencing data, we do not see much additional granularity in separating the tumor samples (Figure 54). Tumors with more moderate LOH did seem to group together, but there continued to be several tumors that appeared to have more significant LOH that speckle the left side of the plot and were grouped with the femurs. These data also make clear that two presumably normal FRs that previously showed LOH by pyrosequencing still appeared to group with normal FRs by methylation, suggesting that changes to the methylome did not act as a precursor to tumorigenesis.

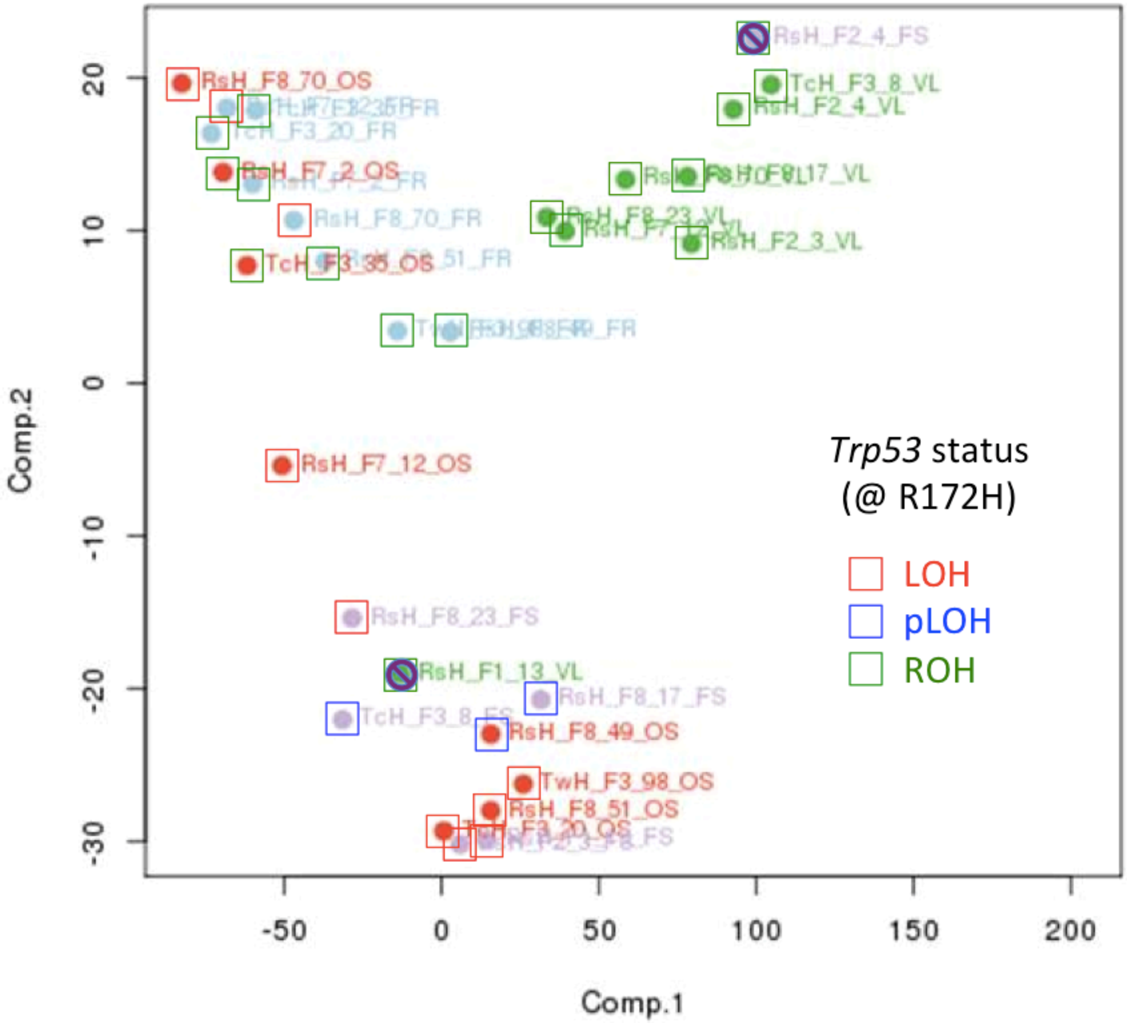
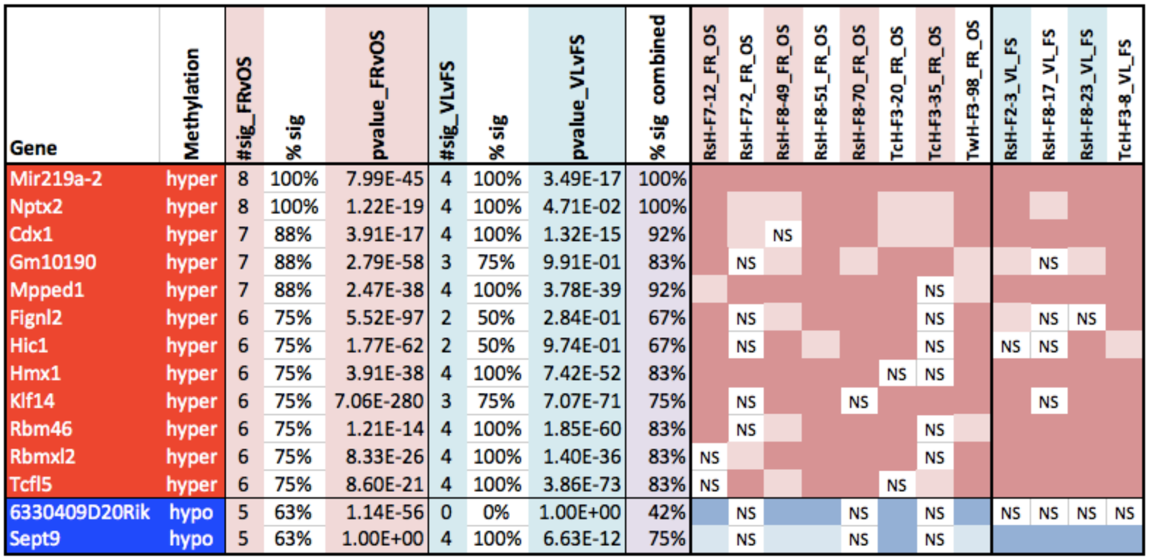


Figure 54: PCA analysis with the pyrosequencing LOH data laid over the top. There is some modicum of grouping of the samples with pLOH that indicate the possibility of a dosage effect, but one FS and three OS sarcoma samples dot the left side.

8.4.2 Recurrence in the methylation data

Unlike for somatic SNV data, we see many of the same changes across different mouse tumors from a methylation perspective. Due to tissue-specific differences, we considered genes that were differentially methylated across all tumors, or across each of FR/OS and VL/FS pairs. Multiple genes show significant hypermethylation across all tumors, including *Hic1* (hypermethylated in cancer, Table 17). However, of more immediate note is that we observed substantial hypermethylation of *Mir219a-2* across all mouse tumors.

Table 17: Methylation data in the mice shows significant recurrence. Overall, recurrence was much higher via for hypermethylation vs. hypomethylation. Key: NS – not significant.



Gene	Methylation	#sig_FRvOS	% sig	pvalue_FRvOS	#sig_VLvFS	% sig	pvalue_VLvFS	% sig combined	RsH-F7-12_FR_OS	RsH-F7-2_FR_OS	RsH-F8-49_FR_OS	RsH-F8-51_FR_OS	RsH-F8-70_FR_OS	TcH-F3-20_FR_OS	TcH-F3-35_FR_OS	TwH-F3-98_FR_OS	RsH-F2-3_VL_FS	RsH-F8-17_VL_FS	RsH-F8-23_VL_FS	TcH-F3-8_VL_FS
Mir219a-2	hyper	8	100%	7.99E-45	4	100%	3.49E-17	100%												
Nptx2	hyper	8	100%	1.22E-19	4	100%	4.71E-02	100%												
Cdx1	hyper	7	88%	3.91E-17	4	100%	1.32E-15	92%			NS									
Gm10190	hyper	7	88%	2.79E-58	3	75%	9.91E-01	83%		NS								NS		
Mpped1	hyper	7	88%	2.47E-38	4	100%	3.78E-39	92%							NS					
Fignl2	hyper	6	75%	5.52E-97	2	50%	2.84E-01	67%		NS					NS			NS	NS	
Hic1	hyper	6	75%	1.77E-62	2	50%	9.74E-01	67%		NS					NS		NS	NS		
Hmx1	hyper	6	75%	3.91E-38	4	100%	7.42E-52	83%						NS	NS					
Klf14	hyper	6	75%	7.06E-280	3	75%	7.07E-71	75%		NS			NS					NS		
Rbm46	hyper	6	75%	1.21E-14	4	100%	1.85E-60	83%		NS					NS					
Rbmxl2	hyper	6	75%	8.33E-26	4	100%	1.40E-36	83%	NS						NS					
Tcfl5	hyper	6	75%	8.60E-21	4	100%	3.86E-73	83%	NS					NS						
6330409D20Rik	hypo	5	63%	1.14E-56	0	0%	1.00E+00	42%		NS			NS		NS		NS	NS	NS	NS
Sept9	hypo	5	63%	1.00E+00	4	100%	6.63E-12	75%		NS			NS		NS					

There have been no published studies on *Mir219a2*. However, in human cancers, a type of sarcoma, malignant peripheral nerve sheath tumors (MPNST) are frequently associated with deep deletions (>6%, Figure 55). Therefore, loss of Mir219a-2/MIR219A2 may represent a novel key player in sarcomagenesis.

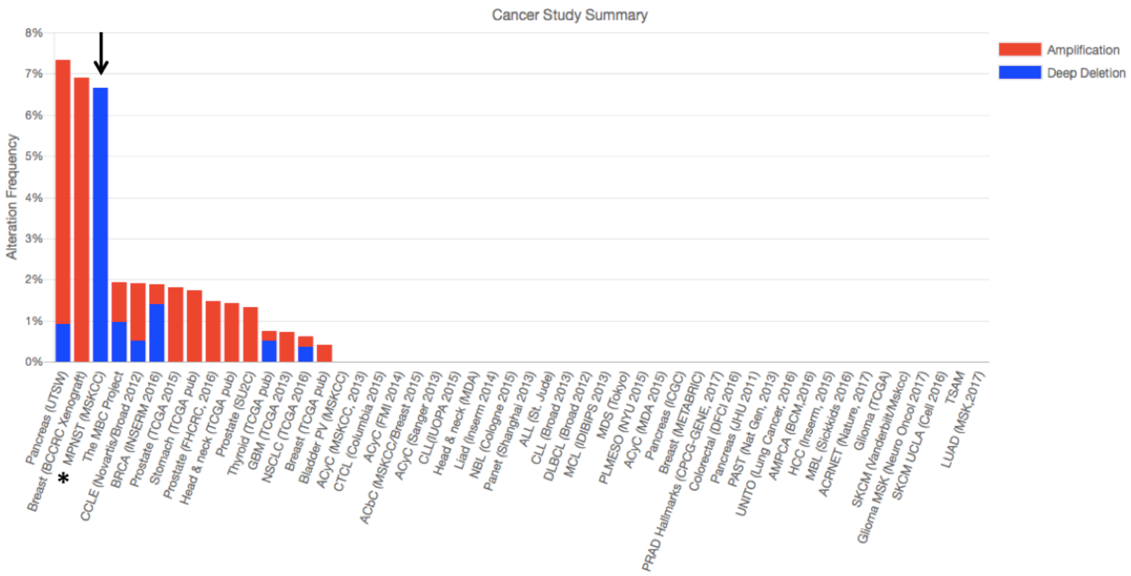


Figure 55: cBio data of *MIR219A2* (22). Although alterations are not particularly common across all tumor types and studies (partially due to not being assayed), deep deletions of MIR219A2 occur in almost 7% of MPNSTs, a type of sarcoma and mesenchymal tumor.

Several other genes also show hypermethylation across multiple tumors, including *Klf14*, *Cdx1* and *Tcf15*.

8.4.2.1 KLF14

KLF14 (Kruppel-like factor 14) has been identified as a tumor suppressor gene. In mice, disruption of the gene leads to aneuploidy and spontaneous tumorigenesis, particularly via centrosome amplification, a common event in cancer resulting the presence of extra centrosomes. *Klf14* directly targets Polo-like kinase 4 *(Plk4)*, considered to be a master regulator of centriole replication. In MEFs, Fan et al. found that loss of KLF14 led to genome instability, and tumorigenesis, while gain of KLF14 led to cell cycle arrest (152). In contrast to published work on *KLF14*, a big-picture view taken from the cBio database, suggests a different role. Instead, *KLF14* appears to be amplified in the majority of cancers, including sarcomas (Figure 56).

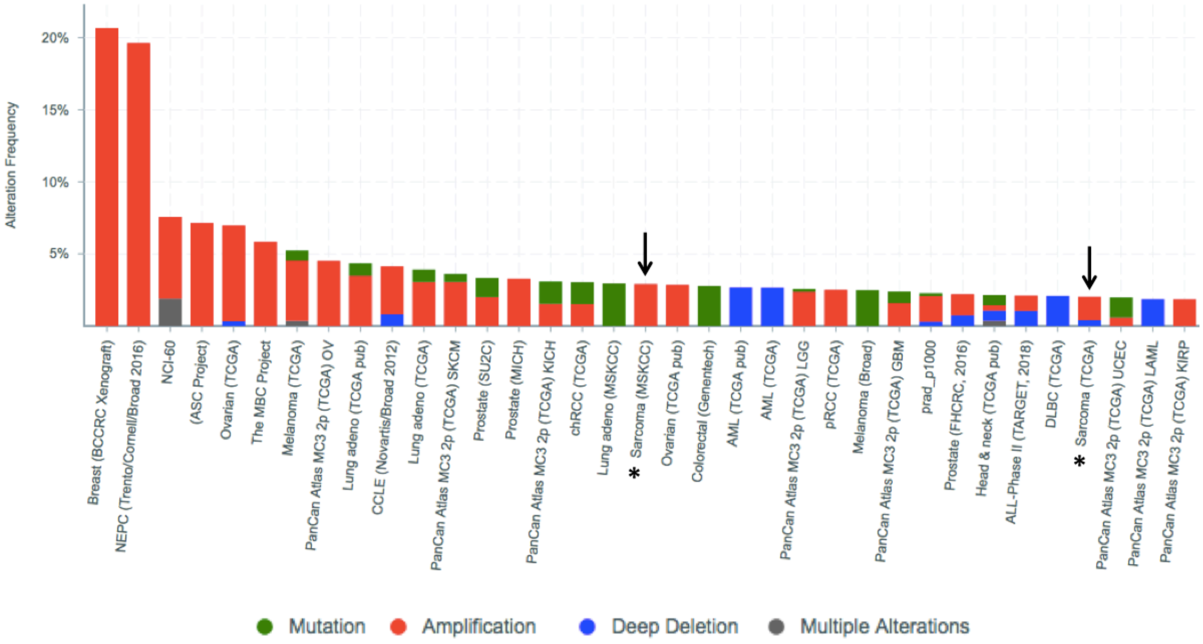


Figure 56: Alteration frequency of *KLF14* in the top 35 tumor studies in cBio ((22)). Contrary to published experiments, suggesting that *KLF14* acts like a tumor suppressor gene, multiple tumors have amplifications.

