THE UNIVERSITY OF CHICAGO

DEFINING LEUKEMOGENIC MECHANISMS IN HEREDITARY HEMATOPOIETIC MALIGNANCIES

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES

AND THE PRITZKER SCHOOL OF MEDICINE

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF HUMAN GENETICS

BY

MICHAEL WILLIAM DRAZER

CHICAGO, ILLINOIS

DECEMBER 2021

Dedication

To my family and my patients.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

5’ UTR	5’ Untranslated Region
AFP	Alpha fetoprotein
AL	Acute Leukemia
ALL	Acute Lymphocytic Leukemia
AML	Acute Myeloid Leukemia
B-ALL	B-cell acute lymphoblastic leukemia
bFGF	Basic fibroblast growth factor
BM	Bone marrow
BMP4	Bone morphogenic protein-4
CBC	Complete blood count
CH	Clonal hematopoiesis
CLIA	Clinical Laboratory Improvement Amendments
CLL	Chronic Lymphocytic Leukemia
CML	Chronic Myeloid Leukemia
CNV	Copy Number Variant
dsDNA	Double-stranded DNA
EB	Embryoid body
EBV	Epstein Barr Virus
Flt3L	fms-like tyrosine kinase 3 ligand
GATK	Genome Analysis Toolkit
gDNA	Genomic deoxyribonucleic acid
gRNA	Guide RNA

HM	Hematopoietic malignancy
HHM	Hereditary Hematopoietic malignancy
HSPC	Hematopoietic stem and progenitor cell
HT/HHM	Hereditary thrombocytopenia/Hereditary Hematopoietic Malignancy syndrome
IL-6	Interleukin-6
iPSC	Induced pluripotent stem cell
iPSC-HSPCs	Induced pluripotent stem cell-derived hematopoietic stem and progenitor cells
KO	Knockout
LCLs	Lymphoblastoid cell lines
LFS	Li-Fraumeni Syndrome
LP/P	Likely pathogenic/pathogenic
MDS	Myelodysplastic syndrome
MEFs	Murine embryonic fibroblasts
MRD	Matched related donor
MSC	Mesenchymal stromal cell
NGS	Next generation sequencing
PB	Peripheral blood
PCR	Polymerase chain reaction
pLoF	Predicted loss of function
RUNX1db	The RUNX1 Database
SCF	Stem cell factor

ssODN	Single stranded oligonucleotide
T-ALL	T-cell acute lymphoblastic leukemia
TF	Transcription factor
TPO	Thrombopoietin
TMB/Mb	Tumor mutation burden/megabase
VAF	Variant allele frequency
VEGF	Vascular endothelial growth factor
WHO	World Health Organization
Wnt3a	Wingless-type MMTV integration site-3a

Acknowledgements

First, I would like to thank my advisor, Dr. Lucy Godley, for her mentorship during my training and my transition to my faculty position at UChicago. Her support has opened numerous doors for me, and I have done my best to make the most of each of these opportunities. Lucy has challenged, and frequently upended, accepted “norms” in hematology and clinical genetics, and these experiences have molded my own approach to science and my career development.

I would also like to thank my patients and my colleagues around the world who have contributed to this research by donating numerous samples, time, patience, and ideas. My patients recognize the “gaps” in the research – you keep us grounded and focused on research that matters to you and your families.

I also would like to thank my thesis committee: Drs. Yoav Gilad, Marcelo Nobrega, Jorge Di Paola, and Amittha Wickrema, for their time and thoughtful support during this period. In particular, I appreciate Dr. Gilad, Jonathan Burnett, and Dr. Ittai Eres, for teaching me “the basics” of the generation, maintenance, and characterization of patient-derived induced pluripotent stem cells. This training has served as the basis for the early stage of my own research program, has informed my future projects, and has spawned exciting and promising collaborations with my colleagues outside of Chicago. I would also like to thank Dr. Tim Graubert, who has served on my mentorship committee since my second year of Hematology/Oncology fellowship and who has provided valuable advice at numerous junctures during my early career.

I owe a tremendous sense of gratitude to my external collaborators, including Drs. Deborah French and Paul Gadue at the Children’s Hospital of Philadelphia, who allowed

me to pursue an external rotation in their groups during my doctoral training. Their collegiality provided me with invaluable training that will impact my science for years to come. In addition, I would like to thank their trainees and staff, including Dr. Sarah Borst, Chintan Jobaliya, and Dr. Jean Ann Maguire, who worked with me during my rotation and taught me the “on the ground” skills that I have implemented successfully in Chicago.

I have been fortunate to establish collaborations and friendships with colleagues invested in translational work involving the hereditary hematopoietic malignancy syndromes, including Drs. Anna Brown, Hamish Scott, Claire Homan, Ulla Wartiovaara-Kautto, Kiran Tawana, Paul Liu, Kai Yu, Christopher Hahn, Amy Trottier, Simone Feurstein, Marcela Cavalcante de Andrade Silva, and Pamela Mora. Much of the work in this thesis was performed during (and despite) the COVID19 pandemic, a testament to the flexibility and strength of our collaborations.

I would like to thank all of the current and past members of the Godley group, including: Shawn Albert, Stephen Arnovitz, Dr. Kristina Bigelow, Dr. Jane Churpek, Anastasia Hains, Aelin Kim, Janet Lepore, Maya Lewinsohn, Dr. Christopher Mariani, Dr. Rafael Marquez, Connie Phung, Matthew Pozsgai, Dr. Sakshi Uppal, Dr. Aparna Vasanthakumar, Dr. Allison West, and Art Wolin. Each of you have proven to be invaluable colleagues during my training.

I would especially like to thank my friends, colleagues, and classmates in the Human Genetics ecosystem. My cohort classmates, Selene Clay and Kathryn Farris, were invaluable in their support as we navigated the first-year curriculum together; Linsin Smith and Dr. Briana Mittleman also served as my “honorary” first-year cohort mates as I started my first-year studies off cycle to account for the peculiarities of my clinical training

schedule. I will always appreciate Drs. Carole Ober and Anna Di Rienzo for supporting my idea that I would benefit from rigorous doctoral training when I approached them while I was still a clinical fellow in Hematology/Oncology, but with an eye toward the distant future. I also owe a tremendous sense of gratitude to Candice Lewis and Sue Levison, both of whom have been adaptable as they navigate the administrative nightmares associated with a doctoral candidate who is also a physician a few blocks over, but in an entirely different administrative landscape.

I will always be grateful for the foundations that support the work of physician scientists, in particular those of us at the beginning of our careers. They understand the unique sacrifices that are required to pursue this work, and I will always appreciate their efforts to blunt some of the threats against our phenotype. I have benefited tremendously from the support of the Damon Runyon Cancer Research Foundation, the Evans P. Evans Foundation, the Cancer Research Foundation, the Cooney Research Foundation, the Michael Reese Society, the University of Chicago Comprehensive Cancer Center, and the National Institutes of Health.

I am especially thankful for Kelsey McNeely, the first member of my own research group, who has proven to be patient and flexible as she has learned new techniques in murine modeling, bioinformatics, and cellular modeling. She has proven to be highly adaptable during the COVID19 pandemic, and I have always appreciated her flexibility whenever I start a request with “I know you didn’t sign up for this, but…”. In addition, I would like to thank the trainees who have spent time in my group during the early part of my faculty position, including Dalein Azaii Calderon, Maria Acevedo Mendez, Dr. Frank Wen, Ava Hollis, Shaili Smith, and Rachel Hodge.

I also owe a tremendous thank you to my science teachers in Hebron, Indiana: Dan Pastrick, Cindy Brown, Terry Bailey, and Dr. Christine Duttlinger. Each of you sparked a love for the sciences in me and allowed me the flexibility to carve my own path, even if it required substantial amounts of extra effort on your part (and obtaining texts from more ‘well to do’ school districts). You each provided me with room to run so that I was well prepared the minute opportunities presented themselves to me. I also owe Dr. Carol Bloom, my coach in 4H livestock judging in Porter County, Indiana, a tremendous thank you for her early instruction and mentorship. A trailblazer in her own right, she encouraged my early interest in science, medicine, and genetics, and provided me with invaluable advice, wisdom, and time during a formative time period in my life.

Finally, words cannot describe my gratitude for John, Mary, Mariah, Kevin, Megan, Tom, Maggie, and Xavier – you each understand the sacrifices we each have made so that I can pursue this path as a physician scientist, and for that I will always be grateful.

Abstract

Hereditary hematopoietic malignancies (HHMs) were once thought to be exceedingly rare, but the development of relatively inexpensive next generation sequencing (NGS) in the early 21st century led to a dramatic increase in the number of recognized HHM syndromes. Now, HHMs are an accepted entity in the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia, and a substantial proportion of individuals are recognized to be at risk for these syndromes.

However, a number of knowledge gaps remain in the field. First, no large-scale natural history study has been completed for individuals with HHMs. Second, individuals with HHMs experience up to a 40% lifetime risk for blood cancers such as acute leukemia (AL) and myelodysplastic syndrome (MDS), but the molecular and genetic steps driving leukemogenesis are not well understood. Third, the field's ability to model HHMs has been limited by a lack of high-fidelity models. Finally, many individuals with an HHM phenotype who undergo clinical-grade sequencing do not have a clear germline genetic driver identified, which suggests additional HHMs remain to be discovered.

During my doctoral work, I have addressed these knowledge gaps by completing the first large-scale natural history study investigating pre-leukemic states in individuals with germline mutations in *DDX41*, *GATA2*, or *RUNX1* in collaboration with colleagues from the University of South Australia and the National Institutes of Health. These studies demonstrate that strikingly distinct patterns of leukemogenesis exist among these syndromes. As a component of this work, I drove our group's efforts to generate the largest international genomics database focused on *RUNX1* germline mutation carriers, which allows genomics data from disparate sources to be analyzed via a uniform

bioinformatics pipeline. I hope that this database will facilitate data sharing in the field of hereditary hematopoietic disorders. I extended this work to perform the largest natural history study investigating leukemogenesis in individuals with a hereditary thrombocytopenia/hereditary hematopoietic malignancy (HT/HHM) phenotype driven by germline mutations in *ANKRD26* or *ETV6*. These two syndromes phenocopy *RUNX1* germline mutations, but it was previously unknown how pre-leukemic states differed in these phenocopies as compared to *RUNX1* germline mutation carriers. I have also been involved in a study comparing induced pluripotent stem cells (iPSCs) derived from patients with *ETV6* or *RUNX1* germline mutations, the first of its kind, as well as the first studies characterizing an iPSC line derived from a patient with a germline *ANKRD26* mutation. Finally, I have also identified a novel HT/HHM driven by germline mutations in *TTLL6* and have investigated the potential role of germline *TUBB1* mutations as a novel HT/HHM syndrome. Overall, my results regarding HHMs have generated novel insights into the leukemogenic mechanisms among individuals with HHM-associated germline mutations, have facilitated data sharing and collaboration across international and institutional boundaries, and have generated patient-derived cellular models that may be used for future work investigating mechanisms of disease and novel therapeutic approaches specific to the HHMs.

Chapter I

Introduction

The majority of this chapter is adapted from a series of reviews that I have written in collaboration with Dr. Lucy Godley and our co-authors: Drs. Imo Akpan, Jane Churpek, Simone Feurstein, Matthew Jones, Afaf Osman, Kiran Tawana, Gregory Roloff, and Allison West. These reviews have been published in *Seminars in Oncology*, *Blood*, *Leukemia*, *Current Hematologic Malignancy Reports*, *JCO Precision Oncology*, and *Human Molecular Genetics.*

Feurstein S, **Drazer MW**, Godley LA. Genetic predisposition to leukemia and other hematologic malignancies. *Seminars in Oncology*. 2016; (43): 598-608. PMID 27899193.

Drazer MW, Feurstein S, West AH, Jones MF, Churpek JE, Godley LA. How I diagnose and manage individuals at risk for inherited myeloid malignancies. *Blood*. 2016; 16: 1800-1831. PMID 27471235.

Tawana K, **Drazer MW**, Churpek JE. Universal genetic testing for inherited susceptibility in children and adults with myelodysplastic syndrome and acute leukemia: are we there yet? *Leukemia*. 2018; 32 (7): 1482-1492. PMID: 29483711.

Akpan IJ, Osman AEG, **Drazer MW**, Godley LA. Hereditary myelodysplastic syndrome and acute myeloid leukemia: diagnosis, questions, and controversies. *Current Hematologic Malignancy Reports*. 2018; 13 (6): 426-434. PMID: 30259338.

Roloff GW, **Drazer MW**, Godley LA. Inherited susceptibility to hematopoietic malignancies in the era of precision oncology. *JCO Precision Oncology*. 2021; (5): 107-122.

Feurstein S, **Drazer MW**, Godley LA. Germline predisposition to hematopoietic malignancies. *Human Molecular Genetics*. 2021.

I have adapted figures, tables, and texts from several of these reviews for this chapter. I have also adapted several figures, tables, and texts from published papers for which I served as an author or co-author:

(i) Data presented in the section of this chapter, “Diagnosis of HHMs” and “Prevalence of HHMs” have been published in: **Drazer MW**, Kadri S, Sukhanova M, Patil SA, West AH, Feurstein S, Calderon DA, Jones MF, Weipert CM, Daugherty CK, Ceballos-Lopez AA, Raca G, Lingen MW, Li Z, Segal JP, Churpek JE, Godley LA. (2018). Prognostic tumor sequencing panels frequently identify germline variants associated with hereditary myeloid malignancies. *Blood Advances*. 2(2):146-50. PMID: 29365323.

(ii) Data presented in the sections of this chapter, “Diagnosis of HHMs” and “Prevalence of HHMs” have been published in: Roloff GW, Godley LA, **Drazer MW** (2021). Assessment of technical heterogeneity among diagnostic tests to detect germline risk variants for hematopoietic malignancies. *Genetics in Medicine*, 23(1):211-14. PMID: 32807974.

Hereditary cancer and hereditary hematopoietic malignancy syndromes: a historical perspective

The first recognized hereditary hematopoietic malignancy (HHM) was described in an 1861 case report from Biermer (**Figure 1.1**)(Biermer, 1861). In this publication, an unaffected woman had three children, each of whom developed splenomegaly and died before three years of age (Biermer, 1861). The woman had four children with her second husband. Again, three of these children developed splenomegaly, and one was diagnosed with acute leukemia at autopsy (Biermer, 1861). Videbaek performed the first large study of familial leukemia in 1947, showing that 8.1% of 209 probands with hematopoietic malignancies (HMs) had a family history of blood cancers (Videbaek, 1947). This publication, however, did not include pedigrees or data consistent with a clear Mendelian inheritance pattern. Intriguingly, Videbaek also demonstrated that a large number of the families also experienced an elevated risk of solid tumors, the first evidence that hereditary blood cancer syndromes did not have purely hematopoietic phenotypes (Videbaek, 1947). In 1975, Gunz *et al.* studied 909 individuals with leukemia and demonstrated their first-degree relatives experienced a 2.8-3.0 fold increased risk of leukemia (Gunz et al., 1975). This risk was largely driven by chronic lymphocytic leukemia (CLL) and acute leukemia (AL) diagnoses. No excess cases of chronic granulocytic leukemia (now referred to as chronic myeloid leukemia [CML]) were diagnosed (Gunz et al., 1975). In the Gunz study, a similar proportion of probands with a family history of leukemia (7.9%) was observed when compared to the original Videbaek study (8.1%). This publication was notable for the first published pedigree of “pure” HMs with a Mendelian inheritance pattern, and the group also noted that multiple families

experienced excess risk for both lymphomas and leukemias, but not an excess risk for solid tumors. This pedigree was notable for a male proband with CML and a maternal lineage that was notable for a striking prevalence of AL, including a maternal aunt and eight maternal relatives (Gunz et al., 1975).

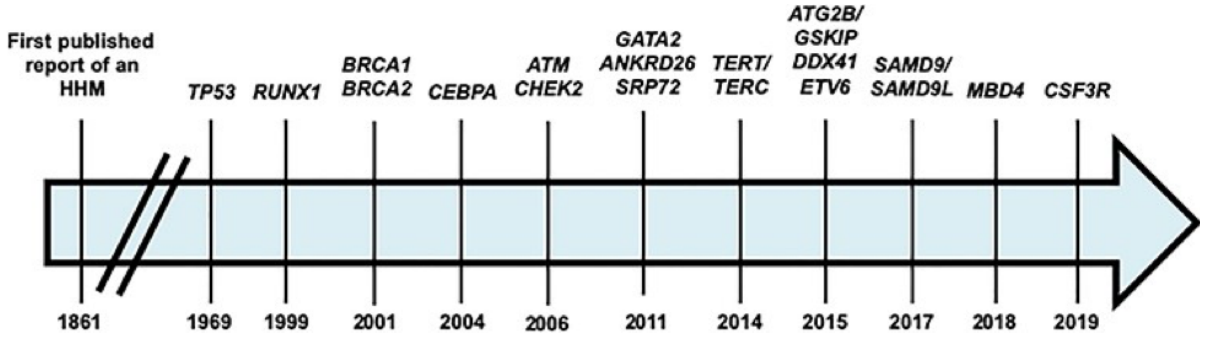
Although these early studies demonstrated the existence of HHMs, the molecular basis of HHMs and other hereditary cancer syndromes remained unknown until the seminal work of Drs. Lynch, Li, Fraumeni, King, and their colleagues during the 20th century. Dr. Henry Lynch and his colleagues established the hereditary basis for hereditary non-polyposis colorectal cancer, now termed Lynch Syndrome (LS), in the early 1960s (Cantor, 2006). Intriguingly, although HMs are not traditionally considered to be a LS-associated tumor, multiple case series have linked HMs to LS-associated germline mutations, with lymphoma in particular being observed frequently in these pedigrees (Bandipalliam, 2005; Bougeard et al., 2003; Scott et al., 2007).

In 1969, Li and Fraumeni laid the groundwork for Li-Fraumeni Syndrome (LFS) (Li & Fraumeni, 1969). Although HMs are not considered to be a malignancy associated with LFS, there are some data suggesting that LFS may include a predisposition to blood cancers. For example, Li and Fraumeni's initial report included two pedigrees that were notable for cases of acute lymphocytic leukemia (ALL) and acute myeloid leukemia (AML) (Li & Fraumeni, 1969). The association between LFS and HHM risk has been further strengthened since the original LFS publication (Bougeard et al., 2015).

Dr. King, her colleagues, and other groups subsequently demonstrated that germline mutations in *BRCA1* or *BRCA2* were responsible for Mendelian breast and ovarian cancer in the late 20th century (Friedman et al., 1994; Hall et al., 1992; Hall et al.,

1990; Wooster et al., 1994). Germline *BRCA1* and *BRCA2* mutations have since been shown to drive HHMs in both humans and murine models (Pouliot et al., 2019; Risch et al., 2001; Vasanthakumar et al., 2016)

Figure 1.1. Milestones in the discovery of hereditary cancer syndromes and hereditary hematopoietic malignancy (HHM) syndromes. The advent of relatively low-cost next generation sequencing has accelerated the discovery of HHMs since 1999.



Hereditary thrombocytopenia/hereditary hematopoietic malignancy syndromes

Germline RUNX1 mutations

Ultimately, the first "pure" HHMs were discovered in the late 1990s, beginning with the autosomal dominant familial platelet disorder with associated myeloid malignancy (FPD/MM, OMIM 601399) caused by germline *RUNX1* mutations. *RUNX1* encodes for a master hematopoietic transcription factor (TF), and one of its targets is the 5' untranslated region (UTR) of *ANKRD26* (see next section discussing germline *ANKRD26* mutations). FPD/MM is notable for qualitative platelet defects, mild to moderate thrombocytopenia, and an increased risk (up to 44%) of HMs. The median age of malignancy diagnosis is young (33 years) as compared to population controls (Godley, 2014; West et al., 2014). The phenotypic spectrum of HMs stemming from germline *RUNX1* mutations includes

MDS, AML, T-cell acute lymphoblastic leukemia (T-ALL), and hairy cell leukemia. Many individuals with germline *RUNX1* mutations are initially misdiagnosed with immune thrombocytopenia (ITP) and are treated with ITP-directed therapies such as systemic steroids, splenectomy, or thrombopoietin agonists. None of these therapeutic responses will produce a durable treatment response (Melazzini et al., 2016; Noris et al., 2011; Tsang et al., 2017).

Of particular interest to my doctoral work is a prior report from the groups of Drs. Godley and Graubert that investigated the presence of clonal hematopoiesis (CH) in a relatively small cohort (n = 9) of unaffected carriers of germline *RUNX1* mutations. CH occurs when hematopoietic stem and progenitor cells (HSPCs) acquire one or more leukemogenic somatic mutation(s), but the bone marrow tissue itself does not show gross morphologic evidence of an HM (Steensma, 2018). This work demonstrated that the prevalence of CH was dramatically increased in *RUNX1* mutation carriers at early ages, with 66% of unaffected carriers under the age of 50 developing CH, as compared to 1% of the general population that develops CH prior to the age of 50 (Churpek et al., 2015; Genovese et al., 2014; Jaiswal et al., 2014). Prior to my doctoral work, similar studies had not been performed in other HHMs.

Germline ANKRD26 mutations

Following the original publication from Song *et al.* demonstrating that germline *RUNX1* mutations predispose to HM with lifelong thrombocytopenia (FPD/MM), the first FPD/MM phenocopy driven by autosomal dominant germline mutations in the 5' UTR of *AKNRD26* was described in 2011 (Thrombocytopenia 2, OMIM 6210855) (Bluteau et al.,

2014; Marquez et al., 2014; Noris et al., 2013; Noris et al., 2011; Pippucci et al., 2011). As with *RUNX1* germline mutations, individuals with germline *ANKRD26* mutations experience decreased platelet counts and carry an elevated lifetime risk (8%) for HMs, including MDS, AML, CLL, and CML (Feurstein et al., 2016; Noris et al., 2013; Noris et al., 2011). The age of onset of HMs is variable, ranging between 30 and 70 years of age (Noris et al., 2013; Noris et al., 2011). Intriguingly, a clear molecular link exists between *ANKRD26* and *RUNX1*, as the 5' UTR of *ANKRD26* is a target of the RUNX1 TF (Bluteau et al., 2014). RUNX1 negatively regulates *ANKRD26* expression, and mutations in the 5' UTR of *ANKRD26* result in *ANKRD26* overexpression. This overexpression persists through the late stages of megakaryopoiesis, ultimately leading to abnormal MAPK signaling during early platelet biogenesis (Bluteau et al., 2014).

A series of publications have demonstrated that germline *ANKRD26* mutations are more common than previously recognized and are the most common of the FPD/MM phenocopies (*ANKRD26*, *ETV6*, and *RUNX1*-driven HHMs). For example, Noris *et al.* demonstrated that 10% of families with unsolved hereditary thrombocytopenia carried germline *ANKRD26* mutations (Noris et al., 2011). Similarly, Downes *et al.* sequenced patients of predominantly European ancestry who presented with various hematopoietic phenotypes and discovered that germline *ANKRD26* mutations were the most common cause of abnormal platelet counts among HHMs driven by *ANKRD26*, *ETV6*, or *RUNX1* germline mutations (Downes et al., 2019).

Germline ETV6 mutations

A third hereditary thrombocytopenia/HM phenocopy driven by autosomal dominant germline *ETV6* mutations has also been described (Noetzli et al., 2015; Noris et al., 2011; Topka et al., 2015; Zhang et al., 2015). This FPD/MM phenocopy, termed thrombocytopenia 5 (OMIM 616216), is notable for a decreased platelet count as well as a more diverse HM phenotype, with affected carriers developing a spectrum of both myeloid and lymphoid malignancies (pre-B-ALL > MDS/AML > multiple myeloma). Intriguingly, pedigrees with *ETV6* mutations have also included affected carriers with solid tumors, including melanoma and colorectal cancer (Noetzli et al., 2015; Topka et al., 2015; Zhang et al., 2015). The lifetime risk for HMs in germline *ETV6* mutation carriers is approximately 33%, with a relatively early median age of onset (33 years) (Drazer et al., 2016; Noetzli et al., 2015; Topka et al., 2015; Zhang et al., 2015).