Therefore, given recurrence in our mouse data, and strong, published functional evidence indicating a role in genome instability, considered a hallmark of many sarcomas, *KLF14* may be particularly interesting for additional follow-up.

8.4.2.2 Cdx1

CDX1 has a known role in gut homeostasis and induction of *Cdx1* speeds up cell proliferation in colon cells, suggesting an *in vitro* oncogenic role. Consistent with this, the majority of colon cancer polyps have highly expressed *CDX1* (153). Moreover, *CDX1* is a target of the oncogenic pathways Ras and Wnt/β-catenin pathway as well as being a known regulator of the Rho, Ras, and PI3-Kinase pathways (153). However, our data, which indicate *Cdx1* is hypermethylated across our mouse tumors, is not consistent with this view.

Recent research indicates the role of *Cdx1* may be considerably more complex. For example, although the majority of colon polyps have highly expressed *Cdx1*, about 20% have lower expression of *Cdx1* (154). Moreover, in mice without both *Cdx1* and *Cdx2*, tumors were found to be less invasive, and less differentiated (154). Coupled with LOF function mutations also being found in other tumor types, such as carcinomas, this suggests that *CDX1* has a complex role in tumorigenesis, with both tumor suppressive and oncogenic effects.

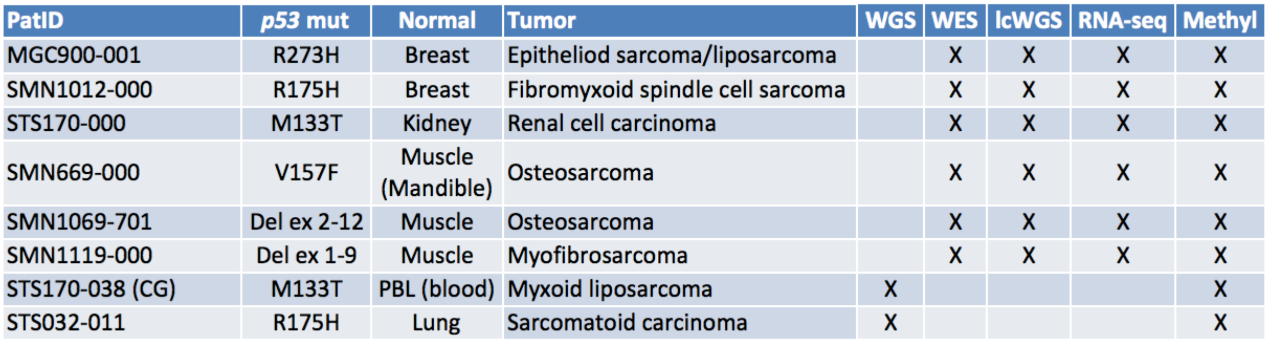
8.4.2.3 TCF15

_*TCF15* (transcription factor 15) is found to be altered in about 1% of all tumors catalogued by cBio (22). However, the available data somewhat contradicts each other – in MPNSTs (a mesenchymal tumor/sarcoma), 6.2% harbor deletions. In contrast, in the TCGA sarcoma (mostly adult STS) study, 2% of cases show amplification (5 cases), while 0.5% (1 case) show deletions. Additional data from a Memorial Sloan Kettering Cancer Center (MSKCC) sarcoma study aligns with the TCGA data (1.2% (4 cases) have amplification, 0.3% (1 case) has a deletion), potentially implying that *TCF15* has tumor subtype specific relevance. Given that we observe hypermethylation in our mice, our data most closely line up with the MPNST data. However, this runs contrast to expectation relative to known function. Previously, *TCF15* has been associated with priming and accelerating pluripotent cells for differentiation through repression of *NANOG* (155). Thus, a lack of *TCF15*, such as via hypermethylation could promote tumorigenesis through evasion of a more limited replicative potential associated with differentiated cells.

9 Results: Sequencing of Human Samples

We received several human sarcomas from Dr. Strong, spanning a variety of *p53* mutations and tumor types (Table 18).

Table 18: Table of human LFS sarcomas and omics analyses performed. The majority of WGS/WES data was generated on Illumina, except for STS170-038 (Complete Genomics).



PatID	*p53* mut	Normal	Tumor	WGS	WES	lcWGS	RNA-seq	Methyl
MGC900-001	R273H	Breast	Epitheliod sarcoma/liposarcoma		X	X	X	X
SMN1012-000	R175H	Breast	Fibromyxoid spindle cell sarcoma		X	X	X	X
STS170-000	M133T	Kidney	Renal cell carcinoma		X	X	X	X
SMN669-000	V157F	Muscle (Mandible)	Osteosarcoma		X	X	X	X
SMN1069-701	Del ex 2-12	Muscle	Osteosarcoma		X	X	X	X
SMN1119-000	Del ex 1-9	Muscle	Myofibrosarcoma		X	X	X	X
STS170-038 (CG)	M133T	PBL (blood)	Myxoid liposarcoma	X				X
STS032-011	R175H	Lung	Sarcomatoid carcinoma	X				X

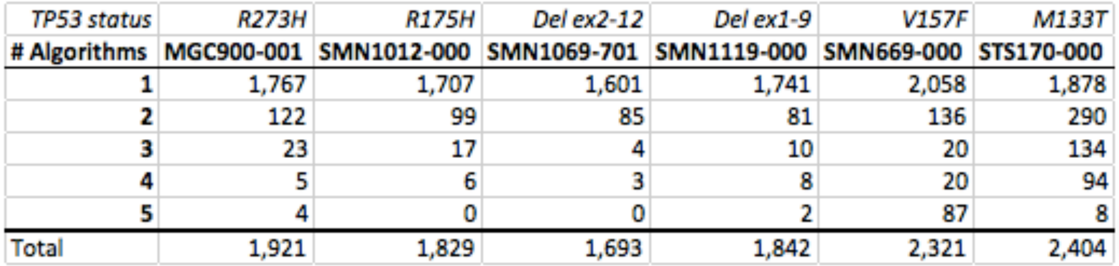
We then looked at several different kinds of next generation sequencing (NGS). For DNA, we looked at either at WGS, or at WES in conjunction with low-coverage WGS (lcWGS). The majority of the normal/tumor pairs were assayed on Illumina at 40X/80X coverage respectively. One sample was assayed on the Complete Genomics platform (STS170-038). Analysis for Illumina samples was done with the pipeline outlined in Figure 48.

For RNA, we were unable to isolate high-quality RNA from fresh tissues, presumably due to handling at the time of tumor extraction (and understandable prioritization of the patient), but we were able to leverage FFPE tissues instead to acquire suitable RNA. For methylation data, we used the EPIC array (Illumina), and analyzed the data via the R-package *minfi* (156). Although we still considered recurrence as an endpoint, analysis of both RNA and methylation data is further complicated by sample availability. For example, for osteosarcomas, we would ideally sequence a tissue-matched normal bone, rather than the muscle, due to tissue specificity in RNA and methylation. We can get around this by sequencing more appropriate tissue controls (e.g. a normal osteoblast cell line), but then these normal cell line controls lack the underlying p53 mutations present in the constitutive tissue. Moreover, recurrence analysis is potentially confounded by the variety of underling tumor types and germline *p53* mutations, making it more likely that they could have different underlying mechanisms of sarcomagenesis. Thus, it was also essential to consider individual tumor analysis (6.2 & 6.3).

9.1 Analysis of human tumors alone reveals minimal recurrence in somatic mutations

Similar to mouse samples, we saw little overlap between the five somatic variant callers. Using similar reasoning as 8.1.1, we used variants that were found by two or more somatic variant callers.

Table 19: Table showing agreement between variant callers on the WES human data.



TP53 status	*R273H*	*R175H*	*Del ex2-12*	*Del ex1-9*	*V157F*	*M133T*
# Algorithms	**MGC900-001**	**SMN1012-000**	**SMN1069-701**	**SMN1119-000**	**SMN669-000**	**STS170-000**
1	1,767	1,707	1,601	1,741	2,058	1,878
2	122	99	85	81	136	290
3	23	17	4	10	20	134
4	5	6	3	8	20	94
5	4	0	0	2	87	8
Total	1,921	1,829	1,693	1,842	2,321	2,404

Initial analysis of the human tumors for recurrent somatic changes in the DNA found no overlap, either at the mutation or gene level, with the exception of *p53*.

The lack of overlap could arise out of several scenarios. First, given the variety of germline mutations and tumor types, these tumors may have distinct etiologies from each other. Secondly, tumor heterogeneity, clonality, and contamination from normal tissue may obfuscate variant calling. Such mutations, occurring in just a fraction of the sample, could be missed without using deep sequencing. Thirdly, tumorigenesis in these patients may be driven by something other than somatically acquired SNVs or indels, such as CNAs or epigenetic changes.

Therefore, to work around the fact that these potentially confounding variables as best as possible, we examined each of the tumors individually for both germline and somatic variants.

9.2 Some human tumors had additional germline variants in cancer genes

Analysis of germline variation did find two individuals with rare mutations in key cancer-related genes. MGC900-001 (*p53 R273H*, breast cancer) had a mutation in the breast cancer predisposition gene *CHEK2* (p.R180C, MAF <0.01%) that was predicted to be damaging (SIFT=0.18, PP2=1). Notably, *CHEK2* had previously been associated with LFS (2.3.1). A second, different individual (SMN119-000, *p53 del ex 1-9*) had a mutation in the DNA repair gene *MLH3 (p.V741F)*, that was also relatively rare (MAF ~2%), and also predicted to be deleterious (SIFT=0.38, PP2=0.933). Although neither tumor showed of the remaining WT allele, they may still contribute to the overall risk profile of the individuals on a mutant *TP53* germline background. These data demonstrate the utility in looking not only at acquired somatic mutations, but also at germline variation.

9.3 Individual human tumors had mutations in known sarcoma genes

Individual analysis of human tumors was useful not only at the germline level, but also at the somatic level. Individual sarcomas had notable somatic mutations in sarcoma related genes. One tumor had an early, heterozygous, somatic stop-gain mutation in *PTEN (p.G20X). G*iven that PTEN is a haploinsufficient TSG (157, 158), such a mutation would be expected to have functional consequences. A second tumor had a stop-gain mutation in *NOTCH1*. *NOTCH1* is a transmembrane receptor that is typically considered to be an oncogene, having been found to be up-regulated in synovial sarcomas and rhabdomyosarcomas (159) and abnormalities in Notch signaling have been linked with pediatric sarcomas (159).

However, in the majority of solid tumors, genetic alterations are rare in any member of the Notch signaling pathway. Instead, deregulation of Notch signaling may play a larger role in tumor maintenance (159). Indeed, the role of Notch signaling is becoming considerably more complex than initially suspected. For example, *NOTCH1* has been shown to be activated in some osteosarcomas (160), but in other cases, it has also been shown to have tumor suppressive properties (161), suggesting a highly context-dependent role for Notch signaling that may be either oncogenic or tumor-suppressive (161).

Overall, the lack of consensus in the human samples, possibly due to tumor type heterogeneity, may argue for a different approach. Rather than use the human samples as a discovery set, and use the mouse model to confirm recurrent changes, do the reverse.

Use tumors from the mouse model, where the mutation, tumor types, and environment are more consistent, to look for recurrence, and follow-up in the human tumors.

9.4 Copy number analysis in human data

Alternatively, other types of alterations may explain the lack of consensus seen in somatic mutations. Chromosomal instability is considered to be hallmark of sarcomas (10). We worked with Dr. Nicholas Navin and a graduate student of his, Naveen Ramesh, to generate copy number analyses using our WES and lcWGS data and their pipeline. Consistent with previously published data, we observed copy number alterations, especially across our tumors with point mutations in *p53* (Figure 57). Our data is generally consistent with reported data for sarcomas in that deletions are more common than amplifications (10). Both MGC900-001 and SMN669-000 tend to have more deletions than amplifications, but this was less true for SMN1012-000 and STS170-000, where amplifications may have been more prevalent.

This effect was less pronounced, or even non-existent in tumors we sequenced that had *p53* deletions (Figure 58). These data may partially explain the lack of somatic point mutations observed in our human tumor; larger events may be driving sarcomagenesis in these patients.

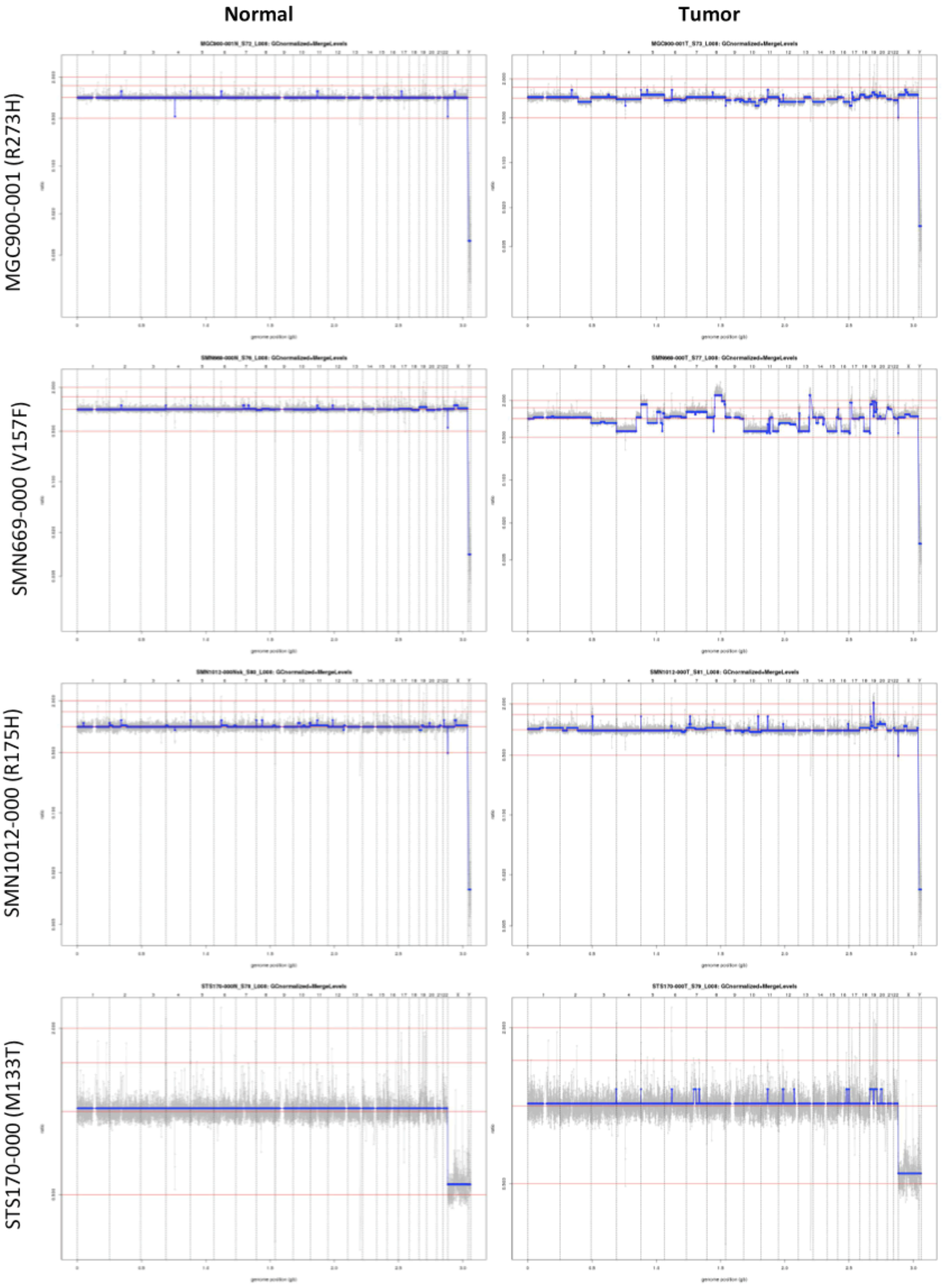
Normal
Tumor
MGC900-001 (R273H)
SMN669-000 (V157F)
SMN1012-000 (R175H)
STS170-000 (M133T)
MGC900-001N_S72_L008: GCnormalized+MergeLevels
MGC900-001T_S73_L008: GCnormalized+MergeLevels
SMN669-000N_S76_L008: GCnormalized+MergeLevels
SMN669-000T_S77_L008: GCnormalized+MergeLevels
SMN1012-000Nsk_S80_L008: GCnormalized+MergeLevels
SMN1012-000T_S81_L008: GCnormalized+MergeLevels
STS170-000N_S78_L008: GCnormalized+MergeLevels
STS170-000T_S79_L008: GCnormalized+MergeLevels
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 2122 X Y
ratio
2.000
0.500
0.100
0.020
0.005
0 0.5 1.0 1.5 2.0 2.5 3.0
genome position (gb)

Figure 57: Copy number analysis of human tumors with point mutations in *p53* indicates significant copy number aberrations in human LFS tumors. Data generated in collaboration with Naveen Ramesh (Dr. Nicholas Navin lab). Grey lines represent ploidy determinations, while blue lines represent copy number calls. Red horizontal lines represent 1N, 2N, 3N, and 4N from bottom to top respectively.

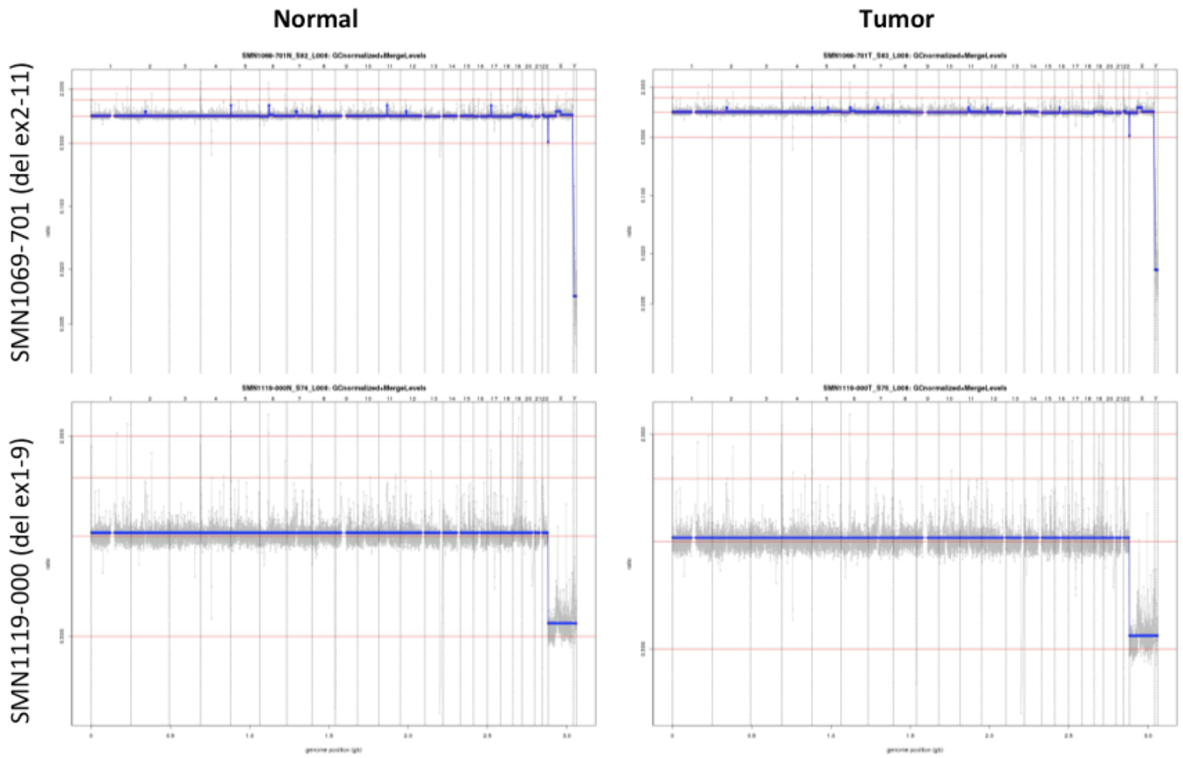


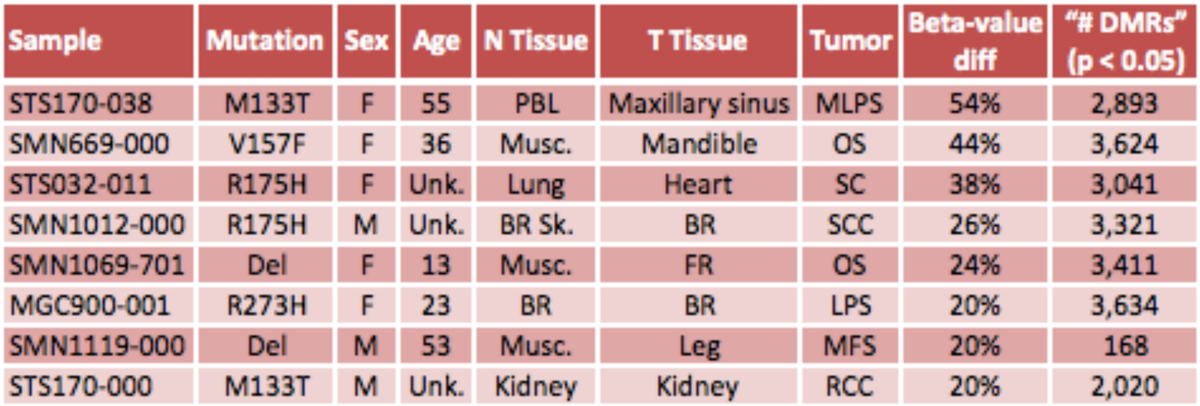
Figure 58: Copy number analysis of human tumors with deletions in *p53* indicates relatively few copy number aberrations in some human LFS tumors. Data generated in collaboration with Naveen Ramesh (Dr. Nicholas Navin lab). Grey lines represent ploidy determinations, while blue lines represent copy number calls. Red horizontal lines represent 1N, 2N, 3N, and 4N from bottom to top respectively.

9.5 Analysis of methylation data showed two distinct profiles

For human methylation data, we used Illumina's HumanMethylationEPIC (EPIC) array on the same DNA that we used to perform WES and lcWGS. To analyze human methylation data, we used the *minfi* R-package and *ssnoob* pre-processing, followed by *bumphunter* to find differentially methylated regions (DMRs) as described by Fortin et al. (162).

The method calls for a minimum default cutoff of 0.2 (i.e. a 20% difference in the beta-values between normal/tumor), but for some samples this led to hundreds of thousands of candidate bumps and long compute times. Since past a certain point, results are less likely to be considered significant after adjusting for multiple testing and permutations, the general consensus is to increase the cutoff (in our case up to about a 50% difference in beta values), in order to drive the number of candidate bumps (without consideration of p-value) down to ~30,000-40,000. We can see then some normal/tumor pairs have very few DMRs (SMN1119-000, and STS170-000), while others must have exceedingly high thresholds for the difference between beta-values (up to 54%, STS170-038), as seen in Table 20.

Table 20: Table of human samples with EPIC array data, indicating sex, age, and normal and tumor tissue types, and the number of bumps and DMRs identified by the pipeline. We note that the # of DMRs may be somewhat artificial since the pipeline uses # of DMRs to help generate appropriate parameters. PBL – peripheral blood leukocytes, BR – breast, FR – femur, MLPS – myxoid liposarcoma, OS – osteosarcoma, SC – sarcomatoid carcinoma, SCC – spindle cell carcinoma, LPS – liposarcoma, MFS – myofibrosarcoma, RCC – renal cell carcinoma.



Sample	Mutation	Sex	Age	N Tissue	T Tissue	Tumor	Beta-value diff	"# DMRs" $(p < 0.05)$
STS170-038	M133T	F	55	PBL	Maxillary sinus	MLPS	54%	2,893
SMN669-000	V157F	F	36	Musc.	Mandible	OS	44%	3,624
STS032-011	R175H	F	Unk.	Lung	Heart	SC	38%	3,041
SMN1012-000	R175H	M	Unk.	BR Sk.	BR	SCC	26%	3,321
SMN1069-701	Del	F	13	Musc.	FR	OS	24%	3,411
MGC900-001	R273H	F	23	BR	BR	LPS	20%	3,634
SMN1119-000	Del	M	53	Musc.	Leg	MFS	20%	168
STS170-000	M133T	M	Unk.	Kidney	Kidney	RCC	20%	2,020

However, some of, or perhaps even the majority of these DMRs may be due to tissue-specific differences (163). As an alternative, it may be better to ascertain DMRs against a tissue-matched control from a different person, or cell line.

9.6 Recurrence analysis in human methylation data

Confounding information aside, there did appear to be some recurrence at the methylation level. We found no DMRs that were consistent across all pairs, with the top five hypomethylated genes including *LHFPL2, PTPRN2, NRG1, ESR1,* and *JARID2*, and the top five hypermethylated genes including *FOXN3, TWIST1, TCF4, ST7*, and *MICAL3*. However, we did not see any overlap with the mouse data.

These methylation data represent an interesting cross-section of genes of the genes that are hypomethylated, LHFPL2 is a transmembrane protein that is lightly altered when examined in cBio (<10% of any tumor type affected), across both amplification, deletions, and mutations, suggesting potentially tumor or environment specific effects. *PTPRN2* is an oncogene that has been associated with promoting migration in breast cancer (164), while *ESR1* is thought to act as a tumor suppressor gene (165). *NRG1* has also been connected to breast cancer where it has been shown to induce stem-like properties (166). Recently, Xi et al. demonstrated that knock down of JARID2 inhibited invasiveness while overexpression promoted in in bladder cancer cells (167). Thus, many of the genes that are hypomethylated appear to have oncogenic functions.

In concert with this, some of the hypermethylated genes generally appear to have tumor suppressor-like functions. Loss of *FOXN3* in colon cancer was found to activate beta-catenin signaling and promote cell growth and migration (168). *ST7* has been identified as a tumor suppressor gene in prostate cancer (169). On the other hand, *TWIST1* is more commonly considered an oncogene that induces EMT (170). *MICAL2* is also considered an oncogene and has been associated with metastatic progression and a potential regulator of EMT (171).

9.7 Individual analysis in human methylation data

SMN669-000 had significant hypermethylation upstream of *MLH1* (shore, p-value=7.30e-4), and a CpG island of *PTEN* that is upstream of the transcription start site by less than 1,500 bp (p-value: 5.26E-04).

10 Discussion and Future Directions

Overall, when looking at the data we saw little recurrence between human samples, but significant recurrence between the mouse samples. This is consistent with the idea that different *p53* alleles have different penetrances, and considering that different sarcomas may have different etiologies (10, 15, 172), this may not be too surprising. An overall examination of the most recurrent hits, including prioritizing for variation found in both humans and mice emphasizes this (Table 21).

Table 21: Recurrence across both human and mouse sarcomas, across all data types.