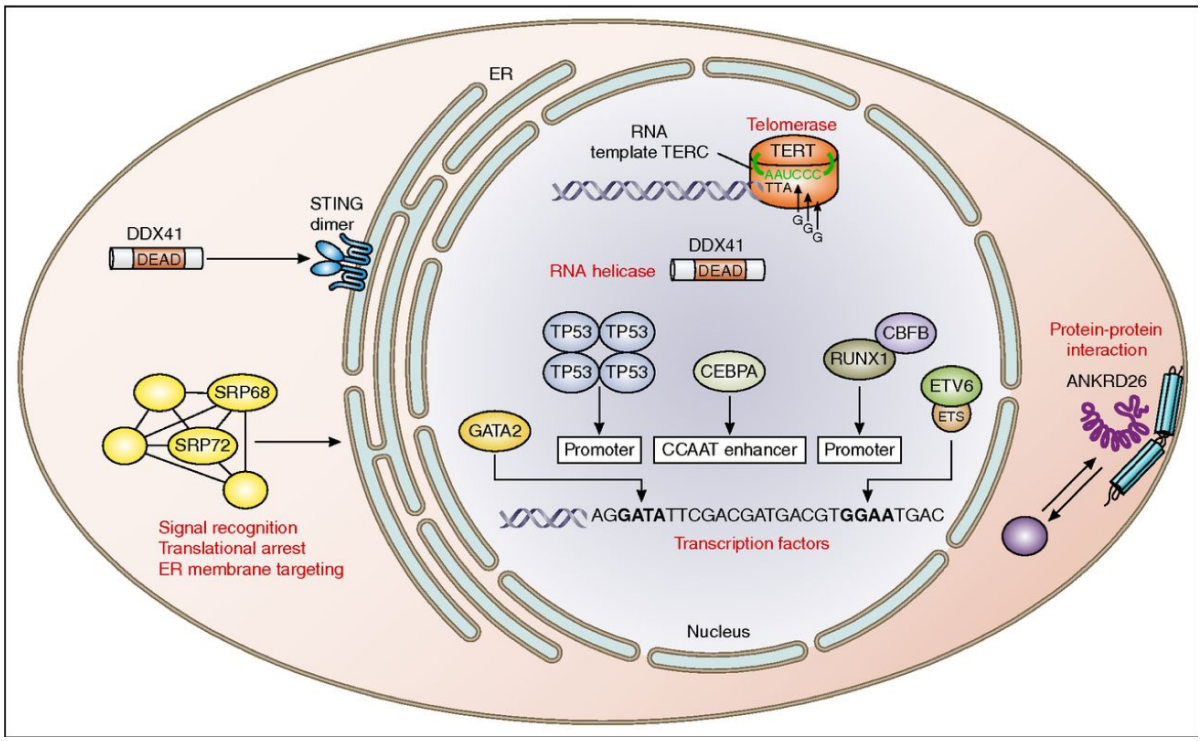
ETV6 is a transcriptional repressor that is involved in myriad translocations in a variety of tumors, including *ETV6-RUNX1*, generated via t(12;21)(p13;q22). This is the most common translocation in pediatric ALL (up to 30% of cases) but is relatively infrequent in AML (Hernández et al., 2000; Odero et al., 2001; Romana et al., 1995; Wlodarska et al., 1996). Although somatic point mutations are relatively infrequent in *ETV6* in HMs (approximately 2%), ETV6 protein expression is diminished in 33% of AML cases (Barjesteh van Waalwijk van Doorn-Khosrovani et al., 2005; Bejar et al., 2011; Papaemmanuil et al., 2009).

HHMs without lifelong thrombocytopenia

The majority of HHMs are not associated with lifelong thrombocytopenia. Non-thrombocytopenic HHMs are driven by germline mutations in *ATM*, *CEBPA*, *CHEK2*, *CSF3R*, *DDX41*, *GATA2*, *SAMD9*, *SAMD9L*, *TERC*, *TERT*, *TP53*, and other syndromes. Instead, the majority of these syndromes are notable for normal peripheral blood counts until the development of an overt hematopoietic process.

The genes involved with these HHMs are involved in myriad cellular processes (**Figure 1.2**) (Drazer et al., 2016). For example, in addition to the previously described TFs *ETV6* and *RUNX1*, germline mutations in *CEBPA*, *GATA2*, *IKZF1*, *PAX5*, and *TP53*, which encode TFs, are also associated with HHMs. Additional HHMs are driven by germline mutations in genes associated with DNA repair, such as *ATM*, *BLM*, *CHEK2*, the Fanconi genes, *MBD4*, *NBN*, *RECQL4*, and *USP45*. Numerous HHMs driven by germline mutations in genes involved in cellular proliferation have also been described: *CBL*, *CSF3R*, *JAK2*, *MPL*, *PTPN11*, *RBBP6*, *SH2B3*, *SAMD9*, and *SAMD9L*. Germline *TET2* mutations, which are involved in epigenetic processes, have also been implicated in HHMs (Drazer et al., 2016).

Figure 1.2. Myriad molecular roles of proteins encoded by HHM-associated genes.



Germline DDX41 mutations

DDX41 encodes for a DEAD-box RNA helicase that is involved in RNA processing and STING signaling, but the function of the protein is otherwise relatively not well understood. The most prevalent HHM stems from germline mutations in *DDX41*, which account for 1%-2.4% of all seemingly sporadic HMs (Polprasert et al., 2015). Germline *DDX41* mutations were initially described in myeloid malignancies (MDS/AML), but the phenotype quickly expanded to also involve lymphoid neoplasms (Hodgkin lymphomas, non-Hodgkin lymphomas, and multiple myeloma)(Lewinsohn et al., 2016). Perhaps the most striking aspects of *DDX41*-driven HHMs are their relatively long latency prior to tumorigenesis (62 years), which is similar to the median age of onset of *de novo* myeloid

malignancies in the general population, their exceptionally high penetrance of malignancies (80%), and the common occurrence of somatic second hits in *DDX41* during tumorigenesis. Despite their high penetrance, *DDX41*-driven HHMs are frequently underdiagnosed secondary to their relatively late age of onset, which reduces the clinical suspicion for this syndrome in both the proband and affected family members (Drazer et al., 2018; Feurstein et al., 2016; Lewinsohn et al., 2015; Polprasert et al., 2015). Intriguingly, *DDX41*-driven HHMs show a sex-based skewing, with male patients possessing a higher risk for leukemogenesis than women (Quesada et al., 2019). Ancestry-informative *DDX41* germline variants have also been described, with the p.D140Gfs and p.M1? mutations observed more frequently in individuals with European ancestry as compared to the p.A500Cfs mutation that is more common in individuals with Asian ancestry (Sanders, 2019).

DDX41-driven HHMs may also represent one of the few HHMs for which a targeted therapy exists, as exceptional responses to lenalidomide-based therapy have been described by multiple groups. *DDX41* is located on chromosome 5q35, an area that is frequently lost in the del(5q) syndrome, which has also been shown to be responsive to lenalidomide monotherapy (Abou Dalle et al., 2020; Polprasert et al., 2015). However, these studies have been small and uncontrolled, and no prospective randomized trials investigating this important clinical question have been performed.

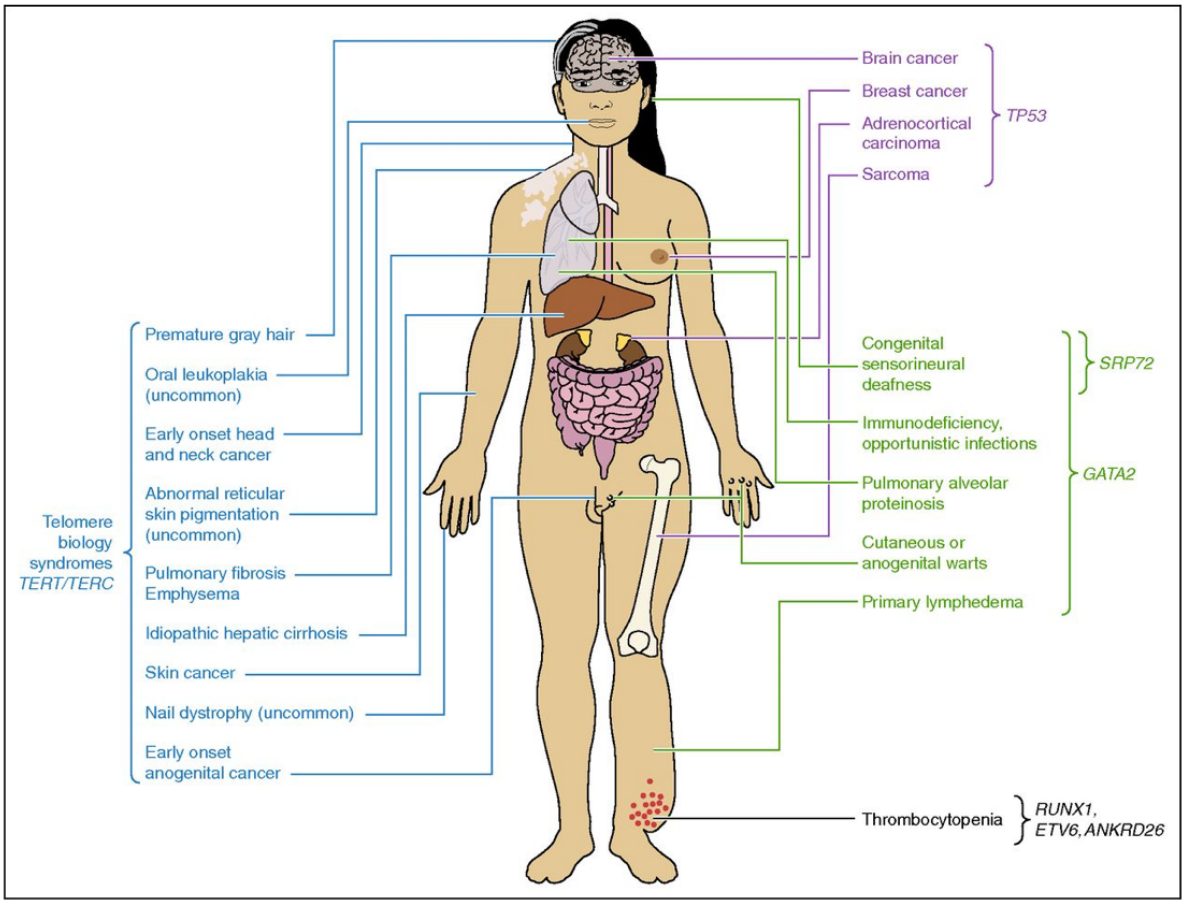
Germline GATA2 mutations

GATA2 encodes for a TF with roles in regulating hematopoiesis and vascular development (Hahn et al., 2011). HHMs driven by *GATA2* are a distinctive set of

syndromes that are remarkable for their profound immunodeficiency in the form of monocytopenia, B-cell lymphocytopenia, natural killer cell deficiency, and neutropenia. These immunodeficiencies generally precede leukemogenesis, and *GATA2*-mutated HHMs represent the rare HHM syndrome for which prophylactic HSPC transplant has been studied, primarily with the goal of reducing the life-threatening infections associated with this syndrome (Burroughs et al., 2017; Cuellar-Rodriguez et al., 2011; Grossman et al., 2014; Parta et al., 2018). *GATA2*-driven HHMs are also notable for their exceptionally high penetrance (90%), which appears to be one of the highest rates of disease penetrance among the known HHMs. In addition to immunodeficiency, the *GATA2*-deficient phenotype is notable for mucocutaneous warts, deafness, lymphedema, and pulmonary proteinosis (Spinner et al., 2014).

Leukemogenesis in *GATA2*-driven HHMs has been shown to occur via distinct patterns, such as HMs driven by somatic monosomy 7 events and/or trisomy 8 (Dickinson et al., 2011; Donadieu et al., 2018).

Figure 1.3. The variable phenotypic manifestations of the HHMs.



Diagnosis of HHMs

Astute physicians have noted the presence of HHMs for more than a century, but the clinical diagnosis of HHMs remains challenging, particularly because the HHMs frequently present with myriad physical abnormalities (**Figure 1.3**), variable expressivity, and variable penetrance. Additionally, many physicians are trained to assume hereditary cancer syndromes are pediatric disorders, but many HHMs are notable for blood cancers with long latencies that arise during adulthood. For example, the median age of HM diagnosis in individuals with germline *DDX41* mutations is 62 years, an age that closely

mirrors the median age at which sporadic blood cancers are diagnosed (Lewinsohn et al., 2015; Polprasert et al., 2015; Tawana & Fitzgibbon, 2016).

An additional difficulty for many patients at risk for an HHM is that tissues that are commonly utilized for genetic testing (e.g., saliva or peripheral blood) are generally contaminated by HM tumor tissue. As such, the most stringent method for germline genetic testing in patients with HMs involves the culture of skin fibroblasts that are obtained via punch biopsy at the time of a bone marrow biopsy and/or during an in-office procedure. Anecdotally, many physicians are uncomfortable performing this procedure and instead will attempt to collect samples during periods of clinical remission following HM-directed therapy. However, this approach is inherently flawed as many patients may still be affected by the presence of variants observed during CH. These variants may be present at relatively high variant allele frequencies (VAFs) in HHM-related genes, a phenomenon which complicates the accurate diagnosis of an HHM.

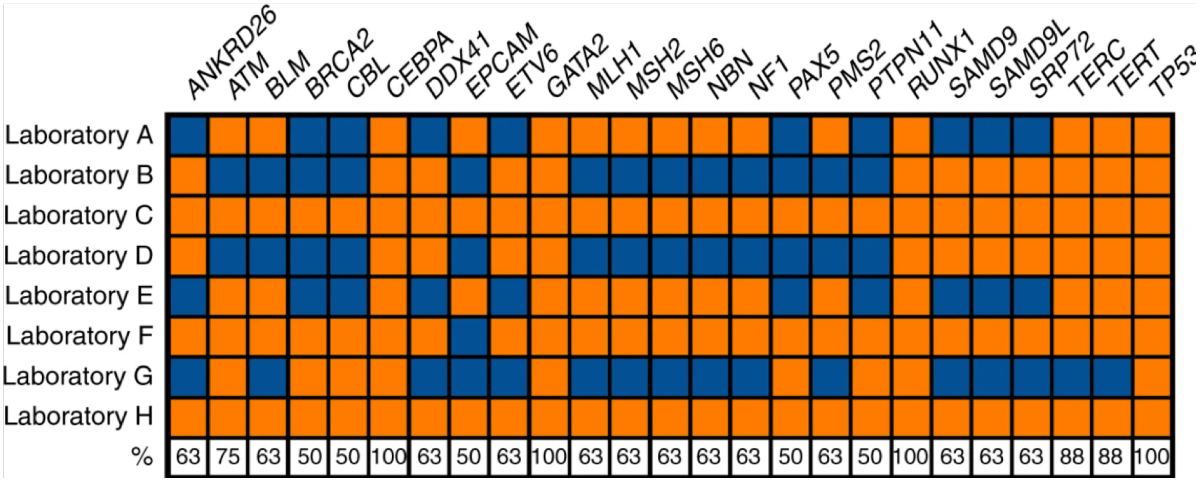
Furthermore, some HHMs are notable for somatic reversion in which the hematopoietic tissue does not carry the variant of interest, for which an accurate diagnosis requires the collection of true germline tissue. For example, the normal function of the SAMD9 or SAMD9L protein product is to suppress cellular proliferation. HHM-associated mutations in *SAMD9* or *SAMD9L* are gain-of-function variants, which are negated via somatic reversion events such as uniparental disomy, loss-of-function mutations, or monosomy in the affected tissue (Chen et al., 2016; Collin, 2017; Davidsson et al., 2018; Nagamachi et al., 2013; Narumi et al., 2016; Tesi et al., 2017). As such, HHMs driven by germline mutations in *SAMD9* or *SAMD9L* are frequently missed unless diagnostic procedures are performed using cultured skin fibroblasts.

An additional pitfall that must be avoided during the diagnosis of HHMs is the use of suboptimal diagnostic assays. Clinically, I observed multiple cases of individuals who were referred to our program for HHM evaluation following a “negative” evaluation at other institutions or practices. On multiple occasions, the diagnostic assays that were sent prior to referral were sent from hematopoietic tissue (peripheral blood or tumor). This placed the patient at risk for misdiagnosis secondary to somatic reversion events. Other individuals had HHM testing performed with panels that did not sequence the full spectrum of genes involved in HHMs – for example, multiple patients were tested with panels that did not sequence *DDX41*. Other patients had panels sent that sequenced HHM-related genes of interest, but not the correct genomic loci that contained the notable mutational “hotspots” of interest (i.e., the 5’ UTR of *ANKRD26*). Similarly, germline CNVs in *RUNX1* and other genes are a known driver of HHMs, but many commercial laboratories do not utilize sequencing techniques that detect CNVs.

However, a systematic analysis of HHM diagnostic assays had never been performed. To address this gap in the field, I collected data from all laboratories that currently offer a commercial diagnostic test for HHMs. I analyzed these assays to determine if they were able to detect the full spectrum of HHM-related genes and HHM-related variants. Shockingly, I demonstrated that the majority of these assays are insufficient for the accurate diagnosis of HHMs (**Figure 1.4**), a shortcoming that leaves patients at risk for misdiagnosis and its subsequent sequelae, including donor-derived leukemias (Roloff et al., 2020). This work, for which I served as senior author, was published in *Genetics in Medicine*. This study ultimately demonstrated that many diagnostic assays used for HHM evaluation are insufficient, and that physicians and

genetic counselors involved in the care of these patients must carefully select a diagnostic assay for evaluation of these patients (Roloff et al., 2020).

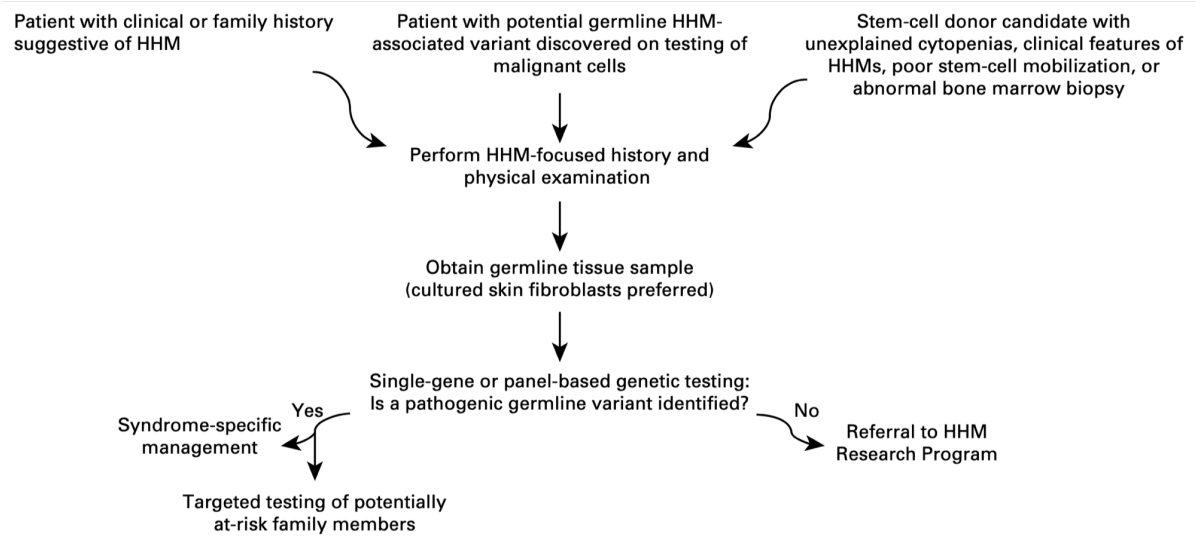
Figure 1.4. Heterogeneity in the genes sequenced by clinical-grade diagnostic assays designed for the diagnosis of HHMs. Orange: gene sequenced; Blue: gene not sequenced.



Despite these challenges, the most common methods for identifying an individual who is at risk for an HHM are threefold (**Figure 1.5**)(Roloff GW, 2021.). First, a patient may present to the clinic with a personal or family history that is highly suggestive of an HHM. Classic examples of this presentation include a personal history of lifelong thrombocytopenia, multiple first-degree relatives with thrombocytopenia, and multiple first- and second-degree relatives with HMs diagnosed at relatively early ages of onset. These histories would be highly suggestive of an HHM driven by germline mutations in *ANKRD26*, *ETV6*, or *RUNX1*, the FPD/MM phenocopies. Second, a patient with an HM has molecular and genetic profiling of their hematopoietic tissue (peripheral blood or bone marrow) performed, and the results from this assay demonstrate the presence of a variant that would be considered likely pathogenic or pathogenic (LP/P) if the variant is present in the germline tissue. Third, a patient with an HM is underdoing hematopoietic stem and

progenitor cell (HSPC) transplant from a matched related donor (MRD), but the MRD is found to be a “poor mobilizer” in whom growth factors do not produce a sufficiently large yield of HSPCs (Rojek et al., 2016).

Figure 1.5. The diagnostic workflow associated with HHMs.



Any of these findings should prompt an HHM-focused history and physical examination, the collection of cultured skin fibroblasts, and HHM-focused genetic testing. At the University of Chicago, any patients who are “negative” for likely LP/P variants in known HHM-related genes, but who have a striking phenotype that is suggestive of an undiagnosed HHM, are offered enrollment on our HHM-specific research program, for which Dr. Lucy Godley serves as the Principal Investigator. I am a co-investigator on this protocol, which has been amended multiple times over the course of my doctoral training to allow for the generation of permanent cell lines, such as Epstein-Barr Virus (EBV)-transformed lymphoblastoid cell lines (LCLs). Research performed using these cell lines is included in the ensuing chapters.

One common clinical scenario occurs when an individual with a seemingly sporadic HM has tumor-directed sequencing performed, but this sequencing demonstrates that the tumor clone carries the presence of a variant in an HHM-related gene and as such would represent an HHM if present in the germline. Previously, it was unknown how frequently this scenario occurred, how often these potential HHM-associated variants were of germline origin, and whether or not the presence of a germline variant could be predicted via sequencing-specific factors such as the VAF of the variant in question.

During the initial portion of my doctoral training, I addressed these questions by identifying 360 individuals with HMs who had tumor-only sequencing performed at the University of Chicago Medical Center. I then analyzed the sequencing data from these individuals in order to identify individuals whose tumor sequencing identified LP/P variants in HHM-related genes. In total, 74 (21%) of these individuals carried 88 LP/P variants in HHM-associated genes. Of these patients, 44 had germline tissue available for sequencing. I then analyzed the data based on VAF, working under the assumption that any variants with a VAF less than 0.4 were unlikely to be of germline origin (which was ultimately confirmed). Using this bioinformatic approach, I was able to reduce the number of patients under investigation from 44 to 25, and 6 of these 25 patients (24%) were ultimately found to carry the variant of interest in their germline tissue (**Figure 1.6**) (Drazer et al., 2018).

Figure 1.6. Schematic summarizing the patient population, tissue availability, and general sequencing results for likely pathogenic/pathogenic variants. VAF: variant allele frequency.

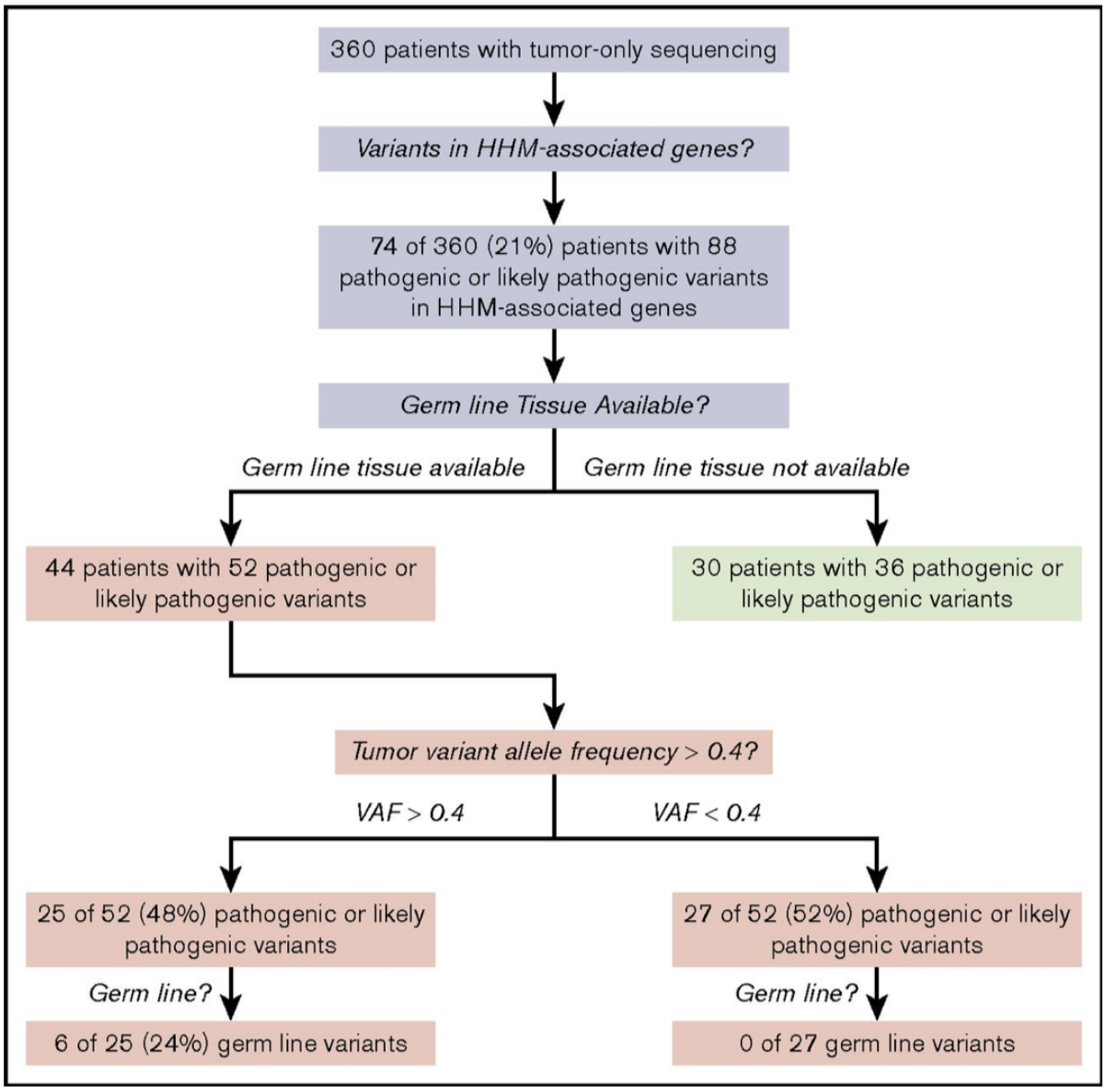
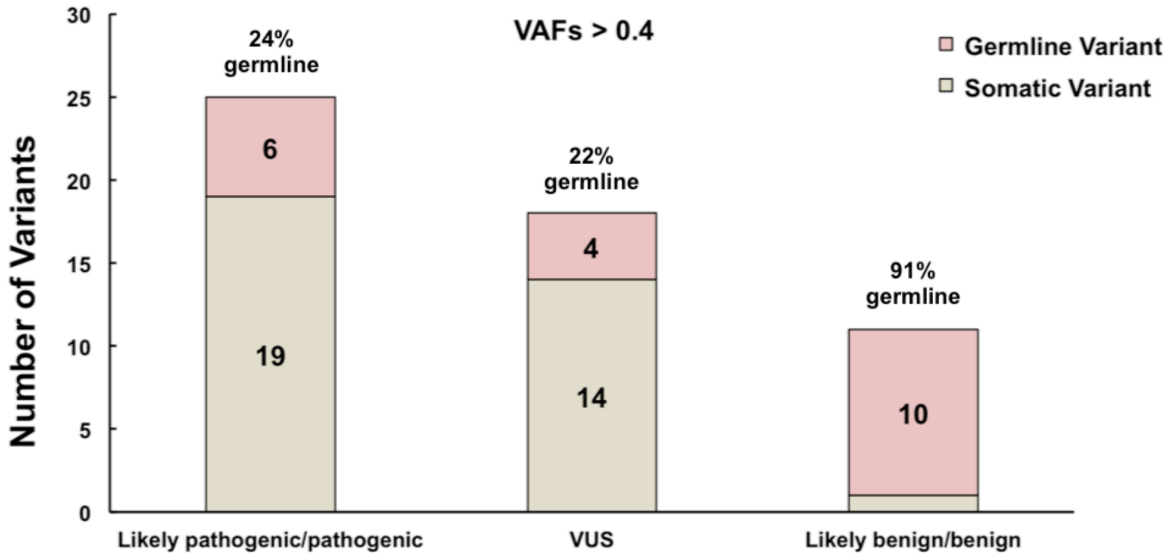


Figure 1.7. Diagnostic yield for germline variants, stratified by pathogenicity (Likely pathogenic/pathogenic; variant of undetermined significance; likely benign/benign). Of note, no germline LP/P variants were identified with a VAF < 0.4.

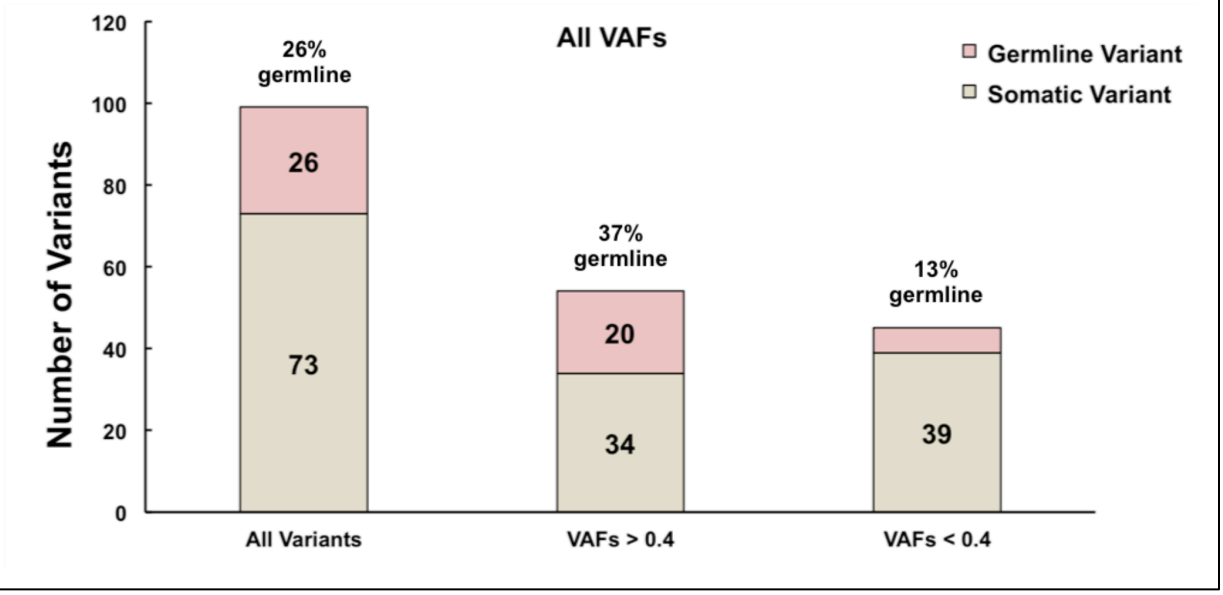


Ultimately, this approach demonstrated that approximately one in five patients will have LP/P variants with a VAF >0.4 identified in HHM-associated genes via tumor-only sequencing that would be causative for HHMs if present in the germline. Among these patients, approximately 1 in 4 were ultimately found to carry the variant in their germline tissue, for an overall diagnostic yield of 5.0% using this approach **(Figure 1.7**).

Of note, some germline variants had a VAF less than 0.4 in the hematopoietic tissue, however none of these variants were LP/P. This finding demonstrated that VAF should not be used as a stringent cutoff for the determination of germline status, a point that I emphasized in the discussion in the final publication in *Blood Advances* (**Figure 1.8**). I have subsequently identified germline LP/P variants in additional patients in which the initial VAF in hematopoietic tissue was less than 0.4, with insertion/deletion (indel) variants being especially prone to VAFs that would be lower than expected for a germline variant. Of note, copy number variants (CNVs) would also potentially skew the VAF of the

non-affected WT allele to VAFs well above what would be expected for a heterozygous variant.

Figure 1.8. Diagnostic yield for variants, stratified by VAF. Of note, no germline LP/P variants were identified with a VAF < 0.4.



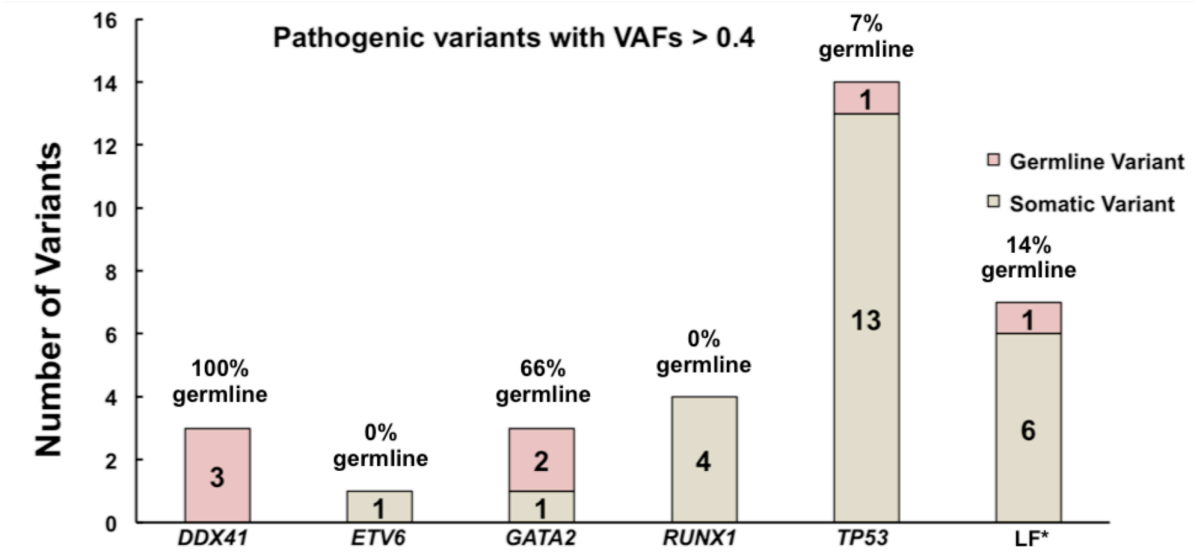
I also investigated the diagnostic yield of this approach using an HHM gene-based approach, with the hypothesis that genes that are frequently mutated during leukemogenesis, such as *TP53*, would have a lower germline diagnostic yield as compared to genes that are less frequently mutated during leukemogenesis. My hypothesis was ultimately proven to be correct, as the majority (93%) of *TP53* variants with a VAF greater than 0.4 were ultimately found to be of somatic origin, although this germline diagnostic yield could be increased to 14% by limiting the analysis only to *TP53* variants that were known to be associated with Li-Fraumeni syndrome (**Figure 1.9**).

Prevalence of HHMs in an unselected population of patients with blood cancers

To date, the prevalence of HHMs in an unselected, ancestrally diverse population

of individuals with blood cancers is unknown, and no prospective study has addressed this question using a sequencing approach that utilizes true paired tumor/germline tissue, sequences all of the HHM-associated genes, and that employs a sequencing platform that is sensitive for the full variant spectrum involved in HHMs, most notably CNVs, which are notoriously difficult to call using panel-based sequencing approaches similar to that utilized on a clinical and research basis at the University of Chicago Medical Center (Kadri et al., 2017).

Figure 1.9. Germline diagnostic yield when stratified by HHM gene of interest.



Although my work above sets a lower bound of approximately 5.0% on the estimated prevalence of HHMs in a relatively unselected patient population, this is almost certainly an underestimate, primarily because the tumor sequencing panels that were used in this study were relatively limited in the number of HHM-related genes they sequenced. For example, the tumor sequencing panels that I utilized in the two earliest phases of the study only sequenced *TP53* from April to October 2014 (n = 29 patients) and sequenced *ETV6*, *RUNX1*, and *TP53* from October 2014 to February 2016 (n = 133

patients). The panel sequenced *ANKRD26*, *CEBPA*, *DDX41*, *ETV6*, *GATA2*, *RUNX1*, *SRP72*, *TERT*, and *TP53* from February 2016 to July 2017 (n = 198 patients). However, even the final iteration of the panel was not comprehensive for all HHM-associated genes, was not specifically designed to identify CNVs, and did not include paired tumor/germline samples so as to detect HHMs for which somatic reversion events are common. Despite these limitations, the most comprehensive panel in my study detected 21 germline variants among 198 patients (germline yield of 11.0%), as compared to 5 germline variants among 162 patients (germline yield of 3.1%) using the two earlier iterations of the tumor sequencing panel that were more limited in the number of HHM-related genes they sequenced **(Table 1.1**). This suggests the lower bound on the prevalence of HHMs in a relatively unselected cohort of patients with blood cancers is closer to 11.0% (**Figure 1.10)** (Drazer et al., 2018).

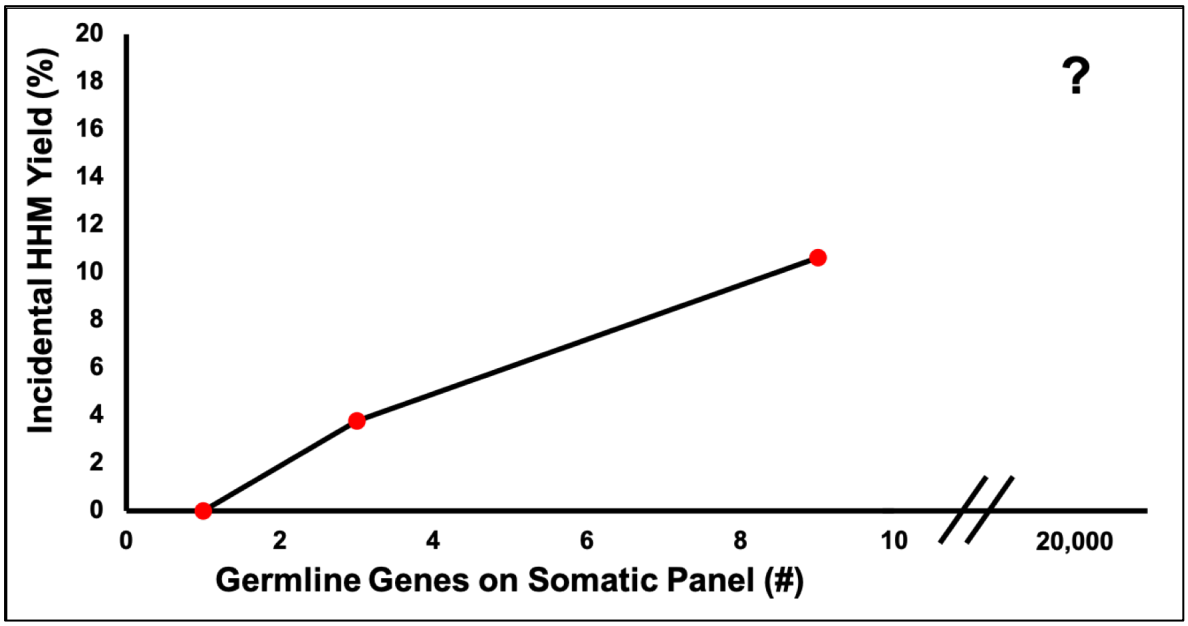
Table 1.1. Germline diagnostic yield when stratified by tumor sequencing panel and number of HHM-related genes sequenced. NGS: next generation sequencing.

NGS Assay	Oncoscreen (n=29)	Oncoheme (n=133)	Oncoplus (n=198)
Time Period	April 2014 - October 2014	October 2014 - February 2016	February 2016 - July 2017
*ANKRD26**	Not assessed	Not assessed	Assessed
CEBPA	Not assessed	Not assessed	Assessed
DDX41	Not assessed	Not assessed	Assessed
ETV6	Not assessed	Assessed	Assessed
GATA2[#]	Not assessed	Not assessed	Assessed
RUNX1	Not assessed	Assessed	Assessed
SRP72	Not assessed	Not assessed	Assessed
TERT	Not assessed	Not assessed	Assessed
TP53	Assessed	Assessed	Assessed
Germline variants Detected (n=26)	**0**	**5**	**21**

The germline yield for patients with blood cancers will also vary based on the phenotype of the patients in question. For example, MDS is typically a later onset disease with a median age of diagnosis of approximately 67 years of age. However, multiple groups have now demonstrated that patients with “early-onset” MDS, variably defined as less than 40 or 45 years of age, are enriched for HHMs, ranging from 13-19% germline

yield (Feurstein et al., 2021; Keel et al., 2016). Similarly, germline *GATA2* mutations alone account for 7% of all MDS cases in patients under the age of 18 (Wlodarski et al., 2016).

Figure 1.10. Germline diagnostic yield is correlated with the number of genes sequenced on the somatic sequencing panel of interest. Germline yield will likely continue to increase as unbiased sequencing techniques are increasingly employed in the clinical setting.



Over time, there has been a shift toward unbiased germline sequencing (whole exome sequencing [WES] and whole genome sequencing [WGS]) among clinical genetics labs; given our local experience of increased germline diagnostic yield using increasingly large sequencing panels (**Table 1.1**, **Figure 1.10**), I anticipate that the germline diagnostic yield will increase as the raw number of genes sequenced increases and the ability to detect CNVs improves utilizing these unbiased methods.

Hypothesis and specific aims

The number of known HHMs has increased significantly since the advent of relatively inexpensive NGS, but the mechanisms driving leukemogenesis in the HHMs

remain relatively unknown. This knowledge gap has persisted because of four major problems: 1) no natural history study has been performed in which individuals with HHM-associated germline mutations were profiled genomically before or after the development of HMs; 2) the number of HHM pedigrees at a single institution is often relatively limited, and high-quality cross sectional “deep phenotyping” studies require collaboration across many research groups; 3) the majority of individuals with HHM phenotypes do not have a germline driver identified; 4) very few high-fidelity models of HHMs have been developed. Additionally, despite a need for collaboration in the field, no common platform for sharing data across research groups has been developed, and no uniform bioinformatic pipeline has been utilized so as to minimize computational differences across research groups.

In order to address these knowledge gaps, the first Specific Aim of my thesis was to perform the first natural history study of individuals with germline mutations in *DDX41*, *GATA2*, or *RUNX1*. In this Aim, I hypothesized that each of these HHMs, which are not phenocopies, would follow unique leukemogenic “steps” in mutation carriers. These HHMs represent the most prevalent HHM (*DDX41*), one of the most highly penetrant HHMs (*GATA2*), and the first described HHM that has the largest described number of kindreds worldwide (*RUNX1*). In order to accomplish this Aim, I was also involved in the development of the RUNX1 database (RUNX1db, https://runx1db.runx1-fpd.org/), the first international database dedicated to the deposition of HHM-focused genomics data (Homan et al., 2021). I utilized the RUNX1db in collaboration with groups from the University of South Australia and the NIH to accomplish Aim 1.

A second knowledge gap in the HHM field is that the HHMs generally can be grouped together by clinical phenotype, as many HHMs phenocopy each other. For example, *ANKRD26*, *ETV6*, and *RUNX1*-driven HHMs share a phenotype of HT/HHMs. However, it is unknown if these HT/HHMs are merely clinical phenocopies or if they share additional genetic properties such as somatic leukemogenic drivers. In theory, the identification of a shared leukemogenic driver among these phenocopies could streamline the development of therapies specific to these HHMs. In order to address this gap, I performed the first systematic study of the HT/HHM phenocopies in order to determine if they shared somatic driver mutations in shared molecular pathways. My hypothesis for this Aim was that unaffected individuals with HT/HHM-associated germline mutations would show similar patterns of CH.

Because no natural history studies have been completed in the field, investigators' ability to generate high-fidelity cellular and murine models of HHMs has also been limited. This lack of models has slowed the development of HHM syndrome-specific therapies, and no treatments are currently available that improve bone marrow function in the HHMs with lifelong cytopenias (i.e., *ANKRD26*, *ETV6*, or *RUNX1*, among others), that abrogate leukemogenesis in unaffected carriers, and that treat malignancies in affected HHM mutation carriers. In order to address this gap, I developed a patient-derived induced pluripotent stem cell (iPSC) line, and then characterized this line during differentiation into iPSC-derived hematopoietic stem and progenitor cells (HSPCs). My hypothesis for this work was that this line would recapitulate the transcriptomic signature of MAPK hyperactivation which was previously seen in megakaryocytes collected from individuals

with *ANKRD26* germline mutations, and that this MAPK hyperactivation was a constitutive effect at all stages of hematopoiesis.

Finally, the majority of individuals with an HHM-like phenotype who undergo comprehensive germline sequencing and clinical evaluation do not have a clear genetic driver identified. In particular, individuals with an HT/HHM phenotype only have a 21% germline diagnostic yield (Guidugli et al., 2017). This finding suggests that myriad genetic drivers of disease remain to be discovered in the HHMs. These drivers likely have not been identified for any combination of the following reasons: 1) novel variants, structural variants in particular, in known HHM-related genes may not be readily detected with the short-read sequencing NGS technology that is most commonly utilized in current research-based efforts, with the first long-read sequencing approach being first used in 2021 for *ANKRD26* variant identification (Wahlster et al., 2021); 2) variants in novel HHM-related genes may not be discovered without unbiased WES or WGS in combination with NGS technology that facilitates the discovery of structural variants; 3) many HHM genetic drivers are ultra-rare in the general population, with population minor allele frequencies on the order of 10^{-4} for even the most "common" pathogenic variant in *DDX41*, which itself is the most common HHM-related driver (**Table 1.2**).

Consequently, identifying multiple at-risk families with the same germline variant is exceedingly unlikely for an individual research group and generally has required collaboration across multiple research groups. As a result, many individuals in the field, including myself, have focused on the development of patient-derived, contemporary model systems for functional analyses of potentially pathogenic variants identified in at-risk families. In order to address this gap, I performed a targeted, unbiased sequencing

approach of individuals and families with an HT/HM phenotype who had enrolled on research protocols at the University of Chicago. I hypothesized that, by performing WES and/or WGS in a population enriched for the HT/HM phenotype, I would identify novel genetic drivers of the HT/HM phenotype.

Table 1.2. Pathogenic germline *DDX41* variants are largely absent from the gnomAD and ExAC population databases despite *DDX41*-driven HHMs being the most common known HHM.

Gene	Variant	gnomAD frequency	ExAC frequency
DDX41	T403fs	Absent	Absent
DDX41	E300fs	Absent	Absent
DDX41	I256fs	Absent	Absent
DDX41	c.1098+2T>G	Absent	Absent
DDX41	R339L	Absent	Absent
DDX41	D336fs	Absent	Absent
DDX41	Q180X	Absent	Absent
DDX41	L237F	Absent	Absent
DDX41	c.572-1G>A	Absent	Absent
DDX41	c.415_418dup	Absent	Absent
DDX41	R124X	Absent	Absent
DDX41	E122X	Absent	Absent
DDX41	K108fs	Absent	Absent
DDX41	K102fs	Absent	Absent
DDX41	P78fs	Absent	Absent
DDX41	R53fs	Absent	Absent
DDX41	Q41X	2.0 x 10-5	2.0 x 10-5
DDX41	M1I	1.0 x 10-4	5.0 x 10-5

Chapter II

Distinct leukemogenic mechanisms exist in HHMs driven by germline *DDX41*, *GATA2*, or *RUNX1* mutations

The data in this chapter are adapted from the following manuscript, which is being prepared for submission: Homan CC*, Drazer MW*, Yu K*, King-Smith SL, Pozsgai MJ, Lawrence DM, Arts P, Feng J, Andrews J, Armstrong M, Ha T, Dobbins J, Bodor C, Cantor A, Cazzola M, Degelman E, Dinardo C, Duployez N, Favier R, Fröhling S, Fitzgibbon J, Klco J, Krämer A, Kurokawa M, Lee J, Malcovati L, Morgan N, Natsoulis G, Owen C, Patel KP, Preudhomme C, Raslova H, Rienhoff H, Ripperger T, Schulte R, Tawana K, Velloso E, Yuan B, Liu P, Schreiber AW, Hahn CN, Brown AL, Godley LA, Scott HS. Clonal hematopoiesis demonstrates syndrome-specific patterns in *DDX41*, *GATA2*, and *RUNX1* mutation carriers. 2021. *co-first authors

In this study, I designed and performed the experiments, analyzed the data, and wrote the manuscript.

Additional data in this chapter are adapted from the following publication in *Haematologica*: Homan CC, King-Smith SL, Lawrence DM, Arts P, Feng J, Andrews J, Armstrong M, Ha T, Dobbins J, Drazer MW, Yu K, Bodor C, Cantor A, Cazzola M, Degelman E, Dinardo C, Duployez N, Favier R, Fröhling S, Fitzgibbon J, Klco J, Krämer A, Kurokawa M, Lee J, Malcovati L, Morgan N, Natsoulis G, Owen C, Patel KP, Preudhomme C, Raslova H, Rienhoff H, Ripperger T, Schulte R, Tawana K, Velloso E, Yuan B, Liu P, Godley LA, Schreiber AW, Hahn CN, Scott HS, Brown AL. The RUNX1 database: a RUNX1 genomics database for mutation data aggregation, analysis, and sharing. *Haematologica.* Online ahead of print. July 2021. PMID: 34233450.

Introduction

Individuals with germline mutations in the HHM-associated genes *DDX41*, *GATA2*, or *RUNX1* experience an elevated lifetime risk for HMs. The penetrance for malignancies is variable: *DDX41* mutation carriers experience an 80% lifetime risk (median age of onset: 62 years), *GATA2* carriers experience a 90% lifetime risk (median age 30 years), and *RUNX1* carriers experience a 44% lifetime risk (median age 33 years) (Feurstein et al., 2016). As a group, these HHMs represent the most common HHM (*DDX41*), one of the two most highly penetrant HHMs (*GATA2*), and the first known HHM with the largest number of known kindreds worldwide (*RUNX1*).

However, no natural history study has been performed to date comparing these HHMs before (unaffected carriers) or after (affected carriers) the development of a malignancy. In particular, it is largely unknown if unaffected carriers of these HHM-associated mutations develop leukemogenic somatic mutations prior to the development of overt malignancies. Prior work addressing these questions has also focused on small numbers of carriers. For example, the Godley and Graubert groups sequenced a group of nine unaffected *RUNX1* carriers and demonstrated that six of these individuals (66%) carried somatic mutations consistent with CH. Remarkably, each individual with CH was under age 50, an age at which less than 1% of the general population is observed to experience CH (Genovese et al., 2014; Jaiswal et al., 2014). However the clinical and biological significance of these mutations was largely unclear aside from an individual with a clearly pathogenic *DNMT3A* p.R882H mutation (Churpek et al., 2015).

A second series included four unaffected carriers with *RUNX1* germline mutations, 3 of whom had evidence of CH (75%). In this second series, the individuals with CH were

49, 56, and 71 years of age. The somatic driver mutations were identified in *EZH2* (one patient), *SF3B1* (one patient), *SH2B3* (one patient), *SRSF2* (one patient), and *TET2* (two patients) (DiFilippo et al., 2020).

To date, six unaffected *GATA2* carriers have been sequenced over two studies (McReynolds et al., 2019; Wang et al., 2015). Only one (16.6%) unaffected carrier in this series of patients was found to have evidence of CH driven by a somatic *CEBPA* mutation (McReynolds et al., 2019).

No studies investigating CH in unaffected *DDX41* carriers have been performed to date. Of note, no study has systematically sequenced unaffected carriers of *DDX41*, *GATA2*, or *RUNX1* mutations using a uniform bioinformatics approach. This lack of a uniform bioinformatics approach is particularly problematic as significant variability exists among the most popular variant calling packages, most notably among insertion/deletions (indels) (Hwang et al., 2015). The variable calling of these indels may result in clinically meaningful discrepancies in terms of variant calls even when utilizing the same clinical sample.