	Human Sarcomas								Mouse Osteosarcomas								Mouse Fibrosarcomas					
Gene	STS032-011	SMN1012-000	STS170-038	STS170-000	MGC900-001	SMN669-000	SMN1069-701	SMN1119-000	RsH-F7-12_FR_OS	RsH-F7-2_FR_OS	RsH-F8-49_FR_OS	RsH-F8-51_FR_OS	RsH-F8-70_FR_OS	TcH-F3-20_FR_OS	TcH-F3-35_FR_OS	TwH-F3-98_FR_OS	RsH-F1-13_VL_FS	RsH-F2-3_VL_FS	RsH-F2-4_VL_FS	RsH-F8-17_VL_FS	RsH-F8-23_VL_FS	TcH-F3-8_VL_FS
TP53	🟩Y	🟩Y	🟩Y	🟩Y	🟩Y	🟩Y	🟩Y	🟩Y	🟩Y	🟩Y	🟩Y	🟩Y	🟩Y	🟩Y	🟩Y	🟩Y	🟩Y	🟩Y	🟩Y	🟩Y	🟩Y	🟩Y
CHEK2		🟪Y								🟩Y												
MROH2A									🟪Y				🟪Y	🟪Y			🟪Y				🟪Y	🟪Y
HIC1									🟥Y	🟥Y	🟥Y	🟪Y	🟥Y	🟥Y	🟥Y	🟥Y	🟦Y	🟦Y	🟥Y	🟥Y	🟥Y	🟦Y
MIR219a-2									🟥Y	🟥Y	🟥Y	🟥Y	🟥Y	🟥Y	🟥Y	🟥Y		🟥Y	🟥Y	🟥Y	🟥Y	🟥Y
AHDC1											🟪Y		🟪Y						🟪Y			🟪Y
PTEN			🟪Y			🟥Y																
TMEM132D									🟦Y	🟥Y	🟥Y	🟥Y	🟥Y			🟪Y		🟥Y		🟥Y	🟥Y	
NOTCH1	🟪Y																🟪Y		🟪Y			
KLF14									🟥Y		🟥Y	🟥Y		🟥Y	🟥Y	🟥Y		🟥Y			🟥Y	🟥Y
FOXS1									🟦Y			🟦Y		🟦Y		🟦Y		🟦Y		🟦Y		

Key 🟩 Germline 🟪 Somatic 🟥 Hyper-CH3 🟦 Hypo-CH3

Thus, these data underscore the power of our comparative –omics approach – not necessarily because of the ability to cross-compare human and mouse data and determine overlapping recurrence between the two, but because the mouse platform is so much more stable. For rare diseases/tumors such as LFS/sarcomas, the reduction of confounding factors (e.g. genetic heterogeneity) can be a powerful discovery space. For example, *MIR219A2* is recurrently hypermethylated in all mouse tumors, which we then found to have a deep deletion in sporadic MPNSTs.

In contrast, many tumors across mice and human showed alterations at the germline, somatic or methylation level in well–known sarcoma genes like *PTEN*. If these are sufficient on their own to lead to sarcomagenesis, it reduces the value in looking for recurrence in trying to identify novel cancer genes. However, this does not mean that a comparative approach is without merit. The Cancer Gene Census currently identifies 610 genes (https://cancer.sanger.ac.uk/census), and represents a good baseline for the low-hanging fruit. However, it is important to note that there likely are additional undiscovered cancer genes that can be found, particularly using a case study approach. Increasingly, it appears that cancer subtypes may have distinct etiologies, including in sarcomas (10, 15).

Additionally, one way to further identify novel cancer genes may be to look for genes that cooperate with known players such as *p53*. Previously, we discussed the possibility that while disruption of key pathways is often considered to be a hallmark of cancer (2, 54, 139), pathways often have built-in redundancies, and a single alteration may be insufficient for tumorigenesis (140). Alternatively, it may be that additional *p53* cooperating mutations may contribute to tumor type and/or spectrum. Although we found no evidence for this in our data – we had no mutations/alterations that were found only in fibrosarcomas vs. osteosarcomas, it is possible that a relative small sample size somewhat obscures these data, or that considering multiple alterations may have more predictive power.

Overall, our data agrees with previously published data – sarcomas have relatively low point mutation burdens, but many have significant copy number aberrations (CNAs) (10, 15). Intriguingly, in our data, patients with germline point mutations in *p53,* tend to have more CNAs than those with large deletions. It is not clear if this could be due to sample purity, or if *p53*-deletion sarcomas are simply more CN-stable. One other possibility is that the type of mutation may matter; while point mutations may be either GOF or LOF, deletions are LOF only. This concept is not addressed in the TCGA paper and may be an interesting point for follow-up.

Overall, the data from TCGA (10) highlight the need to consider tumor subtypes an important factor in all sequencing analyses. This significantly complicates sarcoma research moving forward, requiring expertise across a variety of fields, ranging from accurate sarcoma subtyping from the pathologist to doctors to comprehensive sequencing and analysis by bioinformatics specialists. Moreover, our data and published TCGA data indicate that any of a number of factors from point mutations, to CNAs, to methylation profiles likely contribute to sarcomagenesis.

If all sarcoma subtypes have different drivers, this may make it a cost-prohibitive endeavor or time-insensitive project, due to the need to acquire dozens of samples of the same type. Our data indicate a role for using a mouse model to identify potential sarcoma drivers, but we have focused on osteosarcomas and fibrosarcomas – generating a mouse model for even rarer types may not be possible. Alternatively, given sufficient time clinicians may collect enough samples from human patients to complete such a study. Importantly though, we have used our data to identify potential impact players which are both novel (e.g. *MIR219A2*, and with known roles in other cancers such as *PTEN* and *BUB1B*). Some of these have appeared to be common to both osteosarcomas and fibrosarcomas, suggesting the potential that some drivers may be common to multiple sarcomas. Thus, a comparative omics approach gives tremendous value to focusing on sequencing more common sarcoma types first. Given that the end goal is to identify important genes, which in turn may impact clinical outcomes, this may be sufficient to begin impact treatment.

Secondly, our data indicate that genetic mutations even in known sarcoma genes such as *PTEN* are generally rare across our samples, underlying an important point – the majority of sarcomas have low point mutation loads and instead are considered to have significant copy number aberrations (10). In addition, some sarcoma subtypes, such as synovial sarcoma subtypes had relatively uniform methylation profiles (10). We noticed consistent patterns in our mouse methylation data, including some genes that were recurrently hypermethylated and a PCA analysis that grouped both osteosarcomas and fibrosarcomas together. Moreover several of the top hits in the human methylation data had known functions in cancers. We highlighted four oncogenes that were hypomethylated (e.g., *PTPRN2, JARID2)*, and two tumor suppressor genes were hypermethylated (e.g., *FOXN3, ST7*). Thus, we feel that future sarcoma studies should consider focusing on methylation data over genomic sequence analysis of point mutations.

We continue to look for ways to iterate over this process to better understand this dichotomy of looking for recurrence first, or characterizing an individual tumor.

11 Conclusions

In conclusion, we have identified a novel germline LFSL mutation in *ARHGAP30* that co-segregates with disease across multiple LFSL families (four total to date). Moreover, we have demonstrated that this mutation, *ARGAHP30 (c.G161,017,761A, p.R806Q/p.R1017Q)*, has functional impact when overexpressed *in vitro* for both the short and long isoforms, suggesting that *ARHGAP30* may have *p53*-independent functions. To further strengthen our case, we are currently sequencing additional LFSL families to see if they contain other mutations in elsewhere in *ARHGAP30*, and testing tumors to see if they have LOH. Both results substantially strengthen the identification of *ARHGAP30* as an LFSL gene and an important gene for clinical testing for patients and families with LFS/LFSL phenotypes. Our hope and intention is that this will allow for genetic testing and improved tumor surveillance in individuals who have this genetic germline cancer predisposition.

Secondly, we have demonstrated the utility of a comparative –omics approach to identify potential key players in sarcomagenesis. Recent research by TCGA emphasizes tissue specificity in adult soft tissue sarcomas (10), and suggests significant value in sequencing similar sarcoma types. We were able to leverage a mouse model of LFS to identify two novel genes, *Mroh2a*, and *Mir219a-2* as potential players in sarcomagenesis, the latter of which has relevance to MPNSTs (145). Strikingly, although we observe few overlaps in somatic mutations (at the base or gene level), we do observe significant overlap at the methylation level, including between tumor types, suggesting that epigenetic instability may be critically important, consistent with *p53* as the guardian of the epigenome (18). Across both our mouse and human tumors, we see recurrent epigenetic changes in both novel, and known cancer genes.

However, our data also indicate that all types of -omics approaches can point to alterations in key genes, and that these may be relatively private. In combination with a large existing knowledge base in cancer, these “private” alterations may be considered generally sufficient to explain tumorigenesis. We note that it is possible that these tumors have other mutations that contribute to cancer in smaller ways, or may be involved in cancer initiation as opposed to cancer progression. However, that these mutations are found in bulk tumor suggests that they occur in a large proportion of sequenced cells and may be more likely to be key drivers rather than late-comers. Thus, recurrence should not be used as the only endpoint when considering sequencing studies.

In conclusion, based on our collective data we would argue that to best understand sarcomagenesis moving forward, one should focus on complete tumor profiling in as few subtypes as possible. Furthermore, we would recommend an emphasis on evaluating CNAs and methylation profiles, ultimately suggesting there is still strong utility in using LFS and LFSL as model disorders to identify variation that is important not only for sarcomas, but potentially other tumors as well.

12 Materials and methods

12.1 Subjects

Tissues and genomic DNA were kindly provided by Dr. Louise Strong's lab (MD Anderson Cancer Center), Dr. Henry Lynch's lab (Creighton University), Dr. Albert de la Chapelle's lab (Ohio State University), and Dr. Mai Phuong's lab (National Cancer Institute).

12.2 WGS/WES

WGS data were sequenced at 100bp PE across three platforms GAIIX (STS200-017), HS2000 (STS200-000, STS200-017, STS032-011), and HS4000 (STS200-001, STS200-008, STS200-009, STS200-019, STS200-108) at 30-40X. Remaining normal/tumor pairs on mice were sequenced on HS3000 at 40X/80X respectively. Samples were analyzed according to best practices as described by GATK or individual algorithms.

12.3 Sanger sequencing

Primers were designed in-house, and PCR performed according to standard protocols (HotStarTaq Master Mix kit, QIAGEN). PCR products were cleaned using Diffinity RapidTIps (Sigma) for small experiments, or via ethanol precipitation for full plates before submission to the Sequencing and Microarray Facility (SMF) at MD Anderson Cancer Center. Sequencing was generated in both directions. Traces were analyzed with Mutation Surveyor.

hARHGAP30 R806Q F *CCACAGTTTGCCAAGATGCC*

hARHGAP30 R806Q R *GGTCCTAATCACAGTCCTTCAC*

hARHGAP30-UTR F	tgtaaaacgacggccagtTCTCTTCCTTATTTCCTGACC
hARHGAP30-UTR R	caggaaacagctatgaccCCCTAAGATACCTCCTGTCC
hF11R F	tgtaaaacgacggccagtTCTAAGGAGGAAGTAGGAAAGG
hF11R R	caggaaacagctatgaccTCTGCTCTTCCCAAGTTGTG
hMROH9 F	tgtaaaacgacggccagtATGACCAATATGAACCTCTTCC
hMROH9 R	caggaaacagctatgaccAAGACATTGTTGGACTTCCC

hMUC5B F	tgtaaaacgacggccagtAGTTCCAAAGCCACTTCCTC
hMUC5B R	caggaaacagctatgaccCTGTAAAGCTGGTAGCTGTG
hNDUFS2 F	tgtaaaacgacggccagtAAGACTACAGGGTTTATATGGG
hNDUFS2 R	caggaaacagctatgaccCAGAAGAATTGCTTGAACCTG
hPARP4 F	tgtaaaacgacggccagtCCATAGAATAACAAACTCTGCGTC
hPARP4 R	caggaaacagctatgaccTCTGGATGGAGCATTGAAAGAG
hPKLR F	tgtaaaacgacggccagtTGATACAAATGGTAGGAGTGG
hPKLR R	caggaaacagctatgaccGCCCAGAGAAGTATGATGAC
hRETSAT F	tgtaaaacgacggccagtCCTGTCAGATAGAGGTTGGG
hRETSAT R	caggaaacagctatgaccTGTTTCTGCCCTTTCCTTGAG
hSLAMF1 F	tgtaaaacgacggccagtCAACACAAAGATGGAACGCTG
hSLAMF1 R	caggaaacagctatgaccATGCTTATGCTTGGAAGGGAG
hTP53-UTR3 F	tgtaaaacgacggccagtTTAAATCCCGTAATCCTTGGTGAG
hTP53-UTR3 R	caggaaacagctatgaccTTACATTCTGCAAGCACATCTG

12.4 Mutation Surveyor

We ran both the default, recommended settings, as well as a set of hyper-relaxed criteria that picked to pick some true positives at the cost of a lot of false positives.

Because some genes are highly homologous (e.g. PYHIN1), we supplied known sequence to the program for such genes to force the correct mapping and eliminate erroneous variant calls.

12.5 454 sequencing

Pyrosequencing was performed on a PSQ 96 machine, according to manufacturer's instructions. 20ng of DNA was amplified by PCR before use with pyrosequencer. For FFPE samples, we used three times as much polymerase as we would for a normal PCR, and ran the annealing step in the cycle for two minutes.

hARHGAP30 F	TAAGACGACTCCGGGATCCAG
hARHGAP30 R	GACGGGACACCGCTGATCGTTTATCCCGAGCTTCTCGATCCT
hARHGAP30 Seq	TCAGTACAGGTCTGGGT

mP53H F	GGCCATCTACAAGAAGTCACAGCA
mP53H R	GACGGGACACCGCTGATCGTTTAAGGCGGTGTTGAGGGCTTA
mP53H Seq	GACGGAGGTCGTGAGA

12.6 Cell lines

Four cell lines (HEK293T, Saos-2, U-2 OS, and HT-1080) were acquired from the characterized cell line core facility at MD Anderson Cancer Center. The core provides mycoplasma-free, and fingerprinted cell lines.

12.7 Mice.

Existing *Trp53 R172H* mice were bred to create a total cohort of 50 mice with a heterozygous background. Twenty WT mice were kept as controls. Mice were monitored continuously for tumors and other health issues, and sacrificed when necessary. Original mice were received from Dr. Guillermina Lozano and are on a predominantly C57BL/6 background.