To address this limitation, I performed the first systematic natural history study of both unaffected and affected individuals with germline *DDX41*, *GATA2*, or *RUNX1* mutations in collaboration with colleagues from the NIH and the University of South Australia. The goals of this study were threefold: 1) to determine the prevalence of CH among these HHMs; 2) to describe the mutational spectrum that develops during CH or in leukemic clones in HHM-related mutation carriers; 3) to determine whether CH-related somatic mutations persist in leukemic clones, or if they are “late” mutational events.

I hypothesized that each of these HHMs, which are not phenocopies, would each follow unique leukemogenic "steps" that varied based on the HHM of interest. I also hypothesized that CH would be more prevalent in HHM mutation carriers relative to population-based controls and that CH would occur at earlier ages in each HHM relative to population controls.

Methods

Table 2.1. Cohort of individuals enrolled on the natural history study.

Germline Mutations	# variants	# Individuals	# Probands	Pre-Leukaemic Individuals	Malignancy Individuals
DDX41	22	61	35	24	31
RUNX1	50	124	65	68	51
GATA2	7	21	8	9	12

Study participants

Patients were enrolled from three academic centers: the University of Chicago, the National Institutes of Health, and the University of South Australia. All patients participated on Institutional Review Board (IRB)-approved protocols that allowed for the sharing of anonymized patient sequencing data between the institutions. Each individual on the study provided informed consent. Parents/guardians of minors provided consent for their enrollment, and minors provided assent when appropriate and feasible. In total, 206 individuals were enrolled from 106 families, with the majority being carriers of *RUNX1* or *DDX41* mutations. The disease status of patients was evenly distributed, with 51% of the individuals being unaffected individuals and 49% patients already diagnosed with a

cancer. The age of the patients enrolled ranged from 1-100 years, with median ages of cancer diagnosis ranging from 24 (*GATA2* carriers) to 68 (*DDX41* carriers) (Table 2.1).

Each HHM-related germline variant was also reviewed by multiple team members and classified as pathogenic/likely pathogenic/VUS/likely benign/benign using criteria from the Association for Molecular Pathology and the American College of Human Genetics and Genomics (Richards et al., 2015). These classifications are shown in **Table 2.2**.

Table 2.2. Germline variants identified in *DDX41*, *GATA2*, and *RUNX1* patient cohorts.

c_hgvs	p_hgvs	zygosity	Classification	acmg_criteria	evidence_weights
NM_014915.2(ANKRD26):c.-118C>T	NP_055730.2:p.=	heterozygous	P	PM6_P, PM1, PM2, PP1_S, PS4	1xPP, 2xPM, 2xPS
NM_014915.2(ANKRD26):c.-128G>A	NP_055730.2:p.=	heterozygous	P	PM1, PM2, PP1_S, PS4	2xPM, 2xPS
NM_014915.2(ANKRD26):c.-119C>G	NP_055730.2:p.=	heterozygous	LP	PM1, PM2, PS4_M	3xPM
NM_016222.4(DDX41):c.435-2_435-1delAGinsCA	NP_057306.2:p.=	heterozygous			
NM_016222.4(DDX41):c.1474dupG	NP_057306.2:p.Ala492GlyfsTe	heterozygous	LP	PM2, PVS1	1xPM, 1xPVS
NM_016222.4(DDX41):c.1013G>A	NP_057306.2:p.Cys338Tyr	heterozygous	VUS B (3B)	PP3	1xPP
NM_016222.4(DDX41):c.415_418dupGATG	NP_057306.2:p.Asp140GlyfsTe	heterozygous	P	PS4_P, PVS1	1xPP, 1xPVS
NM_016222.4(DDX41):c.142C>T	NP_057306.2:p.Gln48Ter	heterozygous	VUS A (3A)	PVS1	1xPVS
NM_016222.4(DDX41):c.323delA	NP_057306.2:p.Lys108SerfsTe	heterozygous	P	PS4_P, PM2, PVS1	1xPP, 1xPM, 1xPVS
NM_016222.4(DDX41):c.3G>A	p.Met1?	heterozygous	P	PM5_P, PS3_P, PM4, PS4	2xPP, 1xPM, 1xPS
NM_016222.4(DDX41):c.773C>T	NP_057306.2:p.Pro258Leu	heterozygous	VUS B (3B)	PM1_P, PP3	2xPP
NM_016222.4(DDX41):c.232_233insAA	NP_057306.2:p.Pro78GlnfsTer	heterozygous	P	PS4_P, PM2, PVS1	1xPP, 1xPM, 1xPVS
NM_016222.4(DDX41):c.1285C>T	NP_057306.2:p.Gln429Ter	heterozygous	P	PS4_P, PM2, PVS1	1xPP, 1xPM, 1xPVS
NM_016222.4(DDX41):c.1016G>T	NP_057306.2:p.Arg339Leu	heterozygous	VUS A (3A)	PM1_P, PS4_P, PM2	2xPP, 1xPM
NM_016222.4(DDX41):c.490C>T	NP_057306.2:p.Arg164Trp	heterozygous			
NM_016222.4(DDX41):c.674C>T	NP_057306.2:p.Ala225Val	heterozygous			
NM_016222.4(DDX41)c.1574G>A	NP_057306.2:p.Arg525His	heterozygous			
NM_016222.4(DDX41):c.517G>A	NP_057306.2:p.Gly173Arg	heterozygous	LP	PP3, PM2, PS4	1xPP, 1xPM, 1xPS
NM_016222.4(DDX41):c.644+5G>C	NP_057306.2:p.=	heterozygous	LP	PM2, PVS1	1xPM, 1xPVS
NM_016222.4(DDX41):c.155dupA	NP_057306.2:p.Arg53AlafsTer	heterozygous	P	PS4_P, PM2, PVS1	1xPP, 1xPM, 1xPVS
NM_016222.4(DDX41)c.1547A>G	NP_057306.2:p.Tyr516Cys	heterozygous			
NM_016222.4(DDX41):c.121C>T	NP_057306.2:p.Gln41Ter	heterozygous			
NM_016222.4(DDX41):c.1105C>G	NP_057306.2:p.Arg369Gly	heterozygous			
NM_016222.4(DDX41):c.373+1G>A	NP_057306.2:p.=	heterozygous			
NM_016222.4(DDX41):c.1586_1587delCA	NP_057306.2:p.Thr529ArgfsTe	heterozygous			
ENST00000396373.4(ETV6):c.1106G>A	ENSP00000379658.3:p.Arg369	heterozygous			
NM_032638.5(GATA2):c.1034_1035insTCTTCTTGTG	NP_116027.2:p.Ala345_Gly346	heterozygous	VUS	PS3_P, PM2, PM4	1xPP, 2xPM
NA	Exon 4 deletion	heterozygous	P	PVS1, PS4_Supporting, PM2	
NM_032638.5(GATA2):c.857C>T	NP_116027.2:p.Ala286Val	heterozygous	LP	PP3, PS4_P, PM2, PS3	2xPP, 1xPM, 1xPS
NM_032638.5(GATA2):c.630_643del	NP_116027.2:p.Lys212ThrfsTe	heterozygous	P	PM2, PS4, PVS1	1xPM, 1xPS, 1xPVS
NM_032638.5(GATA2):c.1061C>T	NP_116027.2:p.Thr354Met	heterozygous	P	PP3, PM1, PM2, PM5, PP1_S, PS3	1xPP, 3xPM, 3xPS
NM_032638.5(GATA2):c.1082G>C	NP_116027.2:p.Arg361Pro	heterozygous			
NM_032638.5(GATA2):c.1017+572C>T	NP_116027.2:p.=	heterozygous	P	PM1, PM2, PP1_S, PS3, PS4	2xPM, 3xPS
NM_001754.4(RUNX1):c.958C>T	NP_001745.2:p.Arg320Ter	heterozygous	P	PM2, PS4_M, PP1_S, PVS1_S	2xPM, 2xPS
NM_001754.4(RUNX1):c.334delC	NP_001745.2:p.Leu112CysfsTe	heterozygous	P	PS4_P, PM2, PVS1	1xPP, 1xPM, 1xPVS
NM_001754.4(RUNX1):c.351+1G>A	NP_001745.2:p.=	heterozygous	P	PS4_P, PM2, PVS1	1xPP, 1xPM, 1xPVS
NM_001754.4(RUNX1):c.351+1G>T	NP_001745.2:p.=	heterozygous	P	PS4_P, PM2, PVS1	1xPP, 1xPM, 1xPVS
NM_001754.4(RUNX1):c.352-1G>C	NP_001745.2:p.=	heterozygous	P	PS4_P, PM2, PVS1	1xPP, 1xPM, 1xPVS
NM_001754.4(RUNX1):c.352-1G>A	NP_001745.2:p.=	heterozygous	LP	PS4_M, PVS1	1xPM, 1xPVS
NM_001754.4(RUNX1):c.506G>T	NP_001745.2:p.Arg169Ile	heterozygous	LP	PP1, PP3, PS4_P, PM1, PM2	3xPP, 2xPM
NM_001754.4(RUNX1):c.508+1G>T	NP_001745.2:p.=	heterozygous	P	PS4_P, PM2, PVS1	1xPP, 1xPM, 1xPVS
NM_001754.4(RUNX1):c.508+3delA	NP_001745.2:p.=	heterozygous	P	PP3, PS4_P, PM2, PP1_S, PS3	2xPP, 1xPM, 2xPS
NM_001754.4(RUNX1):c.593A>T	NP_001745.2:p.Asp198Val	heterozygous	LP	PP3, PM1, PM2, PS4_M	1xPP, 3xPM
NM_001754.4(RUNX1):c.98-1G>A	NP_001745.2:p.=	heterozygous	VUS_B	PS4_P, PM2	1xPP, 1xPM
NA	NA	heterozygous	P	PM1, PM4, PVS1	2xPM, 1xPVS
NM_001754.4(RUNX1):c.1412_1413dupGC	NP_001745.2:p.Leu472fs	heterozygous	P	PS4_P, PM2, PS3_M, PP1_S, PVS1	1xPP, 2xPM, 2xPS
NM_001754.4(RUNX1):c.425C>A	NP_001745.2:p.Ala142Asp	heterozygous	VUS_3	PM1_P, PP3, PS4_P, PM2	3xPP, 1xPM
NM_001754.4(RUNX1):c.163dupG	NP_001745.2:p.Ala55GlyfsTer	heterozygous	P	PS4_P, PM2, PP1_M, PVS1	1xPP, 2xPM, 1xPVS
NM_001754.4(RUNX1):c.496C>T	NP_001745.2:p.Arg166Ter	heterozygous	P	PS4_M, PP1_S, PVS1	1xPM, 1xPS, 1xPVS
NM_001754.4(RUNX1):c.601C>T	NP_001745.2:p.Arg201*	heterozygous	P	PM2, PP1_S, PS4, PVS1	1xPM, 2xPS, 1xPVS
NM_001754.4(RUNX1):c.602G>A	NP_001745.2:p.Arg201Gln	heterozygous	P	PP1, PP3, PM1, PM2, PS3, PS4	2xPP, 2xPM, 2xPS
NM_001754.4(RUNX1):c.610C>T	NP_001745.2:p.Arg204Ter	heterozygous	P	PM2, PP1_M, PS4_M, PVS1	3xPM, 1xPVS
NM_001754.4(RUNX1):c.611G>A	NP_001745.2:p.Arg204Gln	heterozygous	P	PP3, PM1, PM2, PS4_M, PP1_S	1xPP, 3xPM, 1xPS
NM_001754.4(RUNX1):c.698G>A	NP_001745.2:p.Arg233His	heterozygous	LB	BS1	1xBS
NA	Deletion of Exon 3-5	heterozygous	P	PS4_P, PM1, PM2, PM4, PVS1_S	1xPP, 3xPM, 1xPS
NM_001754.4(RUNX1):c.593A>T	NP_001745.2:p.Asp198Val	heterozygous	LP	PP3, PM1, PM2, PS4_M	1xPP, 3xPM
NM_001754.4(RUNX1):c.994delG	NP_001745.2:p.Asp332ThrfsTe	heterozygous	LP	PS4_P, PM2, PVS1_S	1xPP, 1xPM, 1xPS
NM_001754.4(RUNX1):c.554_560delAAGTCGC	NP_001745.2:p.Gln185ProfsTe	heterozygous	P	PS4_P, PM2, PVS1	1xPP, 1xPM, 1xPVS
NM_001754.4(RUNX1):c.999_1003dupGCGCC	NP_001745.2:p.Gln335ArgfsTe	heterozygous	LP	PS4_P, PM2, PVS1_S	1xPP, 1xPM, 1xPS
NM_001754.4(RUNX1):c.403G>A	NP_001745.2:p.Gly135Ser	heterozygous	LP	PM1_P, PM5_P, PP3, PS4_P, PM2	4xPP, 1xPM
NM_001754.4(RUNX1):c.404G>T	NP_001745.2:p.Gly135Val	heterozygous	LP	PM1_P, PP3, PM2, PS4_M	2xPP, 2xPM
NM_001754.4(RUNX1):c.506dupG	NP_001745.2:p.Gly170ArgfsTe	heterozygous	P	PS4_P, PM2, PVS1	1xPP, 1xPM, 1xPVS
NM_001754.4(RUNX1):c.649G>A	NP_001745.2:p.Gly217Arg	heterozygous	VUS_C		
NM_001754.4(RUNX1):c.1209_1210dupCC	NP_001745.2:p.His404ProfsTe	heterozygous	LP	PS4_P, PM2, PVS1_S	1xPP, 1xPM, 1xPS
NM_001754.4(RUNX1):c.431T>A	NP_001745.2:p.Leu144Gln	heterozygous	VUS_A	PP3, PM1, PM2	1xPP, 2xPM
NM_001754.4(RUNX1):c.582A>C	NP_001745.2:p.Lys194Asn	heterozygous	LP	PM1_P, PP3, PM2, PS4_M	2xPP, 2xPM
NM_001754.4(RUNX1):c.1163C>A	NP_001745.2:p.Ser388Ter	heterozygous	LP	PS3_P, PS4_P, PM2, PVS1_S	2xPP, 1xPM, 1xPS
NM_001754.4(RUNX1):c.442_449delACCGCAGC	NP_001745.2:p.Thr148fs	heterozygous	P	PS4_P, PM2, PVS1	1xPP, 1xPM, 1xPVS
NM_001754.4(RUNX1):c.524_539dup16	NP_001745.2:p.Thr181Aspfs*3	heterozygous	LP	PM2, PVS1	1xPM, 1xPVS
NM_001754.4(RUNX1):c.586A>G	NP_001745.2:p.Thr196Ala	heterozygous	LP	PP3, PS4_P, PM1, PM2, PM5	2xPP, 3xPM
NM_001754.4(RUNX1):c.587C>G	NP_001745.2:p.Thr196Arg	heterozygous	P	PP3, PM1, PM2, PS4_M, PP1_S	1xPP, 3xPM, 1xPS
NM_001754.4(RUNX1):c.587C>T	NP_001745.2:p.Thr196Ile	heterozygous	LP	PP3, PS4_P, PM1, PM2, PM5	2xPP, 3xPM
NM_001754.4(RUNX1):c.735delC	NP_001745.2:p.Thr246ArgfsTe	heterozygous	P	PS4_P, PM2, PP1_M, PS3, PVS1	1xPP, 2xPM, 1xPS, 1x
NM_001754.4(RUNX1):c.317G>A	NP_001745.2:p.Trp106Ter	heterozygous	P	PS4_P, PM2, PS1, PVS1	1xPP, 1xPM, 1xPS, 1x
NM_001754.4(RUNX1):c.318G>A	NP_001745.2:p.Trp106Ter	heterozygous	P	PM1_P, PS4_P, PM2, PS1, PVS1	2xPP, 1xPM, 1xPS, 1x
NM_001754.4(RUNX1):c.836G>A	NP_001745.2:p.Trp279Ter	heterozygous	P	PS4_P, PM2, PVS1	1xPP, 1xPM, 1xPVS
NM_001754.4(RUNX1):c.861C>A	NP_001745.2:p.Tyr287*	heterozygous	P	PS4_P, PM2, PP1_S, PVS1	1xPP, 1xPM, 1xPS, 1x
NM_001754.4(RUNX1):c.719delC	NP_001745.2:p.Pro240HisfsTe	heterozygous	P	PS4_P, PM2, PVS1	1xPP, 1xPM, 1xPVS
NA	Deletion of Exon 1-2	heterozygous	P	PP1, PM1, PM2, PVS1_M, PS4	1xPP, 3xPM, 1xPS
NM_001754.4(RUNX1):c.167T>C	NP_001745.2:p.Leu56Ser	heterozygous	B	BA1, BS3, BP2	1xBA, 1xBS, 1xBP
NM_001754.4(RUNX1):c.320G>C	NP_001745.2:p.Arg107Pro	heterozygous	LP	PM5_P, PP1, PP3, PS4_P, PM1, PM	4xPP, 2xPM
NM_001754.4(RUNX1):c.987delG	NP_001745.2:p.Phe330SerfsTe	heterozygous	LP	PS4_P, PM2, PVS1_S	1xPP, 1xPM, 1xPS
NA	Deletion of Exons 1-6	heterozygous	P	PM2, PM4, PS4_M, PVS1_S	3xPM, 1xPS

Tissue acquisition

Patients provided samples via peripheral blood (PB) draw, saliva collection, cultured skin fibroblasts, or bone marrow (BM) aspiration. Genomic DNA (gDNA) for University of Chicago samples was extracted using the QIAamp DNA Blood Mini Kit (Qiagen) per the manufacturer's instructions. DNA concentrations were then quantitated via a Nanodrop device (Thermo Scientific) and/or a Qubit fluorometer and reagents (Life Technologies). Of note, only cultured skin fibroblasts were considered to represent germline tissue for downstream analyses. Saliva, PB, and BM samples were considered to represent hematopoietic tissue.

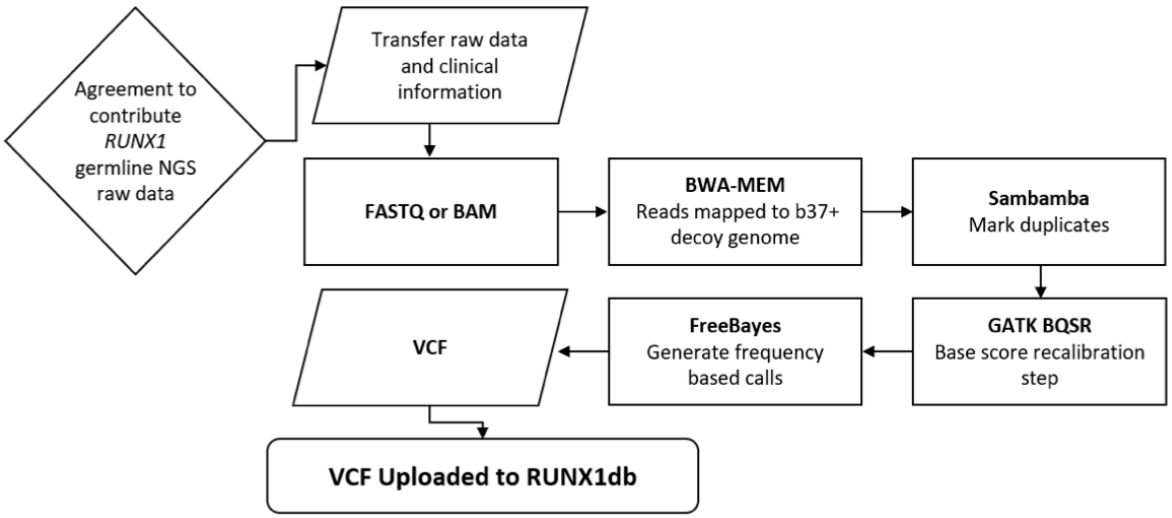
Whole exome sequencing

Genomic DNA from each sample was sheared, selected by size, ligated to adapters, and standard sequencing libraries were generated via PCR amplification. Following library generation, exome capture was performed using the human exome reagent (v3.0; Roche), and an additional PCR amplification with real-time quantitative PCR quantification was performed. An Illumina NovaSeq was used to sequence the capture libraries. Sequencing data were stored on a protected high-performance computing system at the University of Chicago that exceeds requirements for the Health Insurance Portability and Accountability Act. The data were then analyzed via a GATK-based bioinformatics pipeline at the University of Chicago and the NIH (McKenna et al., 2010).

Variant analysis using the RUNX1db bioinformatics pipeline

Following an initial review using the approach detailed in “Whole exome sequencing” above, FASTQ files were transferred to the University of South Australia in order to utilize the RUNX1db bioinformatic framework. The RUNX1 db allows for real-time variant calling and review. Briefly, BWA-MEM was used to map reads to a b37+ decoy genome, and duplicates were marked using Sambamba. A base score recalibration step was performed using GATK BQSR and FreeBayes was then utilized to generate frequency-based calls. The subsequent VCFs were uploaded to the RUNX1db for further review. This workflow has been published and is summarized in **Figure 2.1** (Homan et al., 2021).

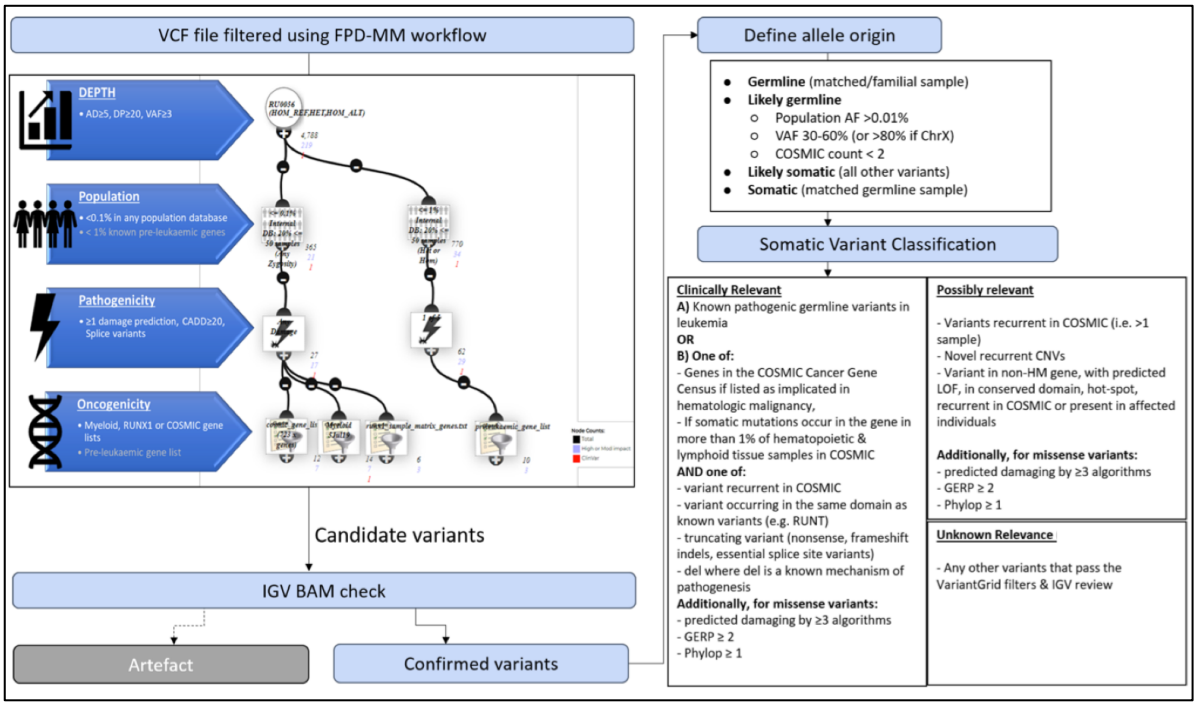
Figure 2.1. Bioinformatic workflow used for the RUNX1 database.



After uploading the VCFs to the RUNX1db, each germline sample was analyzed to confirm the presence of the expected germline variant in question. The samples from hematopoietic tissue (PB, BM, saliva) were then analyzed for the presence of the germline variant in question. The VCF files for each sample from hematopoietic tissue were then filtered for read quality and depth (thresholds: variant allelic depth >= 5, read

depth >= 20), population prevalence (variants that were present at 0.1% or higher in any population database were removed), pathogenicity (missense variants that were not predicted to be damaging in 2 or more in silico predictors were removed; CADD scores with values of 20 or higher were removed), and oncogenicity (variants not in genes with known roles as drivers in myeloid malignancies, not in the Catalogue of Somatic Mutations in Cancer [COSMIC], or *RUNX1* were removed). This filtering process, which is detailed in **Figure 2.2**, produced a list of candidate variants, and each variant was checked in the individual BAM files using IGV.

Figure 2.2. Filtering approach used for all variants in the *DDX41*, *GATA2*, and *RUNX1* natural history study. Filters for read depth, population frequency, pathogenicity, and oncogenicity were utilized to generate a list of candidate variants. Manual IGV BAM check was performed to confirm the presence of the variant in question. Allelic origin – somatic or germline – was determined and all somatic variants were classified as "clinically significant", as "possibly clinically relevant", or as of "unknown relevance".



Variants deemed to be artifacts were removed. Confirmed variants were then analyzed to determine if they were of germline or somatic origin. For samples with

germline sequencing data, variants in hematopoietic tissue were compared directly to data from germline tissue. For any samples with unpaired germline tissue, the potential germline origin of a variant in question was analyzed using a combination of population allelic frequency (minor allele threshold of 0.01% or lower), VAF (likely germline VAFs considered to be between 30 and 60% for genes on autosomal chromosomes and 80% or higher for genes on the X chromosome), and the frequency of the variant in question, with likely germline variants occurring no more than two times in COSMIC.