12.8 Genotyping

Mouse toes are collected at 7-10 days and digested with KAPA enzyme (Sigma) in 15uL total volume. Final volume is brought up to 100uL. DNA is then assayed by PCR and run on an agarose gel to look for band separation in heterozygotes.

mp53H F	ACCTGTAGCTCCAGCACTGG
mp53H R	ACAAGCCGAGTAACGATCAGG

12.9 Tissue collection

Mice were euthanized by cervical dislocation to best preserve RNA (over CO2 inhalation). Tissues were then immediately collected by snap freezing in liquid nitrogen: pectoralis, vastus lateralis, femur, gastrocnemius, kidney, spleen, liver, stomach, duodenum or fixed in paraformaldehyde for pathology: gastrocnemius, reproductive organs, digestive tract, heart, lung, thymus, head/brain, kidney, liver, spleen, leg, sternum, spine. Lymph nodes and tumors were collected where possible. All tissues were sent to Elizabeth Whitley (Pathogenesis LCC) for pathology

12.10 DNA isolation

DNA was harvested from fresh frozen tissues using a mini-prep kit (Zymo) according to manufacturers instruction, with one exception. Lysis was accomplished through gentle rotation in a LabQuake shaker to avoid shearing of high-molecular weight DNA during inversion, or vortexing. DNA was assayed for genomic integrity by agarose gel (0.8%, 20V), and quantitated using PicoGreen (Molecular Probes)

12.11 FFPE RNA isolation

We cut 10, 10um sections with a microtome. RNA was isolated according to kit instructions for ReliaPrep^{TM} FFPE Total RNA Miniprep System.

12.12 Plasmids

We acquired constructs for the short isoform (NM_181720, Origene C/N RC208825) and the long isoform (NM_0010255598, Origene C/N RC217735). Sanger sequencing was used to confirm the appropriate sequence in the original plasmid. Plasmids were expanded using Ultra Competent XL10 cells (Stratagene), and DNA was isolated using Maxi kits (QIAGEN), as per kit instructions.

12.13 Site Directed Mutagenesis

Plasmids containing the mutant were generated through site-directed mutagenesis (Agilent) according to the kit, and checked for correct sequence via Sanger sequencing. Plasmids used for mutagenesis:

F: CTCAAGGAGTTCGGCGAACCCAGACCTGTAC

R: GTACAGGTCTGGGTTTGCCGAACTCCTTGA

12.14 Scratch Assay

Cells were initially seeded in 12-well plates at ~60-70% confluency. Following overnight attachment, transfection reagent (Lipofectamine 2/3000, or FuGene 6) and plasmid were added at 3:1 ratio according to kit instructions. 24 hours post transfection, cells were checked to see if they had achieved near 100% confluency. A single vertical scratch was made with a P20 pipette tip. Cells were then washed, and imaged immediately with a live-cell imaging microscope (Zeiss, 3i). Images were taken every 30 minutes for a period up to 48 hours.

12.15 Proliferation Assay

Cells were seeded in triplicate in 96-well plates at low density at an average concentration of approximately 10,000 cells/well. Once cells attached, we transfected according to the kit instructions (Lipfectamine 2/3000, ThermoFisher; FuGene6, Promega). Cells were then trypsinized, and counted twice/well using a hemacytometer and trypan blue (GIBCO). Cells were counted approximately every 24 hours.

12.16 Western Blots

Protein extracts (30ug) from transfected cells were run on pre-made polyacrylamide gels (Invitrogen), transferred to nitrocellulose (BioRad) and probed with antibodies for ARHGAP30 (Abcam, ab101965) and FLAG (Sigma, F3165). Appropriate secondary antibodies came from Odyssey. Blots were imaged on a Li-Cor (Odyssey).

13 Bibliography

1. Hanahan, D., and R. A. Weinberg. 2000. The hallmarks of cancer. *Cell* 100: 57-70.

2. Hanahan, D., and R. A. Weinberg. 2011. Hallmarks of cancer: the next generation. *Cell* 144: 646-674.

3. Boland, C. R., and A. Goel. 2010. Microsatellite instability in colorectal cancer. *Gastroenterology* 138: 2073-2087 e2073.

4. Stratton, M. R. 2011. Exploring the genomes of cancer cells: progress and promise. *Science* 331: 1553-1558.

5. Bell, D. W. 2010. Our changing view of the genomic landscape of cancer. *J Pathol* 220: 231-243.

6. Portela, A., and M. Esteller. 2010. Epigenetic modifications and human disease. *Nat Biotechnol* 28: 1057-1068.

7. Rhodes, D. R., and A. M. Chinnaiyan. 2005. Integrative analysis of the cancer transcriptome. *Nat Genet* 37 Suppl: S31-37.

8. Hoadley, K. A., M. B. Siegel, K. L. Kanchi, C. A. Miller, L. Ding, W. Zhao, X. He, J. S. Parker, M. C. Wendl, R. S. Fulton, R. T. Demeter, R. K. Wilson, L. A. Carey, C. M. Perou, and E. R. Mardis. 2016. Tumor Evolution in Two Patients with Basal-like Breast Cancer: A Retrospective Genomics Study of Multiple Metastases. *PLoS medicine* 13: e1002174.

9. Arifi, S., R. Belbaraka, R. Rahhali, and N. Ismaili. 2015. Treatment of Adult Soft Tissue Sarcomas: An Overview. *Rare cancers and therapy* 3: 69-87.

10. Network., C. G. A. R. 2017. Comprehensive and Integrated Genomic Characterization of Adult Soft Tissue Sarcomas. *Cell* 171: 950-965 e928.

11. Fletcher, C. D. M., J. A. Bridge, P. C. W. Hogendoorn, and F. Mertens. 2013. *WHO Classification of Tumours of Soft Tissue and Bone*. WHO Press.

12. Blay, J. Y., and I. Ray-Coquard. 2017. Sarcoma in 2016: Evolving biological understanding and treatment of sarcomas. *Nat Rev Clin Oncol* 14: 78-80.

13. Taylor, B. S., J. Barretina, R. G. Maki, C. R. Antonescu, S. Singer, and M. Ladanyi. 2011. Advances in sarcoma genomics and new therapeutic targets. *Nat Rev Cancer* 11: 541-557.

14. Burningham, Z., M. Hashibe, L. Spector, and J. D. Schiffman. 2012. The epidemiology of sarcoma. *Clinical sarcoma research* 2: 14.

15. Chudasama, P., S. S. Mughal, M. A. Sanders, D. Hubschmann, I. Chung, K. I. Deeg, S. H. Wong, S. Rabe, M. Hlevnjak, M. Zapatka, A. Ernst, K. Kleinheinz, M. Schlesner, L. Sieverling, B. Klink, E. Schrock, R. M. Hoogenboezem, B. Kasper, C. E. Heilig, G. Egerer, S. Wolf, C. von Kalle, R. Eils, A. Stenzinger, W. Weichert, H. Glimm, S. Groschel, H. G. Kopp, G. Omlor, B. Lehner, S. Bauer, S. Schimmack, A. Ulrich, G. Mechtersheimer, K. Rippe, B. Brors, B. Hutter, M. Renner, P. Hohenberger, C. Scholl, and S. Frohling. 2018. Integrative genomic and transcriptomic analysis of leiomyosarcoma. *Nat Commun* 9: 144.

16. Ballinger, M. L., D. L. Goode, I. Ray-Coquard, P. A. James, G. Mitchell, E. Niedermayr, A. Puri, J. D. Schiffman, G. S. Dite, A. Cipponi, R. G. Maki, A. S. Brohl, O. Myklebost, E. W. Stratford, S. Lorenz, S.-M. Ahn, J.-H. Ahn, J. E. Kim, S. Shanley, V. Beshay, R. L.

Randall, I. Judson, B. Seddon, I. G. Campbell, M.-A. Young, R. Sarin, J.-Y. Blay, S. I. O'Donoghue, and D. M. Thomas. 2016. Monogenic and polygenic determinants of sarcoma risk: an international genetic study. *The Lancet Oncology* 17: 1261-1271.

17. Lane, D. P. 1992. Cancer. p53, guardian of the genome. *Nature* 358: 15-16.

18. Levine, A. J., and B. Greenbaum. 2012. The maintenance of epigenetic states by p53: the guardian of the epigenome. *Oncotarget* 3: 1503-1504.

19. Kandoth, C., M. D. McLellan, F. Vandin, K. Ye, B. Niu, C. Lu, M. Xie, Q. Zhang, J. F. McMichael, M. A. Wyczalkowski, M. D. M. Leiserson, C. A. Miller, J. S. Welch, M. J. Walter, M. C. Wendl, T. J. Ley, R. K. Wilson, B. J. Raphael, and L. Ding. 2013. Mutational landscape and significance across 12 major cancer types. *Nature* 502: 333-339.

20. Li, F. P., and J. F. Fraumeni, Jr. 1969. Soft-tissue sarcomas, breast cancer, and other neoplasms. A familial syndrome? *Ann Internal Med* 71: 747-752.

21. Garber, J. E., and K. Offit. 2005. Hereditary cancer predisposition syndromes. *J Clin Oncol* 23: 276-292.

22. Cerami, E., J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, C. J. Byrne, M. L. Heuer, E. Larsson, Y. Antipin, B. Reva, A. P. Goldberg, C. Sander, and N. Schultz. 2012. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2: 401-404.

23. Crompton, B. D., C. Stewart, A. Taylor-Weiner, G. Alexe, K. C. Kurek, M. L. Calicchio, A. Kiezun, S. L. Carter, S. A. Shukla, S. S. Mehta, A. R. Thorner, C. de Torres, C. Lavarino, M. Sunol, A. McKenna, A. Sivachenko, K. Cibulskis, M. S. Lawrence, P.

Stojanov, M. Rosenberg, L. Ambrogio, D. Auclair, S. Seepo, B. Blumenstiel, M. DeFelice, I. Imaz-Rosshandler, Y. C. A. Schwarz-Cruz, M. N. Rivera, C. Rodriguez-Galindo, M. D. Fleming, T. R. Golub, G. Getz, J. Mora, and K. Stegmaier. 2014. The genomic landscape of pediatric Ewing sarcoma. *Cancer Discov* 4: 1326-1341.

24. Shern, J. F., L. Chen, J. Chmielecki, J. S. Wei, R. Patidar, M. Rosenberg, L. Ambrogio, D. Auclair, J. Wang, Y. K. Song, C. Tolman, L. Hurd, H. Liao, S. Zhang, D. Bogen, A. S. Brohl, S. Sindiri, D. Catchpoole, T. Badgett, G. Getz, J. Mora, J. R. Anderson, S. X. Skapek, F. G. Barr, M. Meyerson, D. S. Hawkins, and J. Khan. 2014. Comprehensive genomic analysis of rhabdomyosarcoma reveals a landscape of alterations affecting a common genetic axis in fusion-positive and fusion-negative tumors. *Cancer Discov* 4: 216-231.

25. Tirode, F., D. Surdez, X. Ma, M. Parker, M. C. Le Deley, A. Bahrami, Z. Zhang, E. Lapouble, S. Grossetete-Lalami, M. Rusch, S. Reynaud, T. Rio-Frio, E. Hedlund, G. Wu, X. Chen, G. Pierron, O. Oberlin, S. Zaidi, G. Lemmon, P. Gupta, B. Vadodaria, J. Easton, M. Gut, L. Ding, E. R. Mardis, R. K. Wilson, S. Shurtleff, V. Laurence, J. Michon, P. Marec-Berard, I. Gut, J. Downing, M. Dyer, J. Zhang, O. Delattre, P. St. Jude Children's Research Hospital-Washington University Pediatric Cancer Genome, and C. the International Cancer Genome. 2014. Genomic landscape of Ewing sarcoma defines an aggressive subtype with co-association of STAG2 and TP53 mutations. *Cancer Discov* 4: 1342-1353.

26. Brady, C. A., and L. D. Attardi. 2010. p53 at a glance. *Journal of cell science* 123: 2527-2532.

27. Haupt, Y., R. Maya, A. Kazaz, and M. Oren. 1997. Mdm2 promotes the rapid degradation of p53. *Nature* 387: 296-299.

28. Moll, U. M., and O. Petrenko. 2003. The MDM2-p53 interaction. *Molecular cancer research : MCR* 1: 1001-1008.

29. Wan, Y., W. Wu, Z. Yin, P. Guan, and B. Zhou. 2011. MDM2 SNP309, gene-gene interaction, and tumor susceptibility: an updated meta-analysis. *BMC cancer* 11: 208.

30. Canman, C. E., and M. B. Kastan. 1998. Small contribution of G1 checkpoint control manipulation to modulation of p53-mediated apoptosis. *Oncogene* 16: 957-966.

31. Khosravi, R., R. Maya, T. Gottlieb, M. Oren, Y. Shiloh, and D. Shkedy. 1999. Rapid ATM-dependent phosphorylation of MDM2 precedes p53 accumulation in response to DNA damage. *Proc Natl Acad Sci U S A* 96: 14973-14977.

32. Chehab, N. H., A. Malikzay, M. Appel, and T. D. Halazonetis. 2000. Chk2/hCds1 functions as a DNA damage checkpoint in G(1) by stabilizing p53. *Genes Dev* 14: 278-288.

33. Hay, T. J., and D. W. Meek. 2000. Multiple sites of in vivo phosphorylation in the MDM2 oncoprotein cluster within two important functional domains. *FEBS letters* 478: 183-186.

34. Wasylishen, A. R., and G. Lozano. 2016. Attenuating the p53 Pathway in Human Cancers: Many Means to the Same End. *Cold Spring Harb Perspect Med* 6.

35. Peng, Y., L. Chen, C. Li, W. Lu, S. Agrawal, and J. Chen. 2001. Stabilization of the MDM2 oncoprotein by mutant p53. *J Biol Chem* 276: 6874-6878.

36. Bond, G. L., W. Hu, E. E. Bond, H. Robins, S. G. Lutzker, N. C. Arva, J. Bargonetti, F. Bartel, H. Taubert, P. Wuerl, K. Onel, L. Yip, S. J. Hwang, L. C. Strong, G. Lozano, and A. J. Levine. 2004. A single nucleotide polymorphism in the MDM2 promoter attenuates the p53 tumor suppressor pathway and accelerates tumor formation in humans. *Cell* 119: 591-602.

37. Wilkening, S., J. L. Bermejo, and K. Hemminki. 2007. MDM2 SNP309 and cancer risk: a combined analysis. *Carcinogenesis* 28: 2262-2267.

38. Wang, W., M. Du, D. Gu, L. Zhu, H. Chu, N. Tong, Z. Zhang, Z. Xu, and M. Wang. 2014. MDM2 SNP309 polymorphism is associated with colorectal cancer risk. *Sci Rep* 4: 4851.

39. Paulin, F. E., M. O'Neill, G. McGregor, A. Cassidy, A. Ashfield, C. W. Ali, A. J. Munro, L. Baker, C. A. Purdie, D. P. Lane, and A. M. Thompson. 2008. MDM2 SNP309 is associated with high grade node positive breast tumours and is in linkage disequilibrium with a novel MDM2 intron 1 polymorphism. *BMC cancer* 8: 281.

40. Ruijs, M. W., M. K. Schmidt, H. Nevanlinna, J. Tommiska, K. Aittomaki, R. Pruntel, S. Verhoef, and L. J. Van't Veer. 2007. The single-nucleotide polymorphism 309 in the MDM2 gene contributes to the Li-Fraumeni syndrome and related phenotypes. *Eur J Hum Genet* 15: 110-114.

41. Marcel, V., E. I. Palmero, P. Falagan-Lotsch, G. Martel-Planche, P. Ashton-Prolla, M. Olivier, R. R. Brentani, P. Hainaut, and M. I. Achatz. 2009. TP53 PIN3 and MDM2 SNP309 polymorphisms as genetic modifiers in the Li-Fraumeni syndrome: impact on age at first diagnosis. *J Med Genet* 46: 766-772.

42. Oliner, J. D., J. A. Pietenpol, S. Thiagalingam, J. Gyuris, K. W. Kinzler, and B. Vogelstein. 1993. Oncoprotein MDM2 conceals the activation domain of tumour suppressor p53. *Nature* 362: 857-860.

43. Oren, M., and V. Rotter. 2010. Mutant p53 gain-of-function in cancer. *Cold Spring Harb Perspect Biol* 2: a001107.

44. Rotter, V. 1983. p53, a transformation-related cellular-encoded protein, can be used as a biochemical marker for the detection of primary mouse tumor cells. *Proc Natl Acad Sci U S A* 80: 2613-2617.

45. Blandino, G., A. J. Levine, and M. Oren. 1999. Mutant p53 gain of function: differential effects of different p53 mutants on resistance of cultured cells to chemotherapy. *Oncogene* 18: 477-485.

46. Matas, D., A. Sigal, P. Stambolsky, M. Milyavsky, L. Weisz, D. Schwartz, N. Goldfinger, and V. Rotter. 2001. Integrity of the N-terminal transcription domain of p53 is required for mutant p53 interference with drug-induced apoptosis. *EMBO J* 20: 4163-4172.

47. Murphy, K. L., A. P. Dennis, and J. M. Rosen. 2000. A gain of function p53 mutant promotes both genomic instability and cell survival in a novel p53-null mammary epithelial cell model. *FASEB J* 14: 2291-2302.

48. Yap, D. B., J. K. Hsieh, S. Zhong, V. Heath, B. Gusterson, T. Crook, and X. Lu. 2004. Ser392 phosphorylation regulates the oncogenic function of mutant p53. *Cancer Res* 64: 4749-4754.

49. Adorno, M., M. Cordenonsi, M. Montagner, S. Dupont, C. Wong, B. Hann, A. Solari, S. Bobisse, M. B. Rondina, V. Guzzardo, A. R. Parenti, A. Rosato, S. Bicciato, A. Balmain, and S. Piccolo. 2009. A Mutant-p53/Smad complex opposes p63 to empower TGFbeta-induced metastasis. *Cell* 137: 87-98.

50. Wang, S. P., W. L. Wang, Y. L. Chang, C. T. Wu, Y. C. Chao, S. H. Kao, A. Yuan, C. W. Lin, S. C. Yang, W. K. Chan, K. C. Li, T. M. Hong, and P. C. Yang. 2009. p53 controls cancer cell invasion by inducing the MDM2-mediated degradation of Slug. *Nat Cell Biol* 11: 694-704.

51. Lang, G. A., T. Iwakuma, Y. A. Suh, G. Liu, V. A. Rao, J. M. Parant, Y. A. Valentin-Vega, T. Terzian, L. C. Caldwell, L. C. Strong, A. K. El-Naggar, and G. Lozano. 2004. Gain of function of a p53 hot spot mutation in a mouse model of Li-Fraumeni syndrome. *Cell* 119: 861-872.

52. Olive, K. P., D. A. Tuveson, Z. C. Ruhe, B. Yin, N. A. Willis, R. T. Bronson, D. Crowley, and T. Jacks. 2004. Mutant p53 gain of function in two mouse models of Li-Fraumeni syndrome. *Cell* 119: 847-860.

53. Bouaoun, L., D. Sonkin, M. Ardin, M. Hollstein, G. Byrnes, J. Zavadil, and M. Olivier. 2016. TP53 Variations in Human Cancers: New Lessons from the IARC TP53 Database and Genomics Data. *Human mutation* 37: 865-876.

54. Vogelstein, B., N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler. 2013. Cancer Genome Landscapes. *Science* 339: 1546-1558.

55. Ognjanovic, S., M. Oliver, T. L. Bergemann, and P. Hainaut. 2011. Sarcomas in TP53 germline mutation carriers: A review of the IARC TP53 database. *Cancer*.

56. Malkin, D. 2011. Li-fraumeni syndrome. *Genes Cancer* 2: 475-484.

57. Li, F. P., and J. F. Fraumeni, Jr. 1969. Rhabdomyosarcoma in children: epidemiologic study and identification of a familial cancer syndrome. *J Natl Cancer Inst* 43: 1365-1373.

58. Li, F. P., and J. F. Fraumeni, Jr. 1969. Soft-tissue sarcomas, breast cancer, and other neoplasms. A familial syndrome? *Annals of internal medicine* 71: 747-752.

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

60. Li, F. P., J. F. Fraumeni, Jr., J. J. Mulvihill, W. A. Blattner, M. G. Dreyfus, M. A. Tucker, and R. W. Miller. 1988. A cancer family syndrome in twenty-four kindreds. *Cancer Res* 48: 5358-5362.

61. Chompret, A., A. Abel, D. Stoppa-Lyonnet, L. Brugieres, S. Pages, J. Feunteun, and C. Bonaiti-Pellie. 2001. Sensitivity and predictive value of criteria for p53 germline mutation screening. *J Med Genet* 38: 43-47.

62. Tinat, J., G. Bougeard, S. Baert-Desurmont, S. Vasseur, C. Martin, E. Bouvignies, O. Caron, B. Bressac-de Paillerets, P. Berthet, C. Dugast, C. Bonaiti-Pellie, D. Stoppa-Lyonnet, and T. Frebourg. 2009. 2009 version of the Chompret criteria for Li Fraumeni syndrome. *J Clin Oncol* 27: e108-109; author reply e110.

63. Eeles, R. A. 1995. Germline mutations in the TP53 gene. *Cancer surveys* 25: 101-124.

64. Birch, J. M., A. L. Hartley, K. J. Tricker, J. Prosser, A. Condie, A. M. Kelsey, M. Harris, P. H. Jones, A. Binchy, D. Crowther, and et al. 1994. Prevalence and diversity of constitutional mutations in the p53 gene among 21 Li-Fraumeni families. *Cancer Res* 54: 1298-1304.

65. Bachinski, L. L., S. E. Olufemi, X. Zhou, C. C. Wu, L. Yip, S. Shete, G. Lozano, C. I. Amos, L. C. Strong, and R. Krahe. 2005. Genetic mapping of a third Li-Fraumeni syndrome predisposition locus to human chromosome 1q23. *Cancer Res* 65: 427-431.

66. Toguchida, J., T. Yamaguchi, S. H. Dayton, R. L. Beauchamp, G. E. Herrera, K. Ishizaki, T. Yamamuro, P. A. Meyers, J. B. Little, M. S. Sasaki, and et al. 1992. Prevalence and spectrum of germline mutations of the p53 gene among patients with sarcoma. *N Engl J Med* 326: 1301-1308.

67. Bell, D. W., J. M. Varley, T. E. Szydlo, D. H. Kang, D. C. Wahrer, K. E. Shannon, M. Lubratovich, S. J. Verselis, K. J. Isselbacher, J. F. Fraumeni, J. M. Birch, F. P. Li, J. E. Garber, and D. A. Haber. 1999. Heterozygous germ line hCHK2 mutations in Li-Fraumeni syndrome. *Science* 286: 2528-2531.

68. Siddiqui, R., K. Onel, F. Facio, K. Nafa, L. R. Diaz, N. Kauff, H. Huang, M. Robson, N. Ellis, and K. Offit. 2005. The TP53 mutational spectrum and frequency of CHEK2*1100delC in Li-Fraumeni-like kindreds. *Fam Cancer* 4: 177-181.

69. Sodha, N., R. S. Houlston, S. Bullock, M. A. Yuille, C. Chu, G. Turner, and R. A. Eeles. 2002. Increasing evidence that germline mutations in CHEK2 do not cause Li-Fraumeni syndrome. *Human mutation* 20: 460-462.

70. Allinen, M., P. Huusko, S. Mantyniemi, V. Launonen, and R. Winqvist. 2001. Mutation analysis of the CHK2 gene in families with hereditary breast cancer. *Br J Cancer* 85: 209-212.

71. Bougeard, G., J. M. Limacher, C. Martin, F. Charbonnier, A. Killian, O. Delattre, M. Longy, P. Jonveaux, J. P. Fricker, D. Stoppa-Lyonnet, J. M. Flaman, and T. Frebourg. 2001. Detection of 11 germline inactivating TP53 mutations and absence of TP63 and HCHK2 mutations in 17 French families with Li-Fraumeni or Li-Fraumeni-like syndrome. *J Med Genet* 38: 253-257.

72. Lee, S. B., S. H. Kim, D. W. Bell, D. C. Wahrer, T. A. Schiripo, M. M. Jorczak, D. C. Sgroi, J. E. Garber, F. P. Li, K. E. Nichols, J. M. Varley, A. K. Godwin, K. M. Shannon, E. Harlow, and D. A. Haber. 2001. Destabilization of CHK2 by a missense mutation associated with Li-Fraumeni Syndrome. *Cancer Res* 61: 8062-8067.

73. Cybulski, C., T. Huzarski, B. Gorski, B. Masojc, M. Mierzejewski, T. Debniak, B. Gliniewicz, J. Matyjasik, E. Zlowocka, G. Kurzawski, A. Sikorski, M. Posmyk, M. Szwiec, R. Czajka, S. A. Narod, and J. Lubinski. 2004. A novel founder CHEK2 mutation is associated with increased prostate cancer risk. *Cancer Res* 64: 2677-2679.

74. Meijers-Heijboer, H., A. van den Ouweland, J. Klijn, M. Wasielewski, A. de Snoo, R. Oldenburg, A. Hollestelle, M. Houben, E. Crepin, M. van Veghel-Plandsoen, F. Elstrodt, C. van Duijn, C. Bartels, C. Meijers, M. Schutte, L. McGuffog, D. Thompson, D. Easton, N. Sodha, S. Seal, R. Barfoot, J. Mangion, J. Chang-Claude, D. Eccles, R. Eeles, D. G. Evans, R. Houlston, V. Murday, S. Narod, T. Peretz, J. Peto, C. Phelan, H.

X. Zhang, C. Szabo, P. Devilee, D. Goldgar, P. A. Futreal, K. L. Nathanson, B. Weber, N. Rahman, M. R. Stratton, and C. H.-B. C. Consortium. 2002. Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nat Genet* 31: 55-59.

75. Offit, K., H. Pierce, T. Kirchhoff, P. Kolachana, B. Rapaport, P. Gregersen, S. Johnson, O. Yossepowitch, H. Huang, J. Satagopan, M. Robson, L. Scheuer, K. Nafa, and N. Ellis. 2003. Frequency of CHEK2*1100delC in New York breast cancer cases and controls. *BMC medical genetics* 4: 1.

76. Evans, D. G., J. M. Birch, and S. A. Narod. 2008. Is CHEK2 a cause of the Li-Fraumeni syndrome? *J Med Genet* 45: 63-64.

77. Macedo, G. S., I. Araujo Vieira, A. P. Brandalize, J. Giacomazzi, E. Inez Palmero, S. Volc, V. Rodrigues Paixao-Cortes, M. Caleffi, M. Silva Alves, M. I. Achatz, P. Hainaut, and P. Ashton-Prolla. 2016. Rare germline variant (rs78378222) in the TP53 3' UTR: Evidence for a new mechanism of cancer predisposition in Li-Fraumeni syndrome. *Cancer genetics* 209: 97-106.

78. Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. 2002. The human genome browser at UCSC. *Genome Res* 12: 996-1006.

79. Minton, J. A., S. E. Flanagan, and S. Ellard. 2011. Mutation surveyor: software for DNA sequence analysis. *Methods Mol Biol* 688: 143-153.

80. Van der Auwera, G. A., M. O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altshuler, S. Gabriel, and M. A. DePristo. 2013. From FastQ data to high confidence

variant calls: the Genome Analysis Toolkit best practices pipeline. *Current protocols in bioinformatics* 43: 11 10 11-33.

81. Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754-1760.

82. McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20: 1297-1303.

83. Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078-2079.

84. Picard.

85. DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43: 491-498.

86. Ng, P. C., and S. Henikoff. 2003. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* 31: 3812-3814.

87. Adzhubei, I. A., S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev. 2010. A method and server for predicting damaging missense mutations. *Nat Methods* 7: 248-249.

88. Peng, G., Y. Fan, and W. Wang. 2014. FamSeq: a variant calling program for family-based sequencing data using graphics processing units. *PLoS Comput Biol* 10: e1003880.

89. Olivier, M., R. Eeles, M. Hollstein, M. A. Khan, C. C. Harris, and P. Hainaut. 2002. The IARC TP53 database: new online mutation analysis and recommendations to users. *Human mutation* 19: 607-614.