This filtering process then produced a list of variants of likely somatic or somatic origin which was then reviewed for clinical and biological relevance. “Clinically relevant” variants consisted of either known pathogenic germline variants in leukemia or variants that were prevalent in the COSMIC database. Novel driver variants were considered clinically relevant if they occurred in a gene that was recurrently mutated in hematopoietic and lymphoid samples in COSMIC, the variant was in the same domain as known pathogenic variants (for example, the RUNT domain in RUNX1), was a truncating variant (nonsense, frameshift indels, essential splice site variants), or was a deletion where deletion is a known mechanism of disease for the gene in question. Missense variants were only considered to be clinically relevant if they were predicted to be damaging in at least three in silico algorithms and were highly conserved via GERP and Phylop scores.

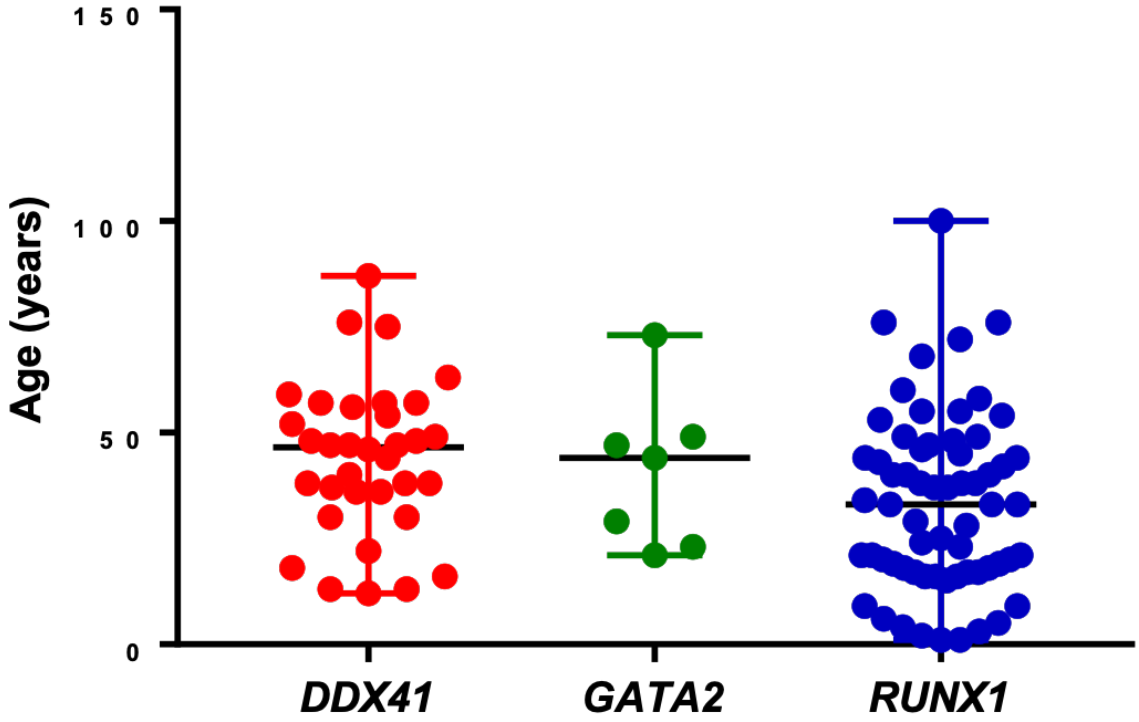
All somatic and likely somatic variants that were not clinically relevant were then categorized as either “possibly relevant” or of “unknown relevance”. Possibly relevant variants were recurrent in COSMIC (one sample or more), were novel recurrent CNVs, or were variants in genes not known to be associated with HMs, but which were predicted to be loss-of-function variants in conserved domains, were mutational hot spots, were

recurrent in COSMIC, or were present in affected individuals. All “possibly relevant” missense variants also needed to be predicted damaging on 3 or more in silico algorithms and highly conserved using the GERP and Phylop scores. Somatic variants of “unknown relevance” were any variants that passed the filtering approach described above and passed IGV review, but did not meet criteria for “clinically relevant” or “possibly relevant” variants.

Results

Among the 206 individuals enrolled in the study, 124 individuals (60%) had germline *RUNX1* variants, 61 individuals (30%) had germline *DDX41* variants, and 21 individuals (10%) had germline *GATA2* variants. In total, there were 79 unique *RUNX1*, *DDX41*, or *GATA2* LP/P variants. The age of pre-leukemic, unaffected carriers at the age of sample collection is shown in **Figure 2.3**; this demonstrates that the age of patients for each HHM were relatively evenly distributed, with the widest age distribution occurring in the *RUNX1*-mutated cohort (1-100 years).

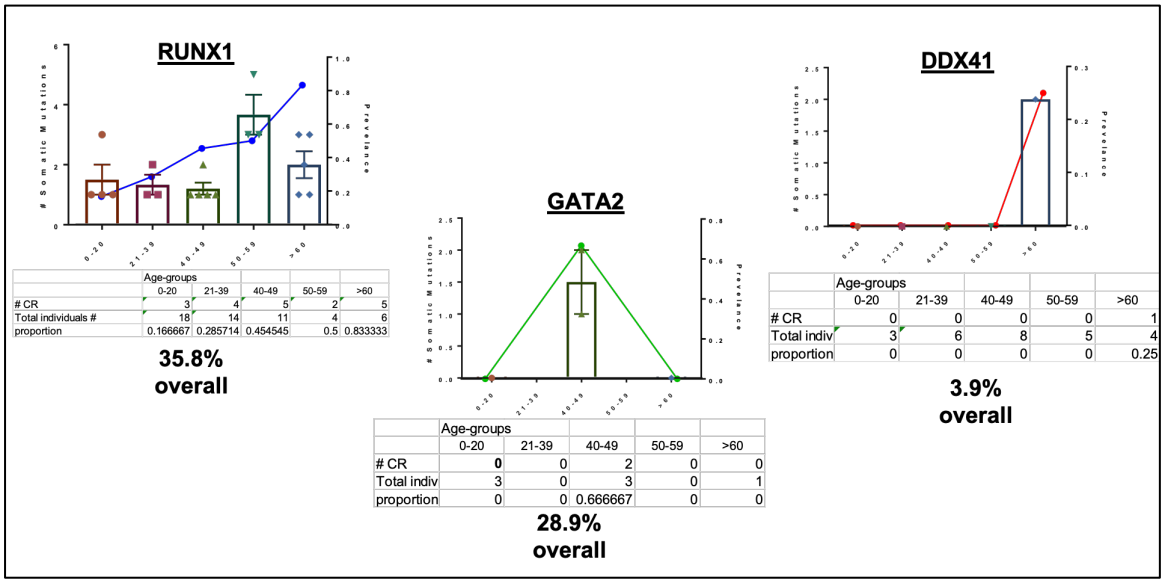
Figure 2.3. Age of pre-leukemic, unaffected individuals at the time of sample collection in the *DDX41*, *GATA2*, and *RUNX1* natural history study.



Patterns of clonal hematopoiesis in individuals with DDX41, GATA2, or RUNX1 germline mutations

Analysis of the unaffected individuals demonstrated that *GATA2* and *RUNX1* mutation carriers were significantly more likely to develop CH as compared to *DDX41* mutation carriers (**Figure 2.4)**. Overall, 35.8% of *RUNX1* mutation carriers and 28.9% of *GATA2* mutation carriers had evidence of CH, as compared to only 3.9% of *DDX41* mutation carriers. Age is the strongest predictor for CH in the general population (Genovese et al., 2014; Jaiswal et al., 2014), with only 1% of individuals under the age of 50 experiencing CH.

Figure 2.4. Prevalence of clonal hematopoiesis by age group and HHM of interest. CH was significantly more common in *RUNX1* mutation carriers (35.8%) and *GATA2* mutation carriers (28.9%) as compared to *DDX41* mutation carriers (3.9%).



	Age-groups				
	0-20	21-39	40-49	50-59	>60
# CR	3	4	5	2	5
Total individuals #	18	14	11	4	6
proportion	0.166667	0.285714	0.454545	0.5	0.833333

35.8% overall

	Age-groups				
	0-20	21-39	40-49	50-59	>60
# CR	**0**	0	2	0	0
Total indiv	3	0	3	0	1
proportion	0	0	0.666667	0	0

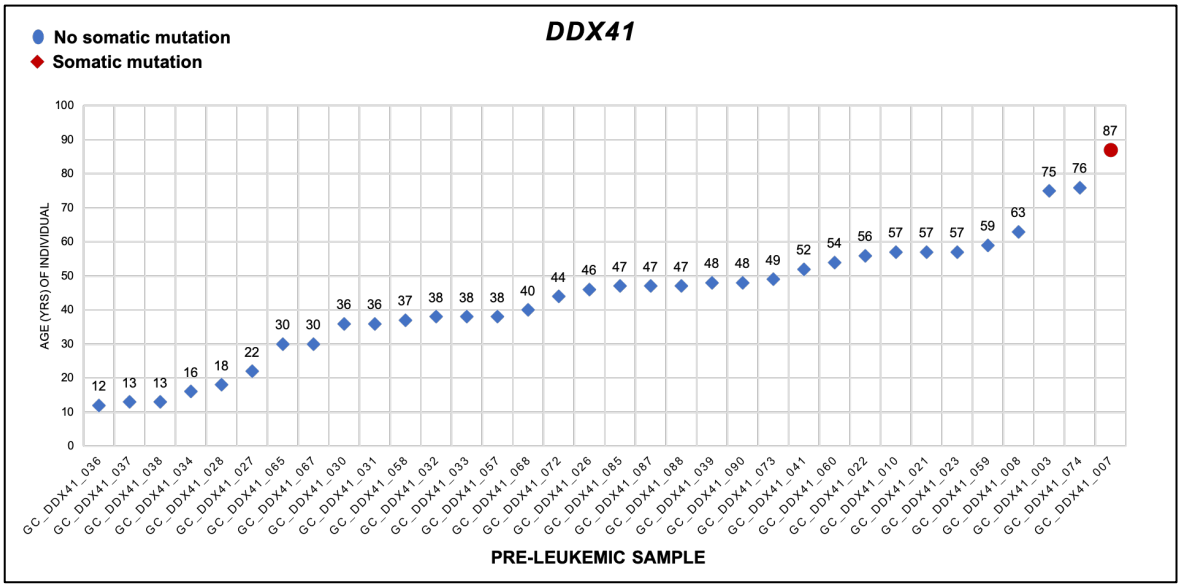
28.9% overall

	Age-groups				
	0-20	21-39	40-49	50-59	>60
# CR	0	0	0	0	1
Total indiv	3	6	8	5	4
proportion	0	0	0	0	0.25

3.9% overall

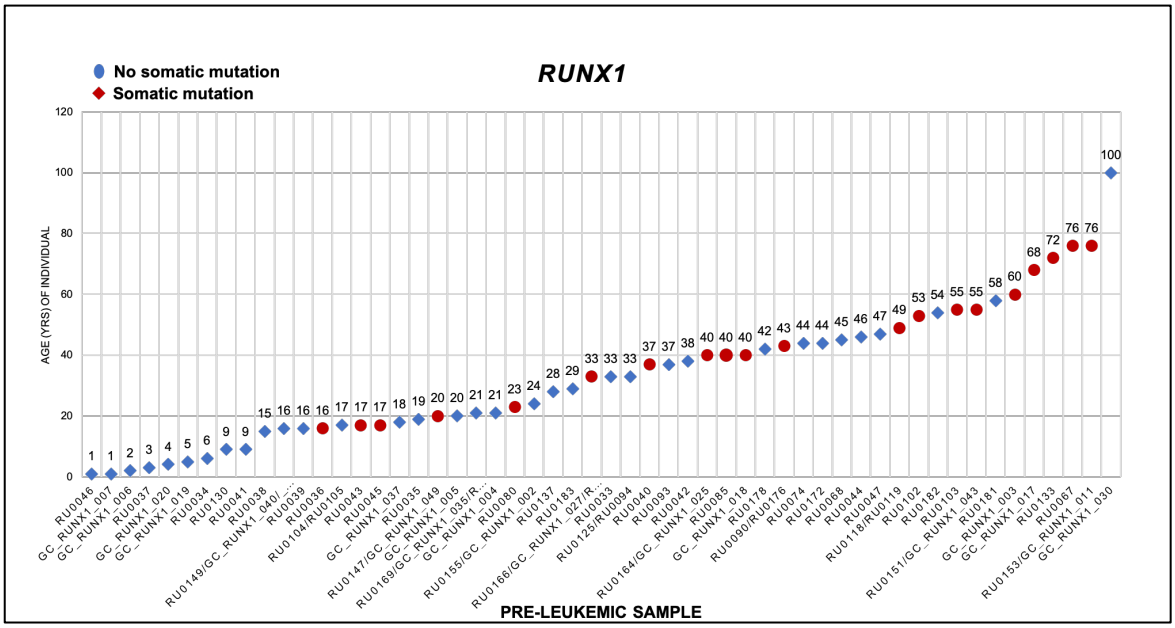
RUNX1 and *DDX41* carriers also showed distinct patterns of CH relative to age of onset. My data demonstrate that CH remains an age-related phenomenon in the HHMs, with the oldest individuals in each cohort being at the highest risk for CH. For example, although *DDX41* carriers were largely unaffected by CH even at older ages, the only individual in the *DDX41* cohort with CH was also the oldest individual in the cohort (age 87). Multiple *DDX41* carriers in their 40s, 50s, 60s, and 70s, were sampled but did not have CH (**Figure 2.5**).

Figure 2.5. Presence of clonal hematopoiesis by age in unaffected *DDX41* carriers. The oldest patient in the cohort (age 87) had evidence of CH (red diamond), but the remainder of patients in the cohort did not have CH (blue).



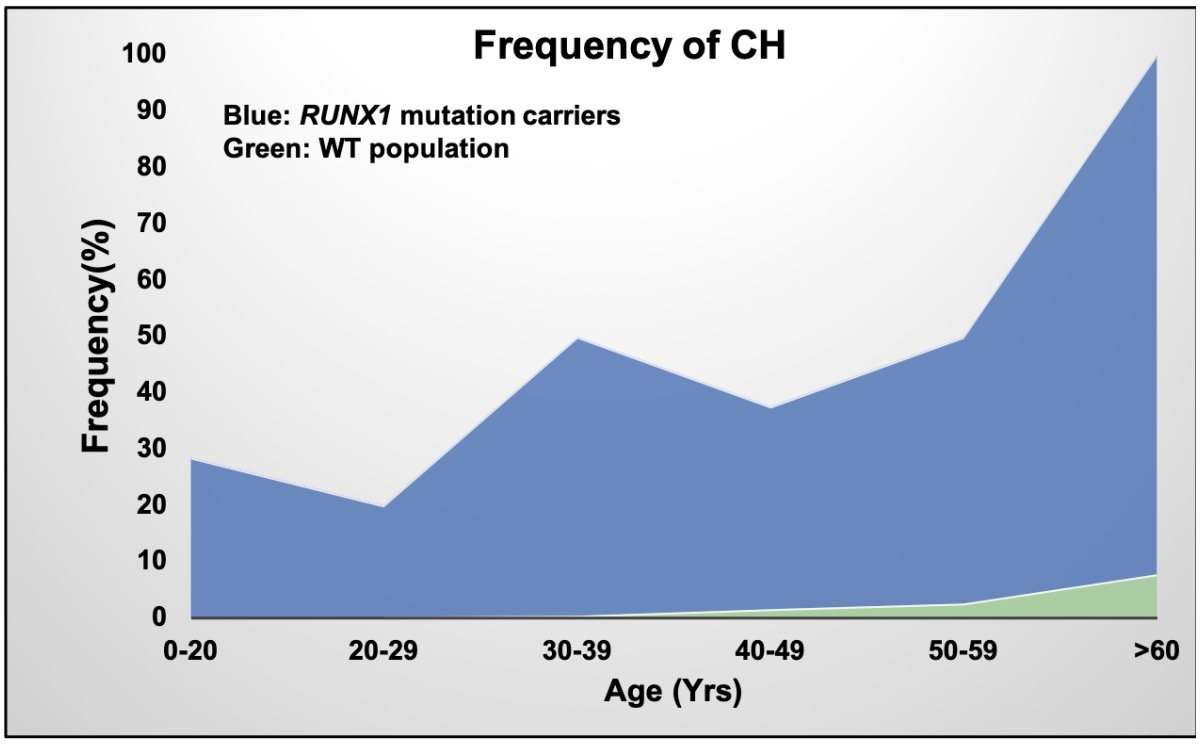
In contrast, *RUNX1* mutation carriers had early onset CH, and the first carrier with CH was identified at age 16 (**Figure 2.6**). Once again, the prevalence of CH increased with age in the *RUNX1* cohort, and CH was more prevalent in *RUNX1* mutation carriers than population controls. For example, a large proportion of *RUNX1* mutation carriers in their teens (16.7%), 20 and 30s (28.6%), and 40s (45.5%) were ultimately found to have CH (**Figure 2.7**). Of particular note, whereas prior work in the field demonstrated that 66% of germline *RUNX1* mutation carriers showed evidence of CH prior to the age of 50 in a relatively small patient series, my work has revised that estimate downward to 27.9% for all *RUNX1* mutation carriers under the age of 50 years and to 35.8% for germline *RUNX1* mutation carriers at any age (Churpek et al., 2015).

Figure 2.6. Presence of clonal hematopoiesis by age in unaffected *RUNX1* carriers. No patients under age 16 had CH (red). While age was associated with CH, the oldest patient in the cohort (100 years) did not have CH (blue).



These findings suggest that early onset CH is not a predominant leukemogenic mechanism in *DDX41* carriers, but early onset CH is common in *RUNX1* mutation carriers. Alternatively, the CH variants in *DDX41* carriers may be difficult to detect via short-read sequencing if the nature of CH-related variants in *DDX41* carriers differs from *RUNX1* or *GATA2* carriers. For example, CH driven by somatic CNVs could explain the relative paucity of CH events detected in pre-leukemic *DDX41* mutation carriers relative to individuals with *GATA2* or *RUNX1* mutations. However, I was unable to assess CNV-driven CH in any of the HHM cohorts, so it is not possible to determine if CNV-driven CH is a phenomenon in *DDX41*, *GATA2*, or *RUNX1* germline mutation carriers.

Figure 2.7. The prevalence of CH in *RUNX1* mutation carriers relative to a control population. CH is significantly more common in each age group relative to population controls.

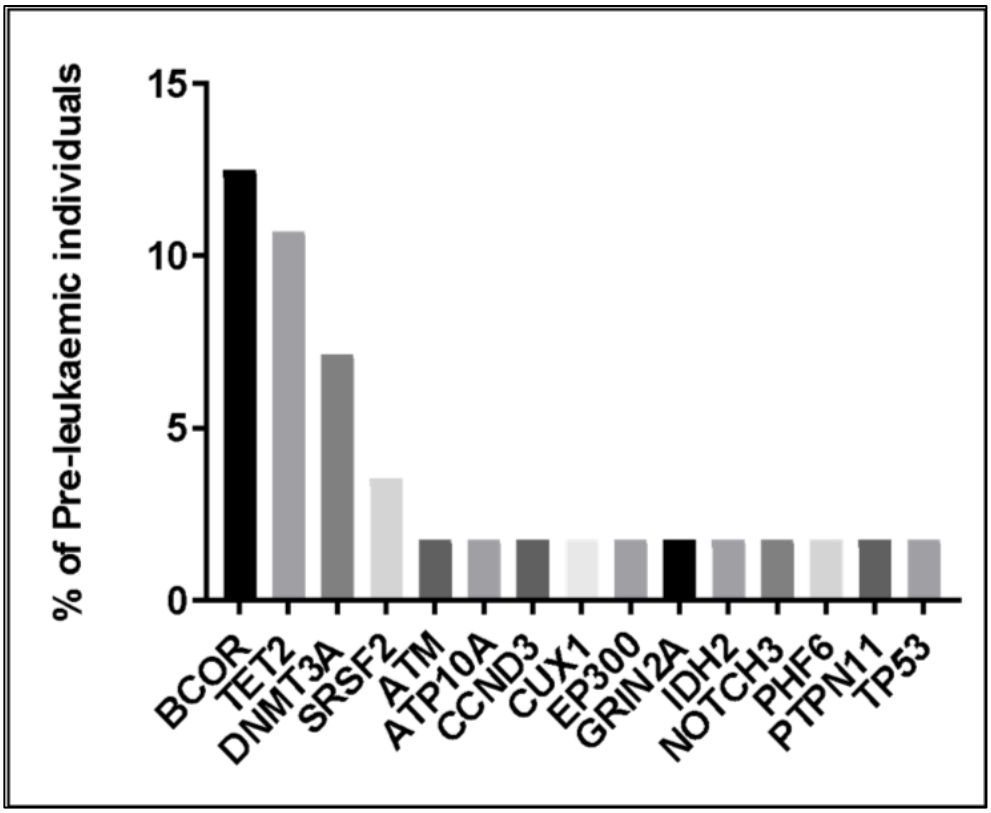


Control/WT Population : Jaiswal S et al. NEJM 2014, McCarroll S *et al.* 2014

Previously, it was unknown if the genes mutated during CH in unaffected individuals with HHM-related mutations (individuals with CH) are similar to those mutated in population controls who develop CH. To address this knowledge gap, I first ranked the genes mutated during CH in individuals with *RUNX1* germline mutations by frequency. This analysis demonstrated that the most prevalent genes involved in CH in *RUNX1* mutation carriers in decreasing order were *BCOR*, *TET2*, *DNMT3A*, and *SRSF2* (**Figure 2.8**). Next, I compared these genes to those that are mutated during CH in the general population. These results demonstrate that *BCOR*-driven CH is the predominant mechanism of CH within the *RUNX1*-mutated cohort, but *BCOR*-driven CH is not a

predominant mechanism of CH or leukemogenesis within population-based controls (**Figure 2.9**).

Figure 2.8. *BCOR* mutations are the most commonly observed mutations in *RUNX1* mutation carriers with CH, followed by *TET2*, *DNMT3A*, and *SRSF2* mutations.



Similarly, *SRSF2*-driven CH is over-represented within the *RUNX1* population as compared to population controls with CH. CH-related mutations in the epigenetic regulators *DNMT3A* and *TET2* are seen in both populations, but *DNMT3A* mutations are under-represented within the *RUNX1*-mutated cohort.

Next, I examined the mutational profile of individual patients with CH within the *RUNX1*-mutated cohort in order to determine how patterns of CH varied based on the age of the individual in question. This analysis demonstrates that younger individuals with *RUNX1* mutations were more likely to have CH driven by only one mutation up to the age

of 50, at which point all individuals in the study with CH had at least two somatic CH-related mutations detected (**Figure 2.10**). Intriguingly, both the youngest and oldest patients with CH in the *RUNX1*-mutated cohort (16 and 76 years of age, respectively) had *BCOR*-driven CH. Furthermore, *BCOR*-driven CH was observed across all age groups between these two extremes of age within our patient cohort. The VAFs of the CH-related variants within the *RUNX1* cohort ranged from 1.2% to 69.4%, with a mean VAF of 16.8%. In contrast, the mean VAF for CH-related somatic mutations in germline *GATA2* mutation carriers was 6.1%, and the VAF for the only *DDX41* germline mutation carrier with a CH-related somatic mutation was 4.1%.

Table 2.3 summarizes the somatic CH-related mutations and germline mutations that were observed in each individual patient in the study.

Figure 2.9. *BCOR* and *SRSF2* mutations are more common in *RUNX1* mutation carriers with CH (n=20) than population-based controls with CH (n = 959). Both cohorts have CH driven by *DNMT3A* or *TET2* mutations, but *DNMT3A* mutations are more common in controls. *TET2* mutations are enriched in *RUNX1* mutation carriers. Control populations from Desai *et al.*, Abelson *et al.*, and Jaiswal *et al.*

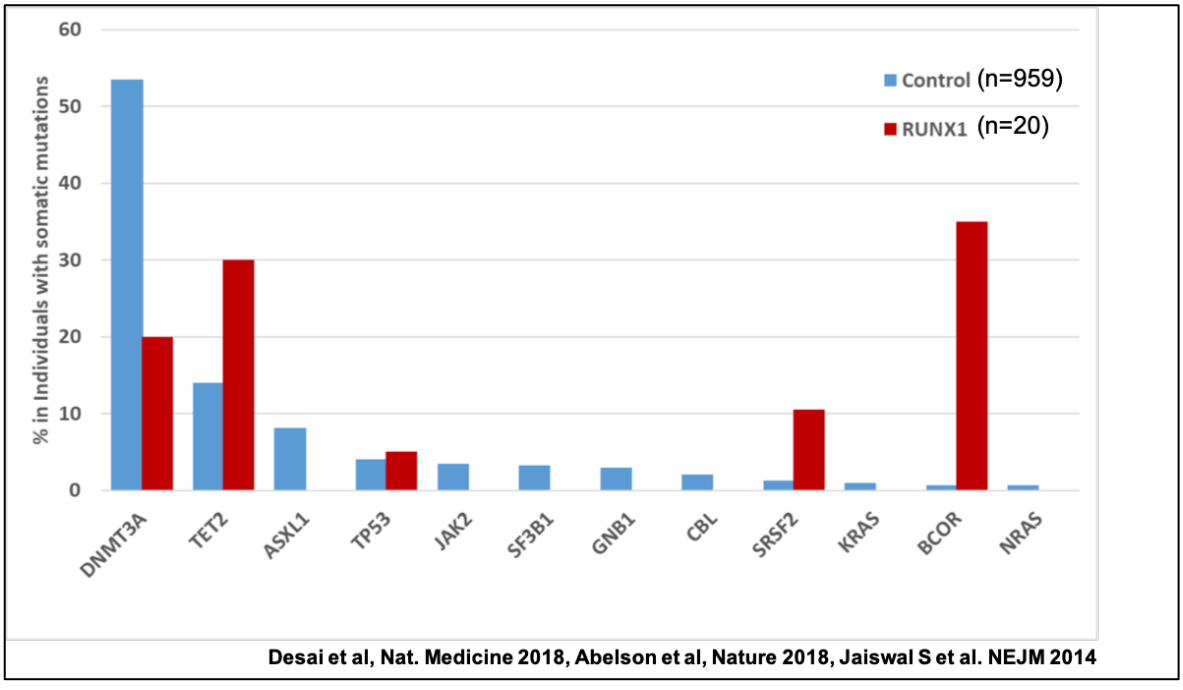
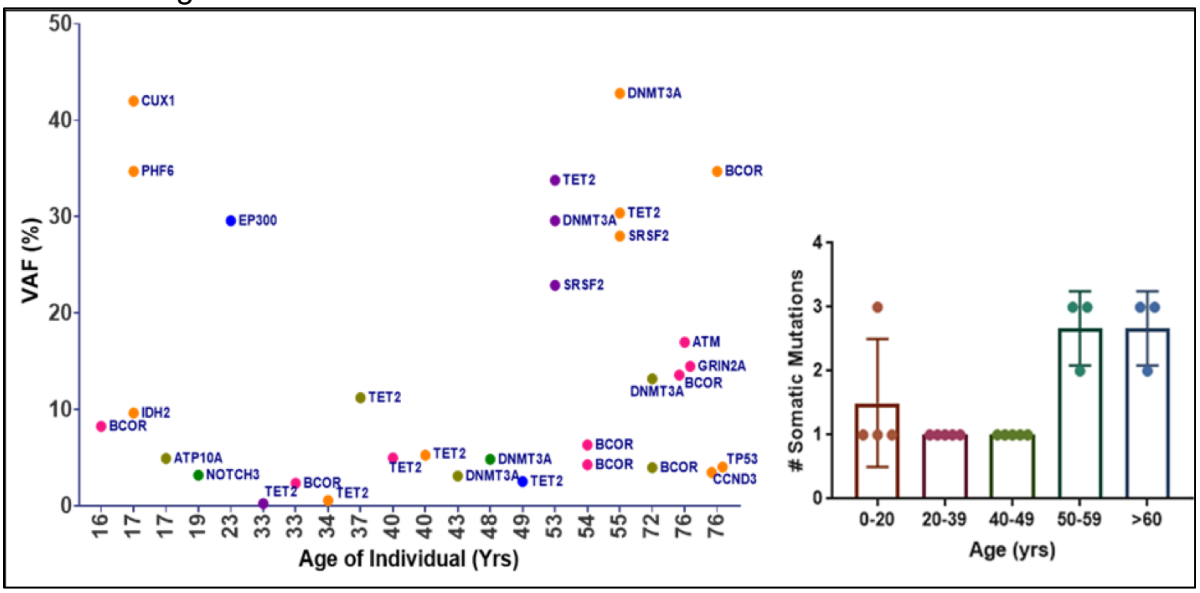


Figure 2.10. Somatic mutational spectrum among individuals in the *RUNX1* cohort with clonal hematopoiesis. *BCOR* mutations were seen in the youngest and oldest patients with CH, as well as in patients from intermediate age groups. Most patients under the age of 50 with CH only carried 1 CH-related somatic mutation; most individuals age 50 and older carried at least two somatic CH-related mutations.



Mutational patterns among individuals with germline mutations in DDX41, GATA2, or RUNX1 who develop hematopoietic malignancies

No group has performed a large-scale analysis of both unaffected (no malignancy diagnosed) and affected (malignancy diagnosed) individuals with HHM-related germline mutations. In order to address this knowledge gap, I sequenced DNA collected from the hematopoietic tissue (peripheral blood or bone marrow aspirate) of individuals with germline HHM-related mutations in *DDX41*, *GATA2*, or *RUNX1* who developed overt blood cancers, such as MDS or AML. The age of each individual at the time of malignant sample collection is shown in **Figure 2.11.** The median age of individuals with germline *DDX41* mutations who developed malignancies was higher than that of individuals with germline *RUNX1* or *GATA2* mutations who developed malignancies. This confirms prior work in the field that has shown germline *DDX41* mutation carriers develop HMs at a later

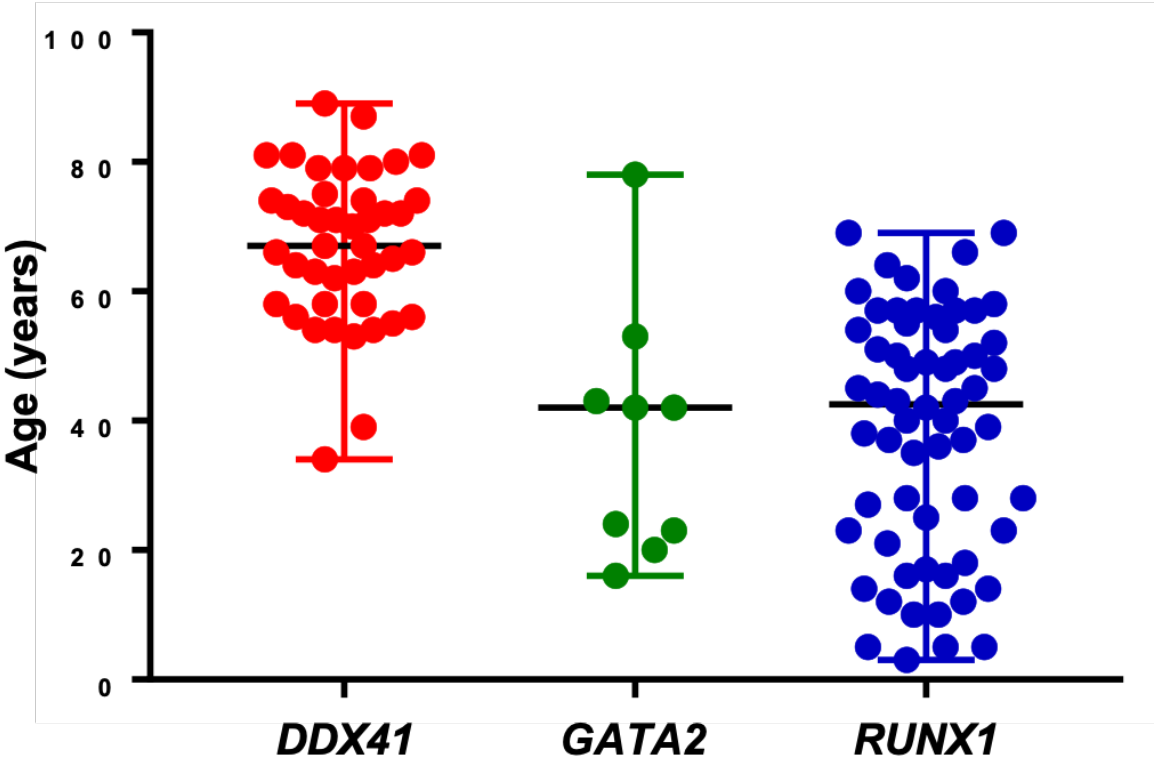
age (median age of 62 years) as compared to other HHMs (Lewinsohn et al., 2015; Polprasert et al., 2015).

Table 2.3. List of somatic and germline mutations identified in each individual with CH in the sequencing cohort.

Gene	Somatic Mutation				Germline Mutation				Pre-Leukaemic Phenotpye		
	Genomic coordinates	c_HGVS	p_HGVS	VAF(%)	Gene	Genomic coordinates	c_HGVS	p_HGVS	Clinical Presentation	Age	Sex
DNMT3A	2:25457171 T>A	NM_022552.5:c.2716A>T	p.Lys906Ter	3.557	*GATA2*	3:128202131 G>A	NM_032638.5:c.1017+572C>T	p.?	Unaffected	44	M
	2:25467022 A>G	NM_022552.5:c.1851+2T>C	p.?	4.109589	*DDX41*	5:176942945 T>TCATC	NM_016222.4:c.415_418dupGATG	p.Asp140GlyfsTer2	Unaffected	87	F
	2:25497943 CG>C	NM_022552.5:c.505del	p.Arg169GlyfsTer56	29.614325	*RUNX1*	21:36259139 C>T	NM_001754.4:c.351+1G>A	p.?	Thrombocytopenia	53	F
	2:25497943 CG>C	NM_022552.5:c.505del	p.Arg169GlyfsTer57	42.84078	*RUNX1*	21:36259139 C>T	NM_001754.4:c.351+1G>A	p.?	Thrombocytopenia	56	F
	2:25464660 C>T	NM_022552.5:c.2053G>A	p.Gly685Arg	3.1910484	*RUNX1*	21:36231773 C>T	NM_001754.4:c.611G>A	p.Arg204Gln	Thrombocytopenia	NA	M
	2:25466797 C>T	NM_022552.5:c.1906G>A	p.Val636Met	24.726706	*RUNX1*	21:36231773 C>T	NM_001754.4:c.611G>A	p.Arg204Gln	Thrombocytopenia	NA	M
	2:25470029 T>C	NM_022552.5:c.1015-2A>G	p.?	13.28486	*RUNX1*	21:36,421,595-36,231,771	NA	Ex1-6 del	Thrombocytopenia	72	F
	2:25463242 AG>A	NM_022552.5:c.2250del	p.Phe751SerfsTer28	3.1189084	*RUNX1*	21:36252856 C>A	NM_001754.4:c.506G>T	p.Arg169Ile	Thrombocytopenia	43	M
	2:25463242 AG>A	NM_022552.5:c.2250del	p.Phe751SerfsTer28	4.8638134	*RUNX1*	21:36252856 C>A	NM_001754.4:c.506G>T	p.Arg169Ile	Thrombocytopenia	48	M
KMT2C	7:151945114 G>T	NM_170606.3:c.2405C>A	p.Ser802Ter	8.0701	*GATA2*	3:128202131 G>A	NM_032638.5:c.1017+572C>T	p.?	Unaffected	44	M
	7:151945114 G>T	NM_170606.3:c.2405C>A	p.Ser802Ter	9.487	*GATA2*	3:128202131 G>A	NM_032638.5:c.1017+572C>T	p.?	Adrenal Ca	47	M
KDM5A	12:416952 C>CT	NM_001042603.3:c.3597dup	p.Gly1200ArgfsTer7	3.22	*GATA2*	3:128202131 G>A	NM_032638.5:c.1017+572C>T	p.?	Unaffected	16.5	F
BCOR	X:39923092 TA>T	NM_001123385.2:c.3615del	p.Lys1207AsnfsTer31	3.6144578	*RUNX1*	21:36231773 C>T	NM_001754.4:c.611G>A	p.Arg204Gln	Thrombocytopenia	NA	M
	X:39911577 GA>G	NM_001123385.2:c.5052del	p.Pro1685GlnfsTer40	4.958678	*RUNX1*	21:36171729 C>T	NM_001754.4:c.836G>A	p.Trp279Ter	Unaffected	55	M
	X:39916476 C>T	NM_001123385.2:c.4527G>A	p.Trp1509Ter	27.272728	*RUNX1*	21:36259327 G>GC	NM_001754.4:c.163dupG	p.Ala55GlyfsTer?	Thrombocytopenia	76	M
	X:39921490 TG>T	NM_001123385.2:c.4329del	p.Thr1444ProfsTer40	69.44444	*RUNX1*	21:36259327 G>GC	NM_001754.4:c.163dupG	p.Ala55GlyfsTer?	Thrombocytopenia	76	M
	X:39921617 CG>C	NM_001123385.2:c.4202del	p.Pro1401ArgfsTer83	41.666668	*RUNX1*	21:36259171 C>G	NM_001754.4:c.320G>C	p.Arg107Pro	Thrombocytopenia	40	M
	X:39923055 C>T	NM_001123385.2:c.3653G>A	p.Trp1218Ter	8.609271	*RUNX1*	21:36171729 C>T	NM_001754.4:c.836G>A	p.Trp279Ter	Unaffected	55	M
	X:39932109 ACT>A	NM_001123385.2:c.2488_2489del	p.Ser830CysfsTer6	17.592592	*RUNX1*	21:36276047-36438799	NA	Deletion of Exon 1-2	Unaffected	60	F
	X:39932334 G>GT	NM_001123385.2:c.2264dup	p.Tyr755Ter	4.7619047	*RUNX1*	21:36231774 G>A	NM_001754.4:c.610C>T	p.Arg204Ter	Thrombocytopenia	33	M
	X:39932898 T>TG	NM_001123385.2:c.1700dup	p.Ala568SerfsTer43	13.333333	*RUNX1*	21:36171729 C>T	NM_001754.4:c.836G>A	p.Trp279Ter	Unaffected	55	M
	X:39933270 CA>C	NM_001123385.2:c.1328del	p.Leu443CysfsTer8	8.292683	*RUNX1*	21:36259139 C>A	p.?	c.351+1G>T	Thrombocytopenia	16	F
	X:39933373 TGCCC>T	NM_001123385.2:c.1222_1225del	p.Gly408MetfsTer33	3.0671296	*RUNX1*	21:36231773 C>T	NM_001754.4:c.611G>A	p.Arg204Gln	Thrombocytopenia	NA	M
	X:39933379 G>T	NM_001123385.2:c.1220C>A	p.Pro407His	3.299941	*RUNX1*	21:36231773 C>T	NM_001754.4:c.611G>A	p.Arg204Gln	Thrombocytopenia	NA	M
	X:39933416 TG>T	NM_001123385.2:c.1182del	p.Lys395ArgfsTer47	1.4778326	*RUNX1*	21:36171729 C>T	NM_001754.4:c.836G>A	p.Trp279Ter	Thrombocytopenia	54	M
	X:39933492 TG>T	NM_001123385.2:c.1106del	p.Ser369Ter	4.004657	*RUNX1*	21:36,421,595-36,231,771	NA	Deletion of Exons 1-6	Thrombocytopenia	72	F
	X:39933593 A>AG	NM_001123385.2:c.1005dup	p.Ser336LeufsTer45	14.285714	*RUNX1*	21:36259171 C>G	NM_001754.4:c.320G>C	p.Arg107Pro	Thrombocytopenia	68	M
	X:39933676 TG>T	NM_001123385.2:c.922del	p.Gln308ArgfsTer70	2.3529413	*RUNX1*	21:36171729 C>T	NM_001754.4:c.836G>A	p.Trp279Ter	Thrombocytopenia	54	M
SRSF2	17:74732959 G>A	NM_003016.4:c.284C>T	p.Pro95Leu	22.93777	*RUNX1*	21:36259139 C>T	NM_001754.4:c.351+1G>A	p.?	Thrombocytopenia	53	F
	17:74732959 G>A	NM_003016.4:c.284C>T	p.Pro95Leu	28.037073	*RUNX1*	21:36259139 C>T	NM_001754.4:c.351+1G>A	p.?	Thrombocytopenia	56	F
	17:74732959 G>T	NM_003016.4:c.284C>A	p.Pro95His	13.946588	*RUNX1*	21:36231791 T>A	NM_001754.4:c.593A>T	p.Asp198Val	Thrombocytopenia	NA	F
TET2	4:106157215 C>T	NM_001127208.3:c.2116C>T	p.Gln706Ter	7.3333335	*RUNX1*	21:36171704 G>T	NM_001754.4:c.861C>A	p.Tyr287*	NA	NA	F
	4:106158510 T>C	NM_001127208.3:c.3409+2T>C	p.?	33.31424	*RUNX1*	21:36259139 C>T	NM_001754.4:c.351+1G>A	p.?	Thrombocytopenia	53	F
	4:106158510 T>C	NM_001127208.3:c.3409+2T>C	p.?	30.469963	*RUNX1*	21:36259139 C>T	NM_001754.4:c.351+1G>A	p.?	Thrombocytopenia	56	F
	4:106162587 G>A	NM_001127208.3:c.3500+1G>A	p.?	2.5641026	*RUNX1*	21:36171704 G>T	NM_001754.4:c.861C>A	p.Tyr287*	Thrombocytopenia	49	M
	4:106162587 G>A	NM_001127208.3:c.3500+1G>A	p.?	1.2269939	*RUNX1*	21:36171704 G>T	NM_001754.4:c.861C>A	p.Tyr287*	NA	NA	M
	4:106164068 G>A	NM_001127208.3):c.3578G>A	p.Cys1193Tyr	11.25	*RUNX1*	21:36231773 C>T	NM_001754.4:c.611G>A	p.Arg204Gln	Thrombocytopenia	37	F
	4:106164787 C>T	NM_001127208.3:c.3655C>T	p.His1219Tyr	5.298013	*RUNX1*	21:36231774 G>A	NM_001754.4:c.610C>T	p.Arg204Ter	Thrombocytopenia	40	M
	4:106196461 T>G	NM_001127208.3:c.4794T>G	p.Tyr1598Ter	5	*RUNX1*	21:36171607 G>A	NM_001754.4:c.958C>T	p.Arg320Ter	Thrombocytopenia	40	F
APC	5:112102943_112102943del	NM_000038.6:c.278_280delinsACG	p.Leu93_Arg94delinsHisGly	40.925266	*RUNX1*	21:36276047-36438799	NA	Deletion of Exon 1-2	Thrombocytopenia	23	M
ATM	11:108213949 G>A	NM_000051.3:c.8269G>A	p.Val2757Met	17.021276	*RUNX1*	21:36259327 G>GC	NM_001754.4:c.163dupG	p.Ala55GlyfsTer?	Thrombocytopenia	76	M
ATP10A	15:25953381 G>A	NM_024490.3:c.2411C>T	p.Ala804Val	4.9504952	*RUNX1*	21:36259394 C>T	NM_001754.4:c.98-1G>A	p.?	Thrombocytopenia	17	M
ATR	3:142281940 A>G	NM_001184.4:c.304T>C	p.Trp102Arg	3.8690476	*RUNX1*	21:36231791 T>A	NM_001754.4:c.593A>T	p.Asp198Val	Thrombocytopenia	NA	F
CCND3	6:41903707 G>A	NM_001760.4:c.850C>T	p.Pro284Ser	3.508772	*RUNX1*	21:36259327 G>GC	NM_001754.4:c.163dupG	p.Ala55GlyfsTer?	Thrombocytopenia	76	M
CUX1	7:101882720 G>A	NM_001202543.2:c.3776G>A	p.Arg1259Gln	42.285713	*RUNX1*	21:36259139 C>A	NM_001754.4:c.351+1G>T	p.?	Thrombocytopenia	17	M
EP300	22:41574723 TCCACACCA	NM_001429.3:c.7014_7028del15	p.His2338_Pro2342del	29.62963	*RUNX1*	21:36276047-36438799	NA	Deletion of Exon 1-2	Thrombocytopenia	23	M
GRIN2A	16:9857831 G>C	NM_000833.4:c.3570C>G	p.His1190Gln	14.52514	*RUNX1*	21:36259327 G>GC	NM_001754.4:c.163dupG	p.Ala55GlyfsTer?	Thrombocytopenia	76	M
IDH2	15:90631934 C>T	NM_002168.3:c.419G>A	p.Arg140Gln	9.67742	*RUNX1*	21:36259139 C>A	NM_001754.4:c.351+1G>T	p.?	Thrombocytopenia	17	M
NOTCH2	19:15276860 GC>G	NM_000435.3:c.5404del	p.Ala1802LeufsTer23	3.2258065	*RUNX1*	21:36171704 G>T	NM_001754.4:c.861C>A	p.Tyr287*	Thrombocytopenia	19	F
PHF6	X:133527608 CTG>C	NM_032458.3:c.321_322del	p.Ala108IlefsTer3	78.84615	*RUNX1*	21:36259139 C>A	NM_001754.4:c.351+1G>T	p.?	Thrombocytopenia	17	M
PTPN11	12:112926887 G>A	NM_002834.4:c.1507G>A	p.Gly503Arg	25.182013	*RUNX1*	21:36164871 T>TGGCGC	NM_001754.4:c.999_1003dupGCGCC	p.Gln335ArgfsTer?	Thrombocytopenia	NA	F
TP53	17:7578442 T>C	NM_000546.6:c.488A>G	p.Tyr163Cys	4.0609136	*RUNX1*	21:36259327 G>GC	NM_001754.4:c.163dupG	p.Ala55GlyfsTer?	Thrombocytopenia	76	M

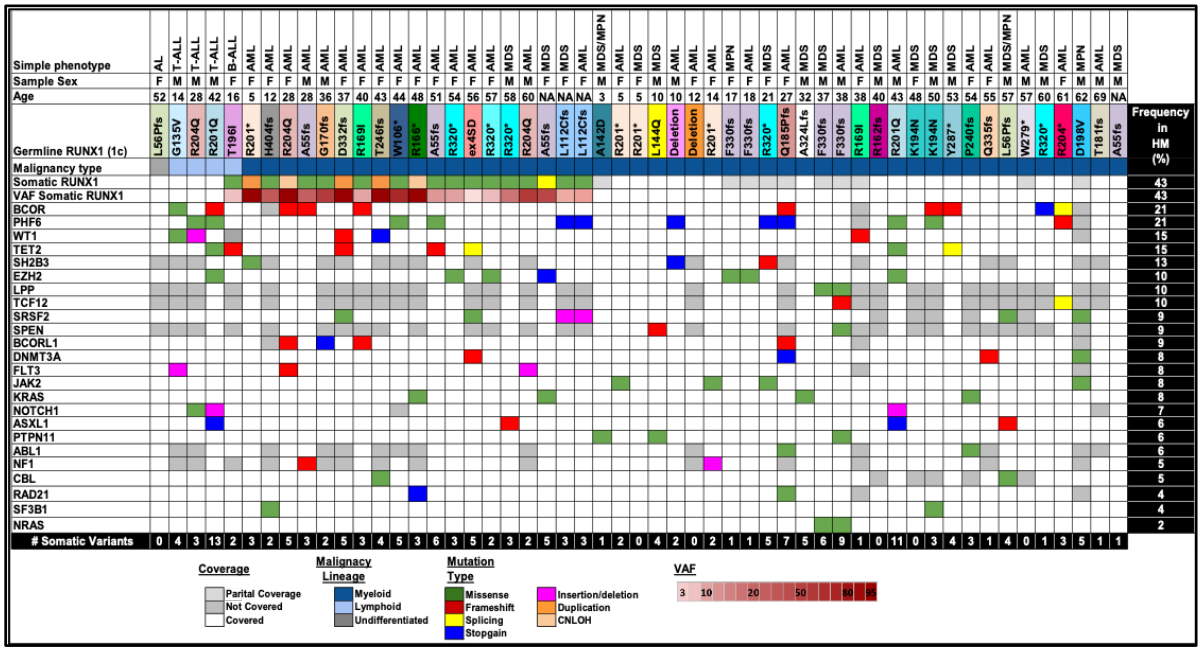
The spectrum of malignancies, germline *RUNX1* mutations, and genes involved in somatic CH mutations are shown in **Figure 2.12**. This figure demonstrates that the genes that are most frequently mutated in blood cancers in germline *RUNX1* mutation carriers are driven by somatic mutations in *BCOR*, *PFH6*, and *RUNX1.* I did not perform single sequencing or allele-specific sequencing, so I was unable to determine if the somatic *RUNX1* mutations occurred in the allele that did not contain the germline *RUNX1* mutation. The VAF of each CH-related somatic mutation, as well as the genes mutated in the CH-related variants, are summarized in the rose plot in **Figure 2.13**.

Figure 2.11. Age of patients with HHM-related malignancies at the time of sample collection.



This analysis demonstrated that the VAF of driver mutations in HMs that develop in germline *RUNX1* mutation carriers can be quite high, frequently in excess of 60%. The distribution of VAFs for driver mutations in malignancies for *DDX41*, *GATA2*, and *RUNX1* germline mutation carriers are shown in **Figure 2.14**. These findings suggest that HHMs occurring in individuals with germline *RUNX1* mutations are inherently more proliferative clones than HHMs in individuals with germline *DDX41* or *GATA2* mutations.

Figure 2.12. The spectrum of malignant phenotypes (top row), germline *RUNX1* mutations (row 4), and genes mutated in *RUNX1* germline mutation carriers with malignancies.