90. Lek, M., K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings, T. Tukiainen, D. P. Birnbaum, J. A. Kosmicki, L. E. Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, J. Berghout, D. N. Cooper, N. Deflaux, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M. I. Kurki, A. L. Moonshine, P. Natarajan, L. Orozco, G. M. Peloso, R. Poplin, M. A. Rivas, V. Ruano-Rubio, S. A. Rose, D. M. Ruderfer, K. Shakir, P. D. Stenson, C. Stevens, B. P. Thomas, G. Tiao, M. T. Tusie-Luna, B. Weisburd, H. H. Won, D. Yu, D. M. Altshuler, D. Ardissino, M. Boehnke, J. Danesh, S. Donnelly, R. Elosua, J. C. Florez, S. B. Gabriel, G. Getz, S. J. Glatt, C. M. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M. I. McCarthy, D. McGovern, R. McPherson, B. M. Neale, A. Palotie, S. M. Purcell, D. Saleheen, J. M. Scharf, P. Sklar, P. F. Sullivan, J. Tuomilehto, M. T. Tsuang, H. C. Watkins, J. G. Wilson, M. J. Daly, D. G. MacArthur, and C. Exome Aggregation. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536: 285-291.

91. Exome Variant Server, N. G. E. S. P. E., Seattle, WA (URL: http://evs.gs.washington.edu/EVS/).

92. de Andrade, K. C., L. Mirabello, D. R. Stewart, E. Karlins, R. Koster, M. Wang, S. M. Gapstur, M. M. Gaudet, N. D. Freedman, M. T. Landi, N. Lemonnier, P. Hainaut, S. A. Savage, and M. I. Achatz. 2017. Higher-than-expected population prevalence of potentially pathogenic germline TP53 variants in individuals unselected for cancer history. *Human mutation* 38: 1723-1730.

93. Amadou, A., M. I. Waddington Achatz, and P. Hainaut. 2018. Revisiting tumor patterns and penetrance in germline TP53 mutation carriers: temporal phases of Li-Fraumeni syndrome. *Curr Opin Oncol* 30: 23-29.

94. Genomes Project, C., A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, and G. R. Abecasis. 2015. A global reference for human genetic variation. *Nature* 526: 68-74.

95. Gottschall, P. E., and M. D. Howell. 2015. ADAMTS expression and function in central nervous system injury and disorders. *Matrix biology : journal of the International Society for Matrix Biology* 44-46: 70-76.

96. Bondeson, J., S. Wainwright, C. Hughes, and B. Caterson. 2008. The regulation of the ADAMTS4 and ADAMTS5 aggrecanases in osteoarthritis: a review. *Clinical and experimental rheumatology* 26: 139-145.

97. Choubey, D., R. Deka, and S. M. Ho. 2008. Interferon-inducible IFI16 protein in human cancers and autoimmune diseases. *Front Biosci* 13: 598-608.

98. Anholt, R. R. 2014. Olfactomedin proteins: central players in development and disease. *Frontiers in cell and developmental biology* 2: 6.

99. Naji, L., D. Pacholsky, and P. Aspenstrom. 2011. ARHGAP30 is a Wrch-1-interacting protein involved in actin dynamics and cell adhesion. *Biochemical and biophysical research communications* 409: 96-102.

100. Wang, J., J. Qian, Y. Hu, X. Kong, H. Chen, Q. Shi, L. Jiang, C. Wu, W. Zou, Y. Chen, J. Xu, and J. Y. Fang. 2014. ArhGAP30 promotes p53 acetylation and function in colorectal cancer. *Nat Commun* 5: 4735.

101. Landrum, M. J., J. M. Lee, G. R. Riley, W. Jang, W. S. Rubinstein, D. M. Church, and D. R. Maglott. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* 42: D980-985.

102. Fu, Y., Z. Liu, S. Lou, J. Bedford, X. J. Mu, K. Y. Yip, E. Khurana, and M. Gerstein. 2014. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol* 15: 480.

103. Ye, K., M. H. Schulz, Q. Long, R. Apweiler, and Z. Ning. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25: 2865-2871.

104. Zhu, M., A. C. Need, Y. Han, D. Ge, J. M. Maia, Q. Zhu, E. L. Heinzen, E. T. Cirulli, K. Pelak, M. He, E. K. Ruzzo, C. Gumbs, A. Singh, S. Feng, K. V. Shianna, and D. B. Goldstein. 2012. Using ERDS to infer copy-number variants in high-coverage genomes. *Am J Hum Genet* 91: 408-421.

105. Denissenko, M. F., J. X. Chen, M. S. Tang, and G. P. Pfeifer. 1997. Cytosine methylation determines hot spots of DNA damage in the human P53 gene. *Proc Natl Acad Sci U S A* 94: 3893-3898.

106. Doyle, B., C. O'Riain, and K. Appleton. 2011. Pyrosequencing of DNA extracted from formalin-fixed paraffin-embedded tissue. *Methods Mol Biol* 724: 181-190.

107. Barretina, J., G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehar, G. V. Kryukov, D. Sonkin, A. Reddy, M. Liu, L. Murray, M. F. Berger, J. E. Monahan, P. Morais, J. Meltzer, A. Korejwa, J. Jane-Valbuena, F. A. Mapa, J. Thibault, E. Bric-Furlong, P. Raman, A. Shipway, I. H. Engels, J. Cheng, G. K. Yu, J. Yu, P. Aspesi, Jr., M. de Silva, K. Jagtap, M. D. Jones, L. Wang, C. Hatton, E. Palescandolo, S. Gupta, S. Mahan, C. Sougnez, R. C. Onofrio, T. Liefeld, L. MacConaill, W. Winckler, M. Reich, N. Li, J. P. Mesirov, S. B. Gabriel, G. Getz, K. Ardlie, V. Chan, V. E. Myer, B. L. Weber, J. Porter, M. Warmuth, P. Finan, J. L. Harris, M. Meyerson, T. R. Golub, M. P. Morrissey, W. R. Sellers, R. Schlegel, and L. A. Garraway. 2012. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483: 603-607.

108. Lin, Y. C., M. Boone, L. Meuris, I. Lemmens, N. Van Roy, A. Soete, J. Reumers, M. Moisse, S. Plaisance, R. Drmanac, J. Chen, F. Speleman, D. Lambrechts, Y. Van de Peer, J. Tavernier, and N. Callewaert. 2014. Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology manipulations. *Nat Commun* 5: 4767.

109. Donehower, L. A., M. Harvey, B. L. Slagle, M. J. McArthur, C. A. Montgomery, Jr., J. S. Butel, and A. Bradley. 1992. Mice deficient for p53 are developmentally normal but susceptible to spontaneous tumours. *Nature* 356: 215-221.

110. Cazier, J. B., and I. Tomlinson. 2010. General lessons from large-scale studies to identify human cancer predisposition genes. *J Pathol* 220: 255-262.

111. Pleasance, E. D., R. K. Cheetham, P. J. Stephens, D. J. McBride, S. J. Humphray, C. D. Greenman, I. Varela, M. L. Lin, G. R. Ordonez, G. R. Bignell, K. Ye, J. Alipaz, M. J. Bauer, D. Beare, A. Butler, R. J. Carter, L. Chen, A. J. Cox, S. Edkins, P. I. Kokko-Gonzales, N. A. Gormley, R. J. Grocock, C. D. Haudenschild, M. M. Hims, T. James, M. Jia, Z. Kingsbury, C. Leroy, J. Marshall, A. Menzies, L. J. Mudie, Z. Ning, T. Royce, O. B. Schulz-Trieglaff, A. Spiridou, L. A. Stebbings, L. Szajkowski, J. Teague, D. Williamson, L. Chin, M. T. Ross, P. J. Campbell, D. R. Bentley, P. A. Futreal, and M. R. Stratton. 2010. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* 463: 191-196.

112. Lawrence, M. S., P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. A. Roberts, A. Kiezun, P. S. Hammerman, A. McKenna, Y. Drier, L. Zou, A. H. Ramos, T. J. Pugh, N. Stransky, E. Helman, J. Kim, C. Sougnez, L. Ambrogio, E. Nickerson, E. Shefler, M. L. Cortes, D. Auclair, G. Saksena, D. Voet, M. Noble, D. DiCara, P. Lin, L. Lichtenstein, D. I. Heiman, T. Fennell, M. Imielinski, B. Hernandez, E. Hodis, S. Baca, A. M. Dulak, J. Lohr, D. A. Landau, C. J. Wu, J. Melendez-Zajgla, A. Hidalgo-Miranda, A. Koren, S. A. McCarroll, J. Mora, B. Crompton, R. Onofrio, M. Parkin, W. Winckler, K. Ardlie, S. B. Gabriel, C. W. M. Roberts, J. A. Biegel, K. Stegmaier, A. J. Bass, L. A. Garraway, M. Meyerson, T. R. Golub, D. A. Gordenin, S. Sunyaev, E. S. Lander, and G. Getz. 2013. Mutational

heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499: 214-218.

113. Watson, I. R., K. Takahashi, P. A. Futreal, and L. Chin. 2013. Emerging patterns of somatic mutations in cancer. *Nat Rev Genet* 14: 703-718.

114. Mardis, E. R., L. Ding, D. J. Dooling, D. E. Larson, M. D. McLellan, K. Chen, D. C. Koboldt, R. S. Fulton, K. D. Delehaunty, S. D. McGrath, L. A. Fulton, D. P. Locke, V. J. Magrini, R. M. Abbott, T. L. Vickery, J. S. Reed, J. S. Robinson, T. Wylie, S. M. Smith, L. Carmichael, J. M. Eldred, C. C. Harris, J. Walker, J. B. Peck, F. Du, A. F. Dukes, G. E. Sanderson, A. M. Brummett, E. Clark, J. F. McMichael, R. J. Meyer, J. K. Schindler, C. S. Pohl, J. W. Wallis, X. Shi, L. Lin, H. Schmidt, Y. Tang, C. Haipek, M. E. Wiechert, J. V. Ivy, J. Kalicki, G. Elliott, R. E. Ries, J. E. Payton, P. Westervelt, M. H. Tomasson, M. A. Watson, J. Baty, S. Heath, W. D. Shannon, R. Nagarajan, D. C. Link, M. J. Walter, T. A. Graubert, J. F. DiPersio, R. K. Wilson, and T. J. Ley. 2009. Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med* 361: 1058-1066.

115. Chatterjee, A., E. J. Rodger, and M. R. Eccles. 2017. Epigenetic drivers of tumourigenesis and cancer metastasis. *Semin Cancer Biol*.

116. Le Deley, M. C., O. Delattre, K. L. Schaefer, S. A. Burchill, G. Koehler, P. C. Hogendoorn, T. Lion, C. Poremba, J. Marandet, S. Ballet, G. Pierron, S. C. Brownhill, M. Nesslbock, A. Ranft, U. Dirksen, O. Oberlin, I. J. Lewis, A. W. Craft, H. Jurgens, and H. Kovar. 2010. Impact of EWS-ETS fusion type on disease progression in Ewing's sarcoma/peripheral primitive neuroectodermal tumor: prospective results from the cooperative Euro-E.W.I.N.G. 99 trial. *J Clin Oncol* 28: 1982-1988.

117. Chen, X., A. Bahrami, A. Pappo, J. Easton, J. Dalton, E. Hedlund, D. Ellison, S. Shurtleff, G. Wu, L. Wei, M. Parker, M. Rusch, P. Nagahawatte, J. Wu, S. Mao, K. Boggs, H. Mulder, D. Yergeau, C. Lu, L. Ding, M. Edmonson, C. Qu, J. Wang, Y. Li, F. Navid, N. C. Daw, E. R. Mardis, R. K. Wilson, J. R. Downing, J. Zhang, M. A. Dyer, and P. St. Jude Children's Research Hospital-Washington University Pediatric Cancer Genome. 2014. Recurrent somatic structural variations contribute to tumorigenesis in pediatric osteosarcoma. *Cell Rep* 7: 104-112.

118. Kovac, M., C. Blattmann, S. Ribi, J. Smida, N. S. Mueller, F. Engert, F. Castro-Giner, J. Weischenfeldt, M. Kovacova, A. Krieg, D. Andreou, P. U. Tunn, H. R. Durr, H. Rechl, K. D. Schaser, I. Melcher, S. Burdach, A. Kulozik, K. Specht, K. Heinimann, S. Fulda, S. Bielack, G. Jundt, I. Tomlinson, J. O. Korbel, M. Nathrath, and D. Baumhoer. 2015. Exome sequencing of osteosarcoma reveals mutation signatures reminiscent of BRCA deficiency. *Nat Commun* 6: 8940.

119. Otano-Joos, M., G. Mechtersheimer, S. Ohl, K. K. Wilgenbus, W. Scheurlen, T. Lehnert, F. Willeke, H. F. Otto, P. Lichter, and S. Joos. 2000. Detection of chromosomal imbalances in leiomyosarcoma by comparative genomic hybridization and interphase cytogenetics. *Cytogenetics and cell genetics* 90: 86-92.

120. Kawaguchi, K., Y. Oda, T. Saito, T. Takahira, H. Yamamoto, S. Tamiya, Y. Iwamoto, and M. Tsuneyoshi. 2005. Genetic and epigenetic alterations of the PTEN gene in soft tissue sarcomas. *Human pathology* 36: 357-363.

121. Loeb, L. A. 2001. A mutator phenotype in cancer. *Cancer Res* 61: 3230-3239.

122. Bielas, J. H., K. R. Loeb, B. P. Rubin, L. D. True, and L. A. Loeb. 2006. Human cancers express a mutator phenotype. *Proc Natl Acad Sci U S A* 103: 18238-18242.

123. Donehower, L. A., M. Harvey, H. Vogel, M. J. McArthur, C. A. Montgomery, Jr., S. H. Park, T. Thompson, R. J. Ford, and A. Bradley. 1995. Effects of genetic background on tumorigenesis in p53-deficient mice. *Mol Carcinog* 14: 16-22.

124. Vos, J. A., S. L. Abbondanzo, C. L. Barekman, J. W. Andriko, M. Miettinen, and N. S. Aguilera. 2005. Histiocytic sarcoma: a study of five cases including the histiocyte marker CD163. *Mod Pathol* 18: 693-704.