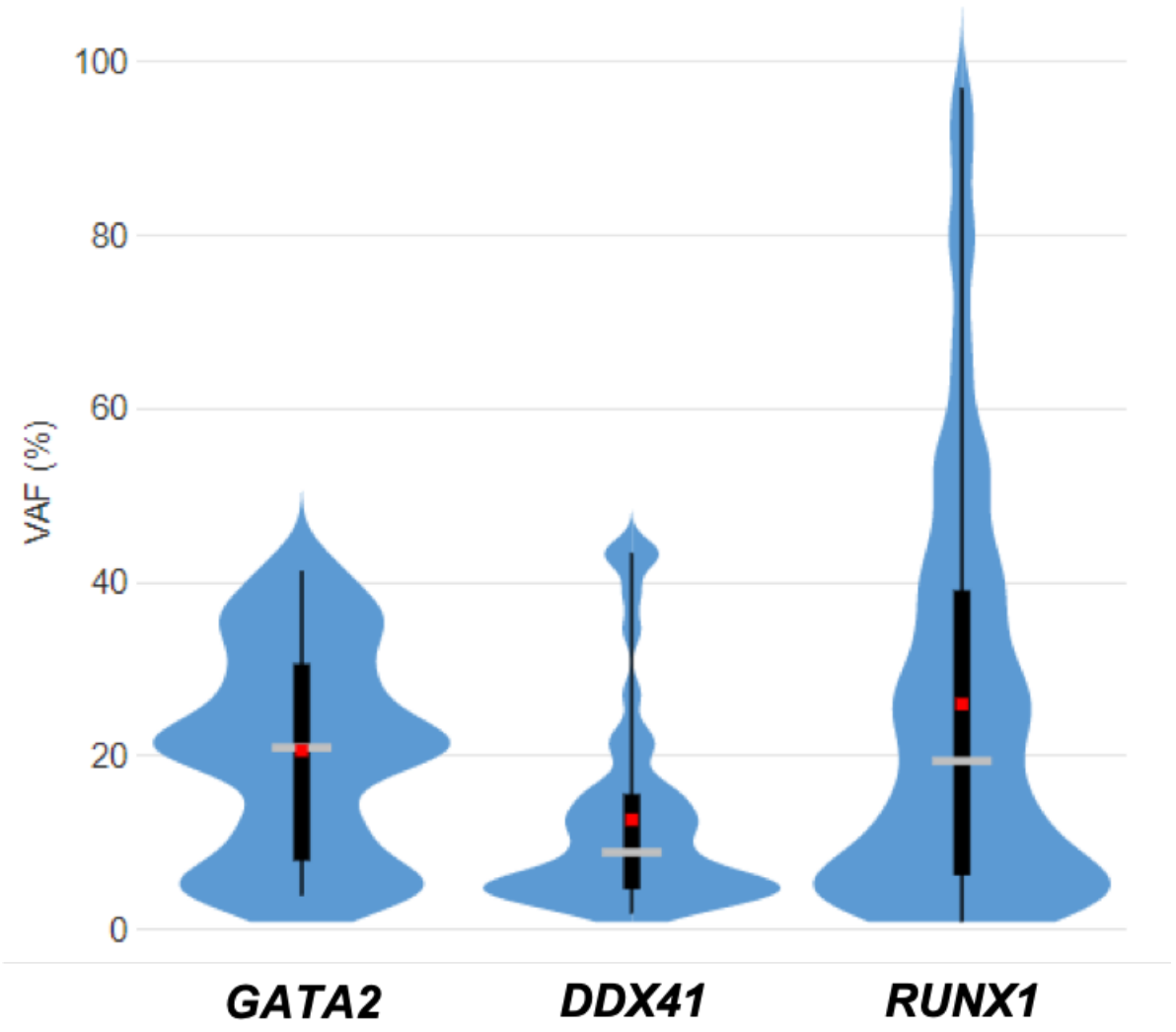
Based on the relative paucity of CH in germline *DDX41* mutation carriers as well as the relatively low VAF in these individuals, I hypothesized that malignancies driven by germline *DDX41* malignancies were more genetically "bland" as compared to those arising in individuals with germline *RUNX1* mutations. To test this, I identified all patients in the *DDX41* and *RUNX1* cohorts for whom I had collected germline tissue (cultured skin fibroblasts) and hematopoietic tissue (peripheral blood and/or bone marrow). I then utilized sequencing data from these paired tumor/normal tissues to calculate the tumor mutation burden per megabase (TMB/Mb) for malignancies arising in individuals with germline *DDX41* or *RUNX1* mutations (**Figure 2.15**). This analysis confirmed that the TMB/Mb of *DDX41*-driven HHMs is significantly lower than in *RUNX1*-driven HHMs.

Figure 2.13. Rose plot summarizing the most frequently mutated genes in *RUNX1* germline mutation carriers with malignancies.

I then compared somatic mutations in individuals with germline *RUNX1* mutations who had developed either CH or HMs so as to determine if a similar set of genes were involved in CH as blood cancers in this cohort (**Figure 2.16**). This analysis demonstrated that *BCOR* mutations were the most common somatic mutations in *RUNX1* mutation carriers with CH, and the proportion of *BCOR* mutations increased in *RUNX1* mutation

carriers who developed frank malignancies. This suggests that *BCOR* somatic mutations are early leukemogenic events that persist in the final leukemic clone.

Figure 2.14. Violin plot with VAFs of somatic mutations in malignancies in *DDX41*, *GATA2*, and *RUNX1* germline mutation carriers. Grey bar: median; red dot: mean.



In contrast, a large number of somatic mutations were observed in *RUNX1* germline mutation carriers who had developed overt hematopoietic cancers (**Figure 2.16**). Second hit mutations in *RUNX1* were the most common somatic mutations in blood cancers arising in *RUNX1* germline mutation carriers, occurring in over 40% of germline *RUNX1* mutation carriers who developed a blood cancer. Intriguingly, somatic *RUNX1*

mutations were observed exclusively in affected individuals and were not detected in any unaffected *RUNX1* mutation carriers with CH. This finding suggests that second hit *RUNX1* mutations are "late" events in individuals with germline *RUNX1* mutations who experience leukemogenesis. Other somatic mutations that were overrepresented in malignancies, but not in *RUNX1* mutation carriers with CH, were detected in *PHF6*, *WT1*, *SH2B3*, *EZH2*, *JAK2*, *LPP*, *NF1*, *PDS5A*, *PTPRD*, *TCF12*, *FLT3*, *KRAS*, *LRP1B*, *SPEN*, *CBL*, *BCORL1*, and *NOTCH1* (**Figure 2.16**).

Figure 2.15. Tumor mutation burden/megabase of malignancies in *DDX41* versus *RUNX1* germline mutation carriers.

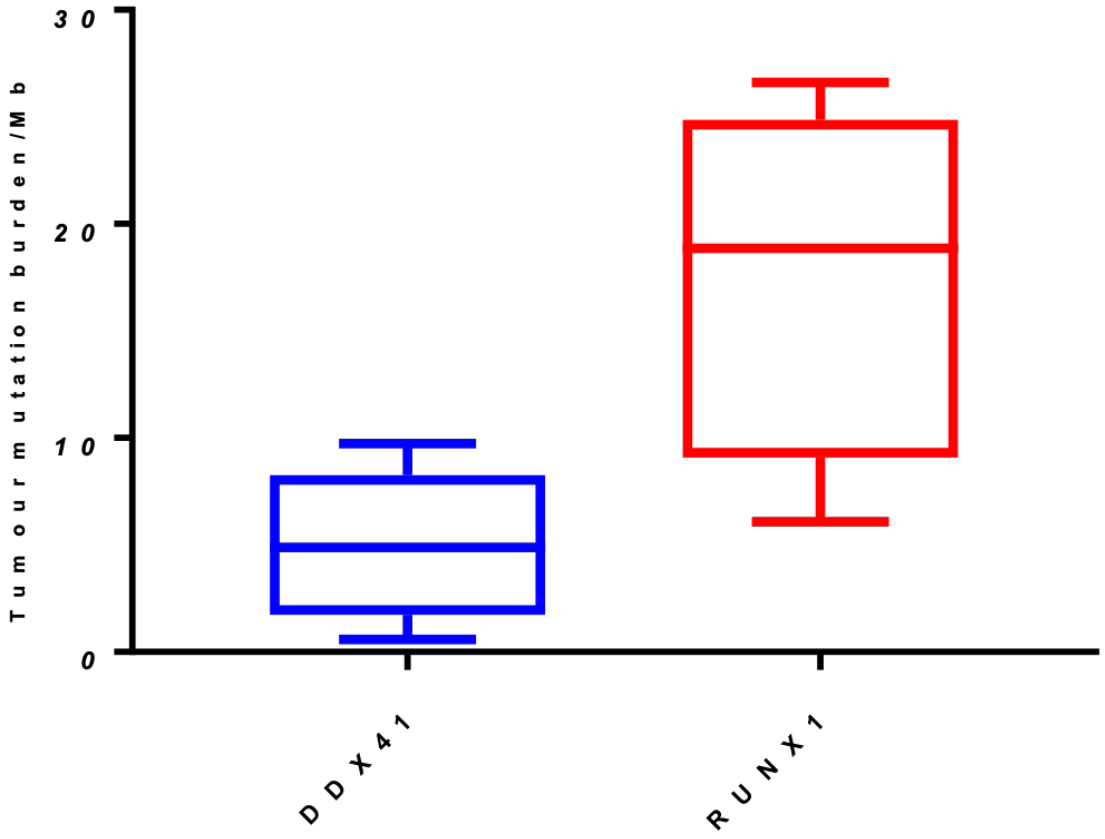
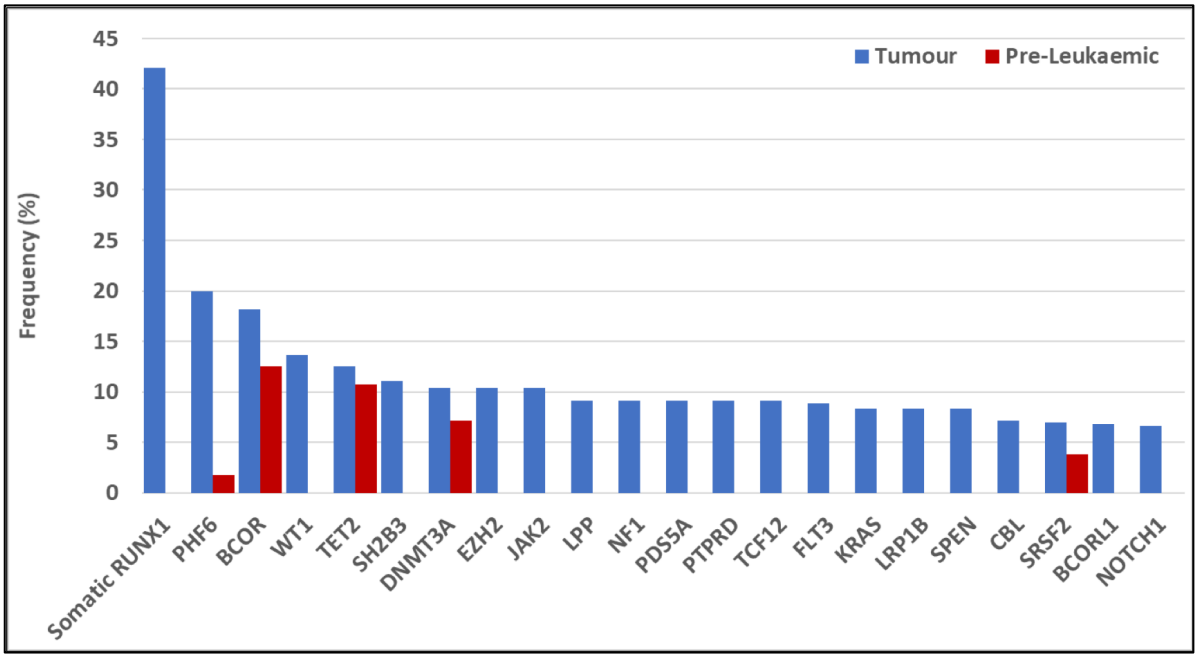


Figure 2.16. Somatic mutations detected in the hematopoietic tissue (peripheral blood, bone marrow) from individuals with germline *RUNX1* mutations who were unaffected ("pre-leukaemic") or affected ("tumour") by blood cancers. Somatic *RUNX1* mutations were the most common mutational event observed in the blood cancers, but were not seen in individuals with clonal hematopoiesis. *BCOR* mutations were the most common mutations observed in individuals with clonal hematopoiesis, and the proportion of individuals with *BCOR* mutations increased in affected individuals.



Discussion

Clonal hematopoiesis in unaffected DDX41, GATA2, or RUNX1 mutation carriers

In summary, I completed the largest natural history study to date focused on the identification of leukemogenic events that occur in individuals with HHMs driven by germline mutations in *DDX41*, *GATA2*, or *RUNX1*. These HHMs represent the most common HHM (*DDX41*), one of the most highly penetrant HHMs (*GATA2*), and the first known HHM for which the largest number of known pedigrees exist worldwide (*RUNX1*). Prior work in the field has focused on small series of patients and have only sequenced individuals with a single HHM (i.e., have sequenced only *RUNX1* carriers but omitted *GATA2* carriers). Because these prior studies utilized different sequencing platforms and

variant calling bioinformatics packages, these disparate studies may have been impacted by sequencing and bioinformatic artifacts that make cross-study comparisons difficult.

To address these limitations in the field, I sequenced both unaffected mutation carriers without malignancies as well as carriers who had been diagnosed with blood cancers. I performed these studies using a shared set of sequencing platforms. I then utilized a uniform bioinformatics pipeline for variant calling and utilized a novel genomics database, the RUNX1 database, to store the data for this project. Ultimately, this approach enabled us to utilize a more uniform approach for my study than what has been performed in the past.

Ultimately, my work demonstrated patterns of CH vary significantly between *RUNX1*, *GATA2*, and *DDX41* mutation carriers. Based on my data, CH is largely not a mechanism of leukemogenesis in *DDX41* mutation carriers, as only 3.9% of *DDX41* carriers develop CH. Furthermore, CH is an age-dependent phenomenon in *DDX41* carriers, as the only patient in the *DDX41* unaffected carrier cohort who developed *DNMT3A*-driven CH was, at 87 years of age, the oldest individual in the unaffected cohort. This pattern of CH closely mirrors the experience in the larger population of non-HHM patients, in which CH is age dependent and dominated by clones in the epigenetic regulators *DNMT3A*, *TET2*, and *ASXL1* (Desai et al., 2018; Genovese et al., 2014; Jaiswal et al., 2014). Therefore, CH is not more common in *DDX41* mutation carriers as compared to population-based controls, CH occurs at “normal” ages of onset in *DDX41* mutation carriers relative to population-based controls, and the sole case of CH involved somatic mutations in *DMT3A*, a gene that is commonly mutated in CH in population-based controls.

An alternative explanation for the findings above is that the prevalence of CH is relatively low in *DDX41* carriers relative to other HHMs because the CH-related variants themselves are not readily identified via short-read sequencing and contemporary variant calling pipelines that have been optimized for short read sequencing. In particular, CNVs at low VAFs would prove difficult to detect using my bioinformatics approach for this project. However, this limitation would also be present for the larger scale population-based studies examining CH and my investigations of CH in carriers of *GATA2* or *RUNX1* germline mutations. Future efforts to investigate leukemogenic events in unaffected *DDX41* carriers could potentially address this limitation by utilizing long-read NGS approaches.

In contrast to *DDX41* mutation carriers, CH is significantly more common in unaffected carriers of *GATA2* or *RUNX1* mutations as compared to population controls. I detected CH in 28.9% of *GATA2* mutation carriers and 35.8% of *RUNX1* mutation carriers, both of which are significantly higher rates compared to age-matched population controls. Both HHMs were notable for a “left-shifted” CH that occurred at significantly younger ages than would be expected relative to population controls. Importantly, prior estimates of CH in younger carriers of germline *RUNX1* mutations showed that approximately 66% of these individuals showed evidence of early-onset CH (Churpek et al., 2015). By expanding the size of my patient cohort significantly, I was able to revise these estimates downward to 27.9% for all *RUNX1* mutation carriers under the age of 50 and to 35.8% for *RUNX1* germline mutation carriers of any age.

CH in *GATA2* or *RUNX1* mutation carriers, however, involved a distinct spectrum of genes for each HHM patient population. *GATA2* carriers were more likely to have CH

driven by somatic mutations in epigenetic regulators (*DNMT3A*, *KMT2C*, or *KDM5A*), and *RUNX1* mutation carriers were more likely to have CH driven by somatic *BCOR*, *TET2*, or *DNMT3A* mutations. Intriguingly, CH in both *GATA2* and *RUNX1* germline mutation carriers was notable for somatic mutations in genes that are not commonly mutated in CH in the general population. Most notably, somatic *BCOR* mutations were the most common CH-related event in *RUNX1* carriers, but *BCOR*-driven CH is uncommon in population-based controls.

Similarly, *KMT2C* and *KDM5A*-mutated clones were identified in *GATA2* carriers with CH, but these genes are infrequently mutated in CH that occurs in population-based controls (Desai et al., 2018; Genovese et al., 2014; Jaiswal et al., 2014). This suggests that CH in HHMs driven by germline *GATA2* or *RUNX1* mutations only partially follows patterns of CH observed in population controls, with CH occurring earlier in these two HHM syndromes and involving a distinct set of genes as compared to population-based controls. This study, therefore, will inform future attempts to model pre-leukemic states in the HHMs.

Leukemogenic mutations in DDX41, GATA2, or RUNX1 germline mutation carriers affected by hematopoietic malignancies

Prior studies have largely focused on the mutational spectrum of single HHMs, but in this study I sequenced diseased tissue (peripheral blood and/or bone marrow at the time of a malignancy diagnosis) from a significantly larger number of individuals with germline *DDX41*, *GATA2*, or *RUNX1* germline mutations who had been affected by HMs. I performed these sequencing studies using an approach that limited the introduction of

technical variability by using the same sequencing devices as I used in my work investigating CH. I also utilized the same bioinformatics pipeline for variant calling as in the CH work. I also utilized the RUNX1 database for all of my data deposition and variant curation so as to maintain a uniform method of data analysis when compared to my CH work described above.

Ultimately, I demonstrated that HHMs affecting both *DDX41* and *RUNX1* carriers were notable for second hit events that drove frank leukemogenesis. This second hit phenomenon was not observed in *GATA2* mutation carriers. Intriguingly, only *RUNX1* carriers experienced CH, whereas *DDX41* mutation carriers largely did not experience CH. This suggests that the second hit events in *DDX41* mutation carriers, the most common of which was the p.Arg525His somatic mutation, are likely sufficient to drive leukemogenesis in the absence of other somatic mutations. In contrast, *RUNX1* mutation carriers most frequently developed malignancies with multiple somatic mutations, suggesting that *RUNX1* somatic mutations are late mutational events that are not sufficient to drive frank leukemogenesis in *RUNX1* germline mutation carriers. In keeping with this finding, I calculated the mutation rate per megabase for malignancies in *DDX41* or *RUNX1* mutation carriers and demonstrated that *DDX41*-driven malignancies are more genetically bland than are *RUNX1*-driven malignancies.

These findings support my hypothesis that distinct leukemogenic mechanisms exist between the HHMs themselves as well as between the HHMs and the general population. Intriguingly, I demonstrated that somatic *TP53* mutations were exceedingly uncommon among malignancies that developed in *DDX41*, *GATA2*, or *RUNX1* mutation carriers, despite *TP53* being mutated in 10.2% of sporadic hematopoietic and lymphoid

malignancies in the COSMIC tumor database (Tate et al., 2019). Data from Niimi *et al.* previously demonstrated that somatic *RUNX1* mutations in MDS are mutually exclusive with somatic *TP53* mutations (Niimi et al., 2006). However, mine is the first study to demonstrate that germline mutations in *RUNX1* appear to be similarly exclusive of somatic *TP53* mutations in malignancies that develop in individuals with germline *RUNX1* mutations.

I also demonstrated that the median age of onset of malignancy diagnosis varies between the HHMs, with the age of diagnosis being significantly younger in affected individuals with germline *GATA2* or *RUNX1* mutations as compared to those with germline *DDX41* mutations.

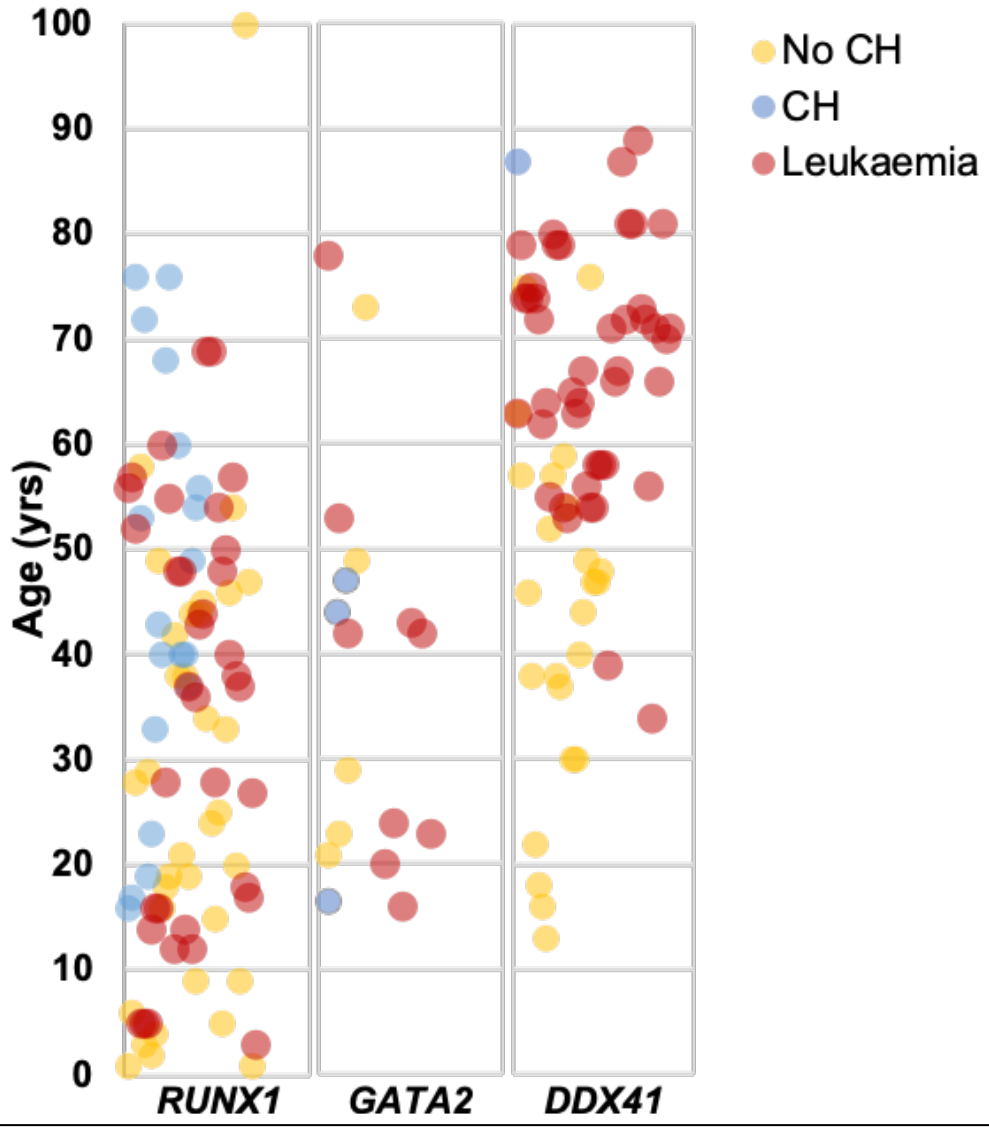
I also showed that the clonal dynamics within malignancies appear to differ based on HHM of interest. For example, while the VAFs of somatic mutations in malignancies driven by germline mutations in *DDX41* or *RUNX1* both have tails skewing toward higher VAFs, the magnitude of VAFs in *RUNX1*-driven HHMs is higher, with some VAFs in excess of 80%. In contrast, the mean VAF of somatic mutations in malignancies driven by germline *DDX41* mutations is lower, less than 20%. This finding suggests the underlying germline genetic drivers of HHMs continue to impact the mutational profiles and clonal dynamics of frank malignancies that develop after many decades in affected individuals.

These analyses ultimately demonstrated distinct patterns of both CH and frank leukemogenesis among these HHMs (**Figure 2.17**), with *DDX41* mutation carriers being largely devoid of CH, developing overt blood cancers later in life, and being diagnosed with malignancies that were largely more genetically “bland” in terms of both the number

of genes mutated as well as the VAFs of somatic mutations that were detected. In contrast, *GATA2* germline mutation carriers were significantly more likely to develop CH than both population-based controls and *DDX41* mutation carriers. The CH-related clones in *GATA2* mutation carriers were largely driven by mutations in epigenetic regulators.

Finally, *RUNX1* germline mutation carriers were the most likely patient cohort to have evidence of CH, with somatic *BCOR* mutations being the most common CH-related event in *RUNX1* mutation carriers. In contrast, *BCOR* mutations were uncommon in the other HHMs and population-based controls. These *BCOR* mutations, however, did not appear sufficient to cause frank leukemogenesis in *RUNX1* germline mutation carriers, as the majority of malignancies in *RUNX1* carriers had multiple somatic mutations detected. The most common of these somatic mutations occurred as second hits in *RUNX1*. Intriguingly, a *RUNX1* HHM phenocopy driven by germline mutations in *ETV6* is also remarkable for second hits in *ETV6* among individuals who develop malignancies (Karastaneva et al., 2020; Topka et al., 2015). The VAFs of somatic variants in these malignancies were significantly higher than in HHMs driven by germline *DDX41* mutations in particular, which suggests that *RUNX1* germline mutations may be associated with a more proliferative malignancy as compared to malignancies driven initially by a germline *DDX41* mutation.

Figure 2.17. Pattern of CH and leukemias between individuals with germline mutations in *DDX41, GATA2*, or *RUNX1.*



Chapter III

Clonal hematopoiesis is uncommon in the *RUNX1* phenocopies driven by *ANKRD26* or *ETV6* germline mutations

The data in this chapter are adapted from the following manuscript, which is being prepared for submission: Drazer MW*, Homan CC*, Yu K*, McNeely KE, Pozsgai MJ, Feng J, King-Smith SL, Kim E, Lawrence DM, Schreiber AW, Hahn CN, Scott HS, Sood R, Brown AL, Liu P, Godley LA. Clonal hematopoiesis is uncommon in the *RUNX1* phenocopies driven by *ANKR26* and *ETV6* germline mutations. 2021. *co-first authors

In this study, I designed and performed the experiments, analyzed the data, and wrote the manuscript.

Additional data in this chapter are adapted from the following publication in *Haematologica*: Homan CC, King-Smith SL, Lawrence DM, Arts P, Feng J, Andrews J, Armstrong M, Ha T, Dobbins J, **Drazer MW**, Yu K, Bodor C, Cantor A, Cazzola M, Degelman E, Dinardo C, Duployez N, Favier R, Fröhling S, Fitzgibbon J, Klco J, Krämer A, Kurokawa M, Lee J, Malcovati L, Morgan N, Natsoulis G, Owen C, Patel KP, Preudhomme C, Raslova H, Rienhoff H, Ripperger T, Schulte R, Tawana K, Velloso E, Yuan B, Liu P, Godley LA, Schreiber AW, Hahn CN, Scott HS, Brown AL. The RUNX1 database: a RUNX1 genomics database for mutation data aggregation, analysis, and sharing. *Haematologica*. Online ahead of print. July 2021. PMID: 34233450.

I have also included modified figures from the following review for which I served as a co-author:

Feurstein S, **Drazer MW**, Godley LA. Germline predisposition to hematopoietic malignancies. *Human Molecular Genetics*. 2021.

Introduction

Currently, there are at least a dozen known HHMs. A number of general phenotypic categories exist among the currently recognized HHMs, with some HHMs phenocopying others. For example, three of the HHMs share a striking phenotype of lifelong thrombocytopenia (low platelet count), bruising, bleeding, and a dramatically increased lifetime risk of leukemogenesis (Roloff GW, 2021.). This phenotype, referred to as the HT/HHM phenotype, stems from germline mutations in *ANKRD26*, *ETV6*, or *RUNX1*. Individuals with germline mutations in these HT/HHM-associated genes experience a variable lifetime risk for HMs that ranges from approximately 8% in *ANKRD26* germline mutation carriers up to 33% in *ETV6* mutation carriers and 44% in *RUNX1* mutation carriers (Feurstein et al., 2016) (Table 3.1).

Table 3.1. Germline *ANKRD26*, *ETV6*, or *RUNX* mutation carriers experience variable lifetime risk for malignancy as well as variable median age of leukemogenesis. MDS: myelodysplastic syndrome; ALL: acute lymphoblastic leukemia; AML: acute myeloid leukemia.

Gene	Malignancy	Lifetime Risk	Median Age of Onset
ANKRD26	**MDS/AML**	**8%**	**30-70 years**
ETV6	**ALL, MDS/AML**	**33%**	**32 years**
RUNX1	**MDS/AML**	**44%**	**33 years**

Prior work from the Godley and Graubert groups demonstrated that 66% of germline *RUNX1* carriers showed evidence of CH prior to the age of 50, a dramatically elevated risk as compared to the general population (Churpek et al., 2015; Genovese et al., 2014; Jaiswal et al., 2014). My work, described in Chapter II of this thesis, investigated the risk of CH in a much larger cohort of germline *RUNX1* mutation carriers. My work has

now revised the prior estimate of 66% CH risk down to 27.9% for all *RUNX1* mutation carriers under the age of 50 and to 35.8% for *RUNX1* germline mutation carriers of any age.

Although germline mutations in *ANKRD26* or *ETV6* phenocopy *RUNX1* germline mutations, no studies to date have focused on the risk of CH in individuals with germline mutations in *ANKRD26* or *ETV6*. This knowledge gap has persisted since the initial studies describing CH in *RUNX1* mutation carriers was published in 2015 (Churpek et al., 2015). Prior work in the field that focused on germline *ANKRD26* mutations, for example, has largely defined the hematologic and bleeding phenotypes of individuals with *ANKRD26* germline mutations or the somatic mutations that develop in these individuals after they are diagnosed with HMs. Although some of the unaffected and affected patients in these studies have had their peripheral blood and/or bone marrow sequenced, the panels used for these analyses have been limited in size (ranging from 5 to 37 genes). Prior work has also utilized heterogeneous sequencing platforms and variant calling packages, and the number of patients studied have been small (5 unaffected and 3 affected, to date) (Kewan et al., 2020; Perez Botero et al., 2016; Tsang et al., 2017).

Similarly, a number of studies have examined cohorts of patients with germline *ETV6* mutations who developed malignancies so as to determine the cooperating mutations that may be present in these individuals. These studies have demonstrated that germline *ETV6* mutations result in a global disruption of the transcriptomic program, whereas somatic *ETV6* mutations that acquire in individuals with sporadic HMs do not cause a similar effect (Nishii et al., 2021). These studies have also shown that second hit mutations in *ETV6* in individuals with germline *ETV6* mutations are more prevalent than

would be expected for sporadic leukemia (Karastaneva et al., 2020; Topka et al., 2015). However, no concerted efforts to genomically profile otherwise unaffected *ETV6* germline mutation carriers prior to the development of HMs have been performed to date. This knowledge gap has limited the field's ability to understand the role of CH in leukemogenesis in unaffected germline *ETV6* mutation carriers and has restricted our ability to develop high-fidelity, patient-derived models of high-risk pre-leukemic states in germline *ETV6* mutation carriers.

To address these knowledge gaps in the HT/HHMs and their pre-leukemic states, I performed the first systematic natural history study of unaffected individuals with either germline *ANKRD26* or germline *ETV6* mutations. My findings investigating pre-leukemic states in *RUNX1* (described in Chapter II of this thesis), were notable for the presence of "early" and "late" somatic mutational events during leukemogenesis. My findings in *RUNX1* were notable for a large number of somatic *BCOR* mutations, which were "early" mutational events that were enriched in the germline *RUNX1* mutation carrier population relative to population controls. My findings were also notable for a large number of second hits in *RUNX1*, which served as "late" mutational events.

The goals of this study, therefore, were to determine: 1) Do *ANKRD26* or *ETV6* germline mutations lead to increased rates of CH among otherwise unaffected individuals, as is observed in their phenocopy driven by germline *RUNX1* mutations? 2) Do *ANKRD26* or *ETV6* germline mutations follow a pattern of "early" and "late" mutational events, as is observed in individuals with germline mutations in *RUNX1*? 3) Are *BCOR* somatic mutations over-represented in CH among *ANKRD26* or *ETV6* germline mutation carriers, as I have demonstrated in *RUNX1* germline mutation carriers?

Given that the penetrance of HMs is lower in *ANKRD26* (8%) and *ETV6* (33%) germline mutation carriers than in *RUNX1* mutation carriers (44%), I hypothesized that unaffected carriers of germline *ANKRD26* or *ETV6* mutations would be notable for an intermediate rate of CH relative to unaffected germline *RUNX1* mutation carriers (Feurstein et al., 2016). I also hypothesized that CH in the *ANKRD26* or *ETV6* mutation carriers would be heavily *BCOR*-driven, as is observed in *RUNX1* HT/HHMs.

These hypotheses were driven by biologic rationale, particularly in regard to germline *ANKRD26* mutations. *RUNX1* encodes for a TF (RUNX1) that binds to the 5' UTR of *ANKRD26* and suppresses *ANKRD26* expression **(Figure 3.1**). *ANKRD26* 5' UTR germline mutations interfere with normal RUNX1 function by disrupting RUNX1 binding. Using an *in vitro* culture system, Bluteau *et al.* demonstrated that *ANKRD26* is expressed in CD34+ hematopoietic stem and progenitor cells (HSPCs) during normal hematopoiesis. However, *ANKRD26* expression then diminishes during normal megakaryopoiesis and is nearly undetectable by day 12 of *in vitro* megakaryopoiesis, when mature megakaryocytes are generated. Germline mutations in the 5' UTR of *ANKRD26*, however, disrupt the interaction with RUNX1. This disruption leads to the persistent expression of *ANKRD26* throughout megakaryopoiesis and ultimately disrupts proplatelet formation, leading to the diminished production of platelets in the peripheral circulation. This mechanism accounts for the thrombocytopenia that is observed in individuals with germline mutations in *RUNX1* or the 5' UTR of *ANKRD26* (Bluteau et al., 2014). A schematic of the 5' UTR of *ANKRD26* is shown in **Figure 3.1**. This figure also details the location of the *ANKRD26* germline mutations that were present in the unaffected individuals enrolled on this natural history study.

Figure 3.1. The 5' UTR region of *ANKRD26* contains a binding site for both RUNX1 and FLI1. Germline mutations in the 5' UTR of *ANKRD26* disrupt this interaction and lead to a hereditary thrombocytopenia/hereditary hematopoietic malignancy (HT/HHM) phenotype. This phenotype phenocopies germline *RUNX1* mutations. Germline *FLI1* mutations lead to hereditary thrombocytopenia but are not known to cause HHMs and are therefore not considered to represent an HT/HHM phenocopy. The mutations shown (c.-128G>A, c.-119C>G, and c.-118C>T) represent the mutations in the cohort of *ANKRD26* germline mutation patients included for Chapter III of this thesis. ClinVar classifies the c.-118C>T and c.-119C>G variants as likely pathogenic and the c.-128G>A variant as pathogenic.

Intriguingly, all of the known LP/P germline mutations in *ANKRD26* submitted to the ClinVar database with supporting evidence are confined to the 5' UTR region of *ANKRD26*. This mutational clustering further supports the role of this region in the dysregulated expression of *ANKRD26* in HT/HHM pathogenesis in *ANKRD26* and *RUNX1* germline mutation carriers. Furthermore, haploinsufficiency of *ANKRD26* has been shown to not be associated with the HT/HHM phenotype, or any clinical phenotype in general, in humans.

Perhaps the most well-known example of an individual who is haploinsufficient for *ANKRD26* is James Watson, who carries a 31.5 Kb deletion in *ANKRD26* that deletes the final five exons of the gene. However, Dr. Watson is not known to have thrombocytopenia (Wheeler et al., 2008). Furthermore, partial inactivation of *Ankrd26* in murine models is associated with only a metabolic phenotype characterized by obesity and insulin resistance instead of a hematopoietic phenotype (Bera et al., 2008).

Bluteau *et al.* also demonstrated that mutations in the 5' UTR of *ANKRD26* lead to overactivation of the mitogen activated protein kinase/extracellular signal-regulated kinase (MAPK/ERK) pathways in megakaryocytes. ERK overactivation leads to dysfunctional megakaryopoiesis and platelet production, which accounts for the hereditary thrombocytopenia that is observed in individuals with germline *ANKRD26* or *RUNX1* mutations (Bluteau et al., 2014).

However, dysregulated activation of the MAPK/ERK pathways may also represent one of the underlying mechanisms driving leukemogenesis in the HT/HHM phenocopies. Dysregulated MAPK/ERK signaling has been implicated in myeloid leukemogenesis previously (Al-Kali et al., 2013). Bluteau *et al.* hypothesized that the MAPK/ERK pathways may be overactivated even in HSPCs in patients with germline *ANKRD26* mutations, which would implicate this mechanism in the increased leukemic risk that is observed in both germline *ANKRD26* and *RUNX1* carriers (Bluteau et al., 2014). For example, the RAS proteins (KRAS and NRAS) sit upstream of ERK in the MAPK/ERK pathways, and activating *KRAS* and *NRAS* mutations are present in 4% of MDS and 15% of AML cases. Even more intriguingly, *RAS* mutations are associated with higher blast counts at the time of blood cancer presentation (Al-Kali et al., 2013). This finding is consistent with my observation in Chapter II of this thesis that individuals with germline *RUNX1* mutations are more likely to present with both higher VAF driver mutations as well as more proliferative blast counts when compared to individuals with HHMs driven by germline mutations in *DDX41* or *GATA2*. Despite this finding and the hypothesis of Bluteau *et al.*, the constitutive activation of MAPK/ERK in *ANKR26*-mutated or *RUNX1*-mutated HSPCs has not been investigated to date. This knowledge gap has largely stemmed from the

difficulty of obtaining primary patient samples of HSPCs, which would require repeated bone marrow biopsies. However, I will address this knowledge gap later in this thesis (Chapter IV).

The hypothesis that dysregulated MAPK/ERK signaling may drive leukemogenesis in the HT/HHM phenocopies is further reinforced by the interaction of MAPK/ERK signaling with the ETS class of TFs, including ETV6. ETV6 has an established role in the MAPK/ERK pathways, where it is an established phosphorylation target of ERK2, JNK2, and, in particular, P38α (Selvaraj et al., 2015). The ETV6 protein contains a number of binding motifs for the MAPK/ERK family of proteins (**Figure 3.2**), and it is very feasible that germline mutations in *ETV6* would potentially disrupt these motifs and/or alter the protein structure of ETV6 such that these bindings sites may be disrupted. Ultimately, these disruptions could lead to disrupted MAPK/ERK signaling, which may ultimately contribute to leukemogenesis in a similar manner as is observed in individuals with germline *ANKRD26* or *RUNX1* mutations. Therefore, I hypothesized that the leukemogenic steps in *ETV6* germline mutation carriers would closely mirror those observed in *ANKRD26* or *RUNX1* germline mutation carriers.

Figure 3.2. The ETV6 protein contains a number of motifs recognized by the mitogen activated protein kinase/extracellular signal-regulated kinase (MAPK/ERK) proteins, including D domains (red), DEF domains (green), LXL domains (blue), and D/LXL domains (red, underlined).

```
         10         20         30         40         50
MSETPAQCSI KQERISYTPP ESPVPSYASS TPLHVPVPRA LRMEEDSIRL
         60         70         80         90        100
PAHLRLQPIY WSRDDVAQWL KWAENEFSLR PIDSNTFEMN GKALLLLTKE
        110        120        130        140        150
DFRYRSPHSG DVLYELLQHI LKQRKPRILF SPFFHPGNSI HTQPEVILHQ
        160        170        180        190        200
NHEEDNCVQR TPRPSVDNVH HNPPTIELLH RSRSPITTNH RPSPDPEQRP
        210        220        230        240        250
LRSPLDNMIR RLSPAERAQG PRPHQENNHQ ESYPLSVSPM ENNHCPASSE
        260        270        280        290        300
SHPKPSSPRQ ESTRVIQLMP SPIMHPLILN PRHSVDFKQS RLSEDGLHRE
        310        320        330        340        350
GKPINLSHRE DLAYMNHIMV SVSPPEEHAM PIGRIADCRL LWDYVYQLLS
        360        370        380        390        400
DSRYENFIRW EDKESKIFRI VDPNGLARLW GNHKNRTNMT YEKMSRALRH
        410        420        430        440        450
YYKLNIIRKE PGQRLLFRFM KTPDEIMSGR TDRLEHLESQ ELDEQIYQED

                                                      EC
```

D domains: red
LXL domains: blue
DEF domains: green
D/LXL domain: red, underlined

Methods

Study participants

Patients with germline *ANKRD26* 5' UTR mutations were enrolled at either the University of Chicago or the National Institutes of Health. All patients with germline *ETV6* mutations were enrolled at the University of Chicago. All patients participated on Institutional Review Board (IRB)-approved protocols that allowed for the sharing of anonymized patient sequencing data between the institutions. Each individual on the

study provided informed consent. Parents of minors provided consent for their enrollment, and minors provided assent when appropriate. In total, 12 individuals were enrolled from four families. Among the *ANKRD26* mutation carriers, there were seven unaffected individuals and one affected individual. There were four unaffected *ETV6* mutation carriers. The age of the patients ranged from 8 to 63 years. The single affected *ANKRD26* mutation carrier had been diagnosed with AML, with 23% blasts present in their peripheral blood at the time of sample collection (**Table 3.2**).

Each HT/HHM-related germline variant was reviewed by multiple team members and classified as pathogenic/likely pathogenic/VUS/likely benign/benign using criteria from the Association for Molecular Pathology and the American College of Human Genetics and Genomics (Richards et al., 2015). The *ANKRD26* 5' UTR variants were classified as likely pathogenic (c.-118C>T; c.-119C>G) or pathogenic (c.-128G>A), in keeping with their ClinVar classifications. The *ETV6* p.R369Q was classified as pathogenic, also in keeping with its ClinVar classification.

Table 3.2. Cohort of patients enrolled on the *ANKRD26* and *ETV6* natural history study.

Germline Gene of Interest	Variants	Individual	Family	Phenotype	Age(s) at Sample Collection (years)
ANKRD26	5' UTR c.-118C>T NM_014915.2	UCMC_40.1	Family_0040	Thrombocytopenia	56, 60
ANKRD26	5' UTR c.-118C>T NM_014915.2	UCMC_40.2	Family_0040	Thrombocytopenia	55
ANKRD26	5' UTR c.-118C>T NM_014915.2	UCMC_40.3	Family_0040	Thrombocytopenia	25
ANKRD26	5' UTR c.-128G>A NM_014915.2	UCMC_B2.23	Family_00B2	Thrombocytopenia	22
ANKRD26	5' UTR c.-128G>A NM_014915.2	UCMC_B2.24	Family_00B2	AML (23% blasts)	48
ANKRD26	5' UTR c.-128G>A NM_014915.2	UCMC_B2.25	Family_00B2	Thrombocytopenia	8
ANKRD26	5' UTR c.-119C>G NM_014915.2	UCMC_311.4	Family_0311	Thrombocytopenia	43, 44
ANKRD26	5' UTR c.-119C>G NM_014915.2	UCMC_311.8	Family_0311	Thrombocytopenia	13
ETV6	p.R369Q NM_001987.4	UCMC_75.5	Family_0075	Thrombocytopenia	59, 62, 63
ETV6	p.R369Q NM_001987.4	UCMC_75.6	Family_0075	Thrombocytopenia	36, 38
ETV6	p.R369Q NM_001987.4	UCMC_75.7	Family_0075	Thrombocytopenia	33
ETV6	p.R369Q NM_001987.4	UCMC_75.41	Family_0075	Thrombocytopenia	62

Tissue acquisition

All patient samples consisted of germline tissue (*e.g.*, cultured skin fibroblasts) or hematopoietic tissue (peripheral blood, bone marrow, or saliva). Serial samples were

collected from the research subjects when possible so as to compare clonal dynamics over time. Genomic DNA (gDNA) was extracted with the QIAamp DNA Blood Mini Kit (Qiagen) using the manufacturer's instructions. DNA concentrations were measured via a Nanodrop (Thermo Scientific) and/or Qubit fluorometer (Life Technologies).

Panel sequencing/whole exome sequencing

At least 100 ng of genomic DNA from each sample was sheared, selected by size, ligated to adapters, and standard sequencing libraries were generated via PCR amplification. Panel-based sequencing was performed using the Oncoplus assay at the University of Chicago as previously described (Kadri et al., 2017). In brief, following library generation, genomic capture was performed using a custom SeqCap EZ capture panel that covered 1212 genes (Roche), and an additional PCR amplification with real-time quantitative PCR quantification was performed. An Illumina HiSeq was used to sequence the pooled capture libraries.

Sequencing data were stored on a protected high-performance computing system at the University of Chicago that exceeds requirements for the Health Insurance Portability and Accountability Act. The data were initially analyzed via a bioinformatics pipeline that melded publicly available packages built off of the GATK package and a custom bioinformatics pipeline developed at the University of Chicago. These data were initially reviewed by MWD and KY for driver mutations.

Following an initial round of review, the raw FASTQ files were then transferred to the University of South Australia in order to analyze the data for *ANKRD26* and *ETV6* mutation carriers using the freebayes-based RUNX1db bioinformatics pipeline (Homan

et al., 2021). Following mapping via BWA-MEM and de-duplication via sambamba, I performed base recalibration using GATK BQSR and generated frequency-based calls using freebayes. I then uploaded the ensuing VCFs to the RUNX1db for review. Additionally, I compared VCFs generated via GATK and freebayes in order to verify the presence of known germline and somatic variants in each sample and to confirm that each data file matched the research subject of interest. This workflow (**Figure 2.1**) has been published previously (Homan et al., 2021).

I used an identical set of filters for read quality and depth as was described in Chapter II. My thresholds were as follows: variant allelic depth >= 5, read depth >= 20, population prevalence (variants at 0.1% or higher in any population database were removed), pathogenicity (missense variants that were not predicted to be damaging in 2 or more *in silico* predictors were removed; CADD scores with values of 20 or higher were removed), and oncogenicity (variants not in genes with known roles as drivers in myeloid malignancies, not in the Catalogue of Somatic Mutations in Cancer [COSMIC], or *RUNX1* variants were removed). I analyzed the subsequent list of candidate variants and used IGV to review each variant of interest manually in the individual BAM files. I removed any variants labeled as artifacts after the aforementioned steps.

I then analyzed the remaining, IGV-confirmed variants to label each variant as germline or somatic in origin. For individuals with sequencing data from cultured skin fibroblasts, I compared variants identified in hematopoietic tissue directly to data obtained from skin fibroblasts. Samples without paired germline tissue were analyzed using a combination of population allelic frequency (minor allele threshold of 0.01% or lower), VAF (with likely germline VAFs considered to be between 30 and 60% for genes on

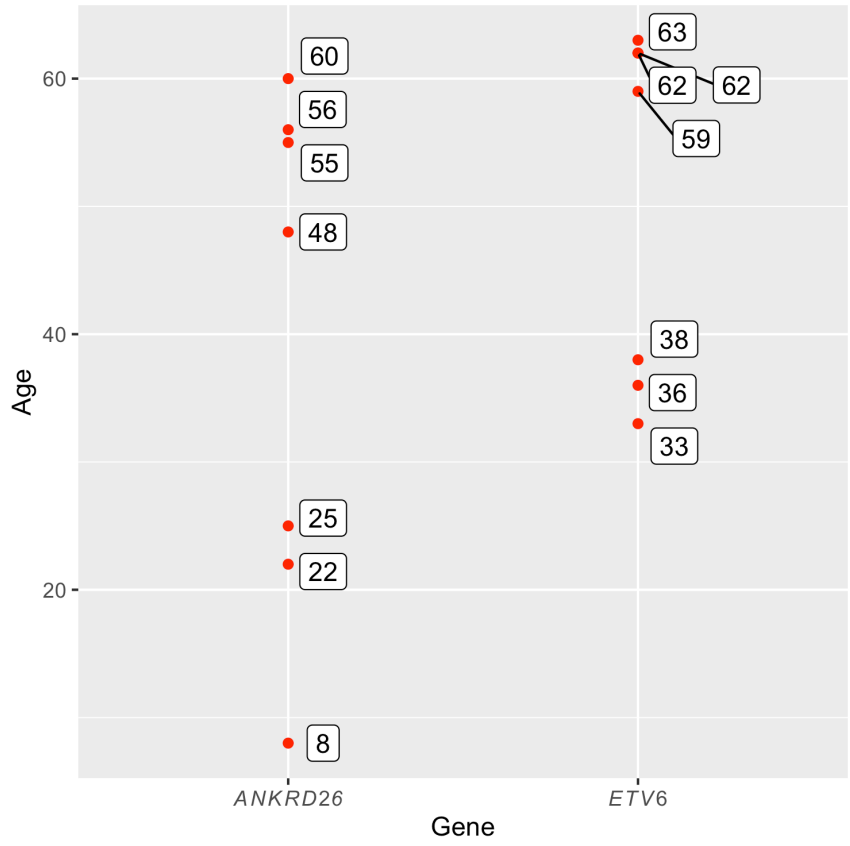
autosomal chromosomes and 80% or higher for genes on the X chromosome), and the frequency of the variant in question in tumor databases such as COSMIC. Any variant passing the above population filters, but which still occurred more than twice in COSMIC, was considered not to be of germline origin.

This filtering process produced a list of variants of likely somatic or somatic origin which I reviewed manually for clinical and biological relevance. The determination of "likely somatic" or "somatic" origin adhered to criteria defined in the original RUNX1 database manuscript (Homan et al., 2021). "Clinically relevant" variants were known pathogenic germline variants in leukemia or variants that were present more than two times in COSMIC in hematopoietic and lymphoid samples (H&L samples). Novel driver variants were clinically relevant if they were present in a gene known to be recurrently mutated in COSMIC H&L samples, were a truncating variant (nonsense, frameshift indels, essential splice site variants), were in the same domain as known pathogenic variants (for example, the RUNT domain in RUNX1), or were a deletion in a gene where deletion is a known mechanism of disease. Missense variants were considered to be clinically relevant if they were damaging in at least 3 *in silico* algorithms and were highly conserved via GERP and Phylop scores.

All somatic and likely somatic variants that did not meet criteria for clinical relevance were categorized as "possibly relevant" or "of unknown relevance". Possibly relevant variants were recurrent in COSMIC (one or more samples), novel recurrent CNVs, or variants in genes not associated with HMs previously, but which were loss-of-function variants in conserved domains, mutational hot spots, were recurrent in COSMIC, or present in multiple affected individuals. All "possibly relevant" missense variants also

needed to be predicted damaging on 3 or more *in silico* algorithms and highly conserved using the GERP and Phylop scores. Somatic variants of “unknown relevance” were any variants that passed the filtering approach described above and passed IGV review, but did not meet criteria for “clinically relevant” or “possibly relevant” variants.

Figure 3.3. Age of individuals at the time of sample collection in the *ANKRD26* and *ETV6* natural history study.



Results