125. Vanden Berghe, T., P. Hulpiau, L. Martens, R. E. Vandenbroucke, E. Van Wonterghem, S. W. Perry, I. Bruggeman, T. Divert, S. M. Choi, M. Vuylsteke, V. I. Shestopalov, C. Libert, and P. Vandenabeele. 2015. Passenger Mutations Confound Interpretation of All Genetically Modified Congenic Mice. *Immunity* 43: 200-209.

126. Roadmap Epigenomics, C., A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y.-C. C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K.-H. H. Farh, S. Feizi, R. Karlic, A.-R. R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A.

Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L.-H. H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, and M. Kellis. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* 518: 317-330.

127. Rivlin, N., R. Brosh, M. Oren, and V. Rotter. 2011. Mutations in the p53 Tumor Suppressor Gene: Important Milestones at the Various Steps of Tumorigenesis. *Genes Cancer* 2: 466-474.

128. Venkitaraman, A. R. 2002. Cancer susceptibility and the functions of BRCA1 and BRCA2. *Cell* 108: 171-182.

129. Cibulskis, K., M. S. Lawrence, S. L. Carter, A. Sivachenko, D. Jaffe, C. Sougnez, S. Gabriel, M. Meyerson, E. S. Lander, and G. Getz. 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* 31: 213-219.

130. Koboldt, D. C., Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, L. Lin, C. A. Miller, E. R. Mardis, L. Ding, and R. K. Wilson. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* 22: 568-576.

131. Larson, D. E., C. C. Harris, K. Chen, D. C. Koboldt, T. E. Abbott, D. J. Dooling, T. J. Ley, E. R. Mardis, R. K. Wilson, and L. Ding. 2012. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* 28: 311-317.

132. Fan, Y., L. Xi, D. S. Hughes, J. Zhang, J. Zhang, P. A. Futreal, D. A. Wheeler, and W. Wang. 2016. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol* 17: 178.

133. Wang, Q., P. Jia, F. Li, H. Chen, H. Ji, D. Hucks, K. B. Dahlman, W. Pao, and Z. Zhao. 2013. Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. *Genome Med* 5: 91.

134. Kroigard, A. B., M. Thomassen, A. V. Laenkholm, T. A. Kruse, and M. J. Larsen. 2016. Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data. *PLoS One* 11: e0151664.

135. Cai, L., W. Yuan, Z. Zhang, L. He, and K. C. Chou. 2016. In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. *Sci Rep* 6: 36540.

136. Wang, K., M. Li, and H. Hakonarson. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38: e164.

137. Choi, Y., and A. P. Chan. 2015. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* 31: 2745-2747.

138. Wu, J. N., and C. W. Roberts. 2013. ARID1A mutations in cancer: another epigenetic tumor suppressor? *Cancer Discov* 3: 35-43.

139. Vogelstein, B., and K. W. Kinzler. 2004. Cancer genes and the pathways they control. *Nat Med* 10: 789-799.

140. Vandin, F., E. Upfal, and B. J. Raphael. 2012. De novo discovery of mutated driver pathways in cancer. *Genome Res* 22: 375-385.

141. Varela, I., P. Tarpey, K. Raine, D. Huang, C. K. Ong, P. Stephens, H. Davies, D. Jones, M. L. Lin, J. Teague, G. Bignell, A. Butler, J. Cho, G. L. Dalgliesh, D. Galappaththige, C. Greenman, C. Hardy, M. Jia, C. Latimer, K. W. Lau, J. Marshall, S. McLaren, A. Menzies, L. Mudie, L. Stebbings, D. A. Largaespada, L. F. Wessels, S. Richard, R. J. Kahnoski, J. Anema, D. A. Tuveson, P. A. Perez-Mancera, V. Mustonen, A. Fischer, D. J. Adams, A. Rust, W. Chan-on, C. Subimerb, K. Dykema, K. Furge, P. J. Campbell, B. T. Teh, M. R. Stratton, and P. A. Futreal. 2011. Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature* 469: 539-542.

142. Nargund, A. M., C. G. Pham, Y. Dong, P. I. Wang, H. U. Osmangeyoglu, Y. Xie, O. Aras, S. Han, T. Oyama, S. Takeda, C. E. Ray, Z. Dong, M. Berge, A. A. Hakimi, S. Monette, C. L. Lekaye, J. A. Koutcher, C. S. Leslie, C. J. Creighton, N. Weinhold, W. Lee, S. K. Tickoo, Z. Wang, E. H. Cheng, and J. J. Hsieh. 2017. The SWI/SNF Protein PBRM1 Restrains VHL-Loss-Driven Clear Cell Renal Cell Carcinoma. *Cell Rep* 18: 2893-2906.

143. TCGA-Research-Network. 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455: 1061-1068.

144. Guan, J., R. Gupta, and F. V. Filipp. 2015. Cancer systems biology of TCGA SKCM: efficient detection of genomic drivers in melanoma. *Sci Rep* 5: 7857.

145. Brohl, A. S., E. Kahen, S. J. Yoder, J. K. Teer, and D. R. Reed. 2017. The genomic landscape of malignant peripheral nerve sheath tumors: diverse drivers of Ras pathway activation. *Sci Rep* 7: 14992.

146. Futreal, P. A., L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton. 2004. A census of human cancer genes. *Nat Rev Cancer* 4: 177-183.

147. Hanks, S., K. Coleman, S. Reid, A. Plaja, H. Firth, D. Fitzpatrick, A. Kidd, K. Mehes, R. Nash, N. Robin, N. Shannon, J. Tolmie, J. Swansbury, A. Irrthum, J. Douglas, and N. Rahman. 2004. Constitutional aneuploidy and cancer predisposition caused by biallelic mutations in BUB1B. *Nat Genet* 36: 1159-1161.

148. Yamamoto, Y., H. Matsuyama, Y. Chochi, M. Okuda, S. Kawauchi, R. Inoue, T. Furuya, A. Oga, K. Naito, and K. Sasaki. 2007. Overexpression of BUBR1 is associated with chromosomal instability in bladder cancer. *Cancer Genet Cytogenet* 174: 42-47.

149. Stolz, A., N. Ertych, and H. Bastians. 2011. Tumor suppressor CHK2: regulator of DNA damage response and mediator of chromosomal stability. *Clin Cancer Res* 17: 401-405.

150. Stolz, A., N. Ertych, A. Kienitz, C. Vogel, V. Schneider, B. Fritz, R. Jacob, G. Dittmar, W. Weichert, I. Petersen, and H. Bastians. 2010. The CHK2-BRCA1 tumour suppressor pathway ensures chromosomal stability in human somatic cells. *Nat Cell Biol* 12: 492-499.

151. Grigorova, M., J. M. Staines, H. Ozdag, C. Caldas, and P. A. Edwards. 2004. Possible causes of chromosome instability: comparison of chromosomal abnormalities in

cancer cell lines with mutations in BRCA1, BRCA2, CHK2 and BUB1. *Cytogenetic and genome research* 104: 333-340.

152. Fan, G., L. Sun, P. Shan, X. Zhang, J. Huan, X. Zhang, D. Li, T. Wang, T. Wei, X. Zhang, X. Gu, L. Yao, Y. Xuan, Z. Hou, Y. Cui, L. Cao, X. Li, S. Zhang, and C. Wang. 2015. Loss of KLF14 triggers centrosome amplification and tumorigenesis. *Nat Commun* 6: 8450.

153. Domon-Dell, C., A. Schneider, V. Moucadel, E. Guerin, D. Guenot, S. Aguillon, I. Duluc, E. Martin, J. Iovanna, J. F. Launay, B. Duclos, M. P. Chenard, C. Meyer, P. Oudet, M. Kedinger, M. P. Gaub, and J. N. Freund. 2003. Cdx1 homeobox gene during human colon cancer progression. *Oncogene* 22: 7913-7921.

154. Hryniuk, A., S. Grainger, J. G. Savory, and D. Lohnes. 2014. Cdx1 and Cdx2 function as tumor suppressors. *J Biol Chem* 289: 33343-33354.

155. Davies, O. R., C. Y. Lin, A. Radzisheuskaya, X. Zhou, J. Taube, G. Blin, A. Waterhouse, A. J. Smith, and S. Lowell. 2013. Tcf15 primes pluripotent cells for differentiation. *Cell Rep* 3: 472-484.

156. Aryee, M. J., A. E. Jaffe, H. Corrada-Bravo, C. Ladd-Acosta, A. P. Feinberg, K. D. Hansen, and R. A. Irizarry. 2014. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 30: 1363-1369.

157. Kwabi-Addo, B., D. Giri, K. Schmidt, K. Podsypanina, R. Parsons, N. Greenberg, and M. Ittmann. 2001. Haploinsufficiency of the Pten tumor suppressor gene promotes prostate cancer progression. *Proc Natl Acad Sci U S A* 98: 11563-11568.

158. Berger, A. H., A. G. Knudson, and P. P. Pandolfi. 2011. A continuum model for tumour suppression. *Nature* 476: 163-169.

159. Rota, R., R. Ciarapica, L. Miele, and F. Locatelli. 2012. Notch signaling in pediatric soft tissue sarcomas. *BMC medicine* 10: 141.

160. Tao, J., M. M. Jiang, L. Jiang, J. S. Salvo, H. C. Zeng, B. Dawson, T. K. Bertin, P. H. Rao, R. Chen, L. A. Donehower, F. Gannon, and B. H. Lee. 2014. Notch activation as a driver of osteogenic sarcoma. *Cancer Cell* 26: 390-401.

161. Nowell, C. S., and F. Radtke. 2017. Notch as a tumour suppressor. *Nat Rev Cancer* 17: 145-159.

162. Fortin, J. P., T. J. Triche, Jr., and K. D. Hansen. 2017. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics* 33: 558-560.

163. Romanoski, C. E., C. K. Glass, H. G. Stunnenberg, L. Wilson, and G. Almouzni. 2015. Epigenomics: Roadmap for regulation. *Nature* 518: 314-316.

164. Sengelaub, C. A., K. Navrazhina, J. B. Ross, N. Halberg, and S. F. Tavazoie. 2016. PTPRN2 and PLCbeta1 promote metastatic breast cancer cell migration through PI(4,5)P2-dependent actin remodeling. *EMBO J* 35: 62-76.

165. Hishida, M., S. Nomoto, Y. Inokawa, M. Hayashi, M. Kanda, Y. Okamura, Y. Nishikawa, C. Tanaka, D. Kobayashi, S. Yamada, G. Nakayama, T. Fujii, H. Sugimoto, M. Koike, M. Fujiwara, S. Takeda, and Y. Kodera. 2013. Estrogen receptor 1 gene as a tumor suppressor gene in hepatocellular carcinoma detected by triple-combination array analysis. *Int J Oncol* 43: 88-94.

166. Jeong, H., J. Kim, Y. Lee, J. H. Seo, S. R. Hong, and A. Kim. 2014. Neuregulin-1 induces cancer stem cell characteristics in breast cancer cell lines. *Oncol Rep* 32: 1218-1224.

167. Zhu, X. X., Y. W. Yan, C. Z. Ai, S. Jiang, S. S. Xu, M. Niu, X. Z. Wang, G. S. Zhong, X. F. Lu, Y. Xue, S. Tian, G. Li, S. Tang, and Y. Z. Jiang. 2017. Jarid2 is essential for the maintenance of tumor initiating cells in bladder cancer. *Oncotarget* 8: 24483-24490.

168. Dai, Y., M. Wang, H. Wu, M. Xiao, H. Liu, and D. Zhang. 2017. Loss of FOXN3 in colon cancer activates beta-catenin/TCF signaling and promotes the growth and migration of cancer cells. *Oncotarget* 8: 9783-9793.

169. Hooi, C. F., C. Blancher, W. Qiu, I. M. Revet, L. H. Williams, M. L. Ciavarella, R. L. Anderson, E. W. Thompson, A. Connor, W. A. Phillips, and I. G. Campbell. 2006. ST7-mediated suppression of tumorigenicity of prostate cancer cells is characterized by remodeling of the extracellular matrix. *Oncogene* 25: 3924-3933.

170. Zhu, Q. Q., C. Ma, Q. Wang, Y. Song, and T. Lv. 2016. The role of TWIST1 in epithelial-mesenchymal transition and cancers. *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine* 37: 185-197.

171. Mariotti, S., I. Barravecchia, C. Vindigni, A. Pucci, M. Balsamo, R. Libro, V. Senchenko, A. Dmitriev, E. Jacchetti, M. Cecchini, F. Roviello, M. Lai, V. Broccoli, M. Andreazzoli, C. M. Mazzanti, and D. Angeloni. 2016. MICAL2 is a novel human cancer gene controlling mesenchymal to epithelial transition involved in cancer growth and invasion. *Oncotarget* 7: 1808-1825.

172. Barretina, J., B. S. Taylor, S. Banerji, A. H. Ramos, M. Lagos-Quintana, P. L. Decarolis, K. Shah, N. D. Socci, B. A. Weir, A. Ho, D. Y. Chiang, B. Reva, C. H. Mermel, G. Getz, Y.

Antipin, R. Beroukhim, J. E. Major, C. Hatton, R. Nicoletti, M. Hanna, T. Sharpe, T. J. Fennell, K. Cibulskis, R. C. Onofrio, T. Saito, N. Shukla, C. Lau, S. Nelander, S. J. Silver, C. Sougnez, A. Viale, W. Winckler, R. G. Maki, L. A. Garraway, A. Lash, H. Greulich, D. E. Root, W. R. Sellers, G. K. Schwartz, C. R. Antonescu, E. S. Lander, H. E. Varmus, M. Ladanyi, C. Sander, M. Meyerson, and S. Singer. 2010. Subtype-specific genomic alterations define new targets for soft-tissue sarcoma therapy. *Nat Genet* 42: 715-721.

14 Vita

Justin Wai-Chun Wong was born in Boston, Massachusetts. After completing his work at Belmont High School, in Belmont, Massachusetts in 2004, he entered Franklin W. Olin College of Engineering, where he graduated with a Bachelor of Science with a focus in biomedical engineering and tissue engineering. For the next two years, he worked as an application engineer at Massachusetts General Hospital to scale up a circulating tumor cell diagnostic. In August of 2010, he entered the University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences.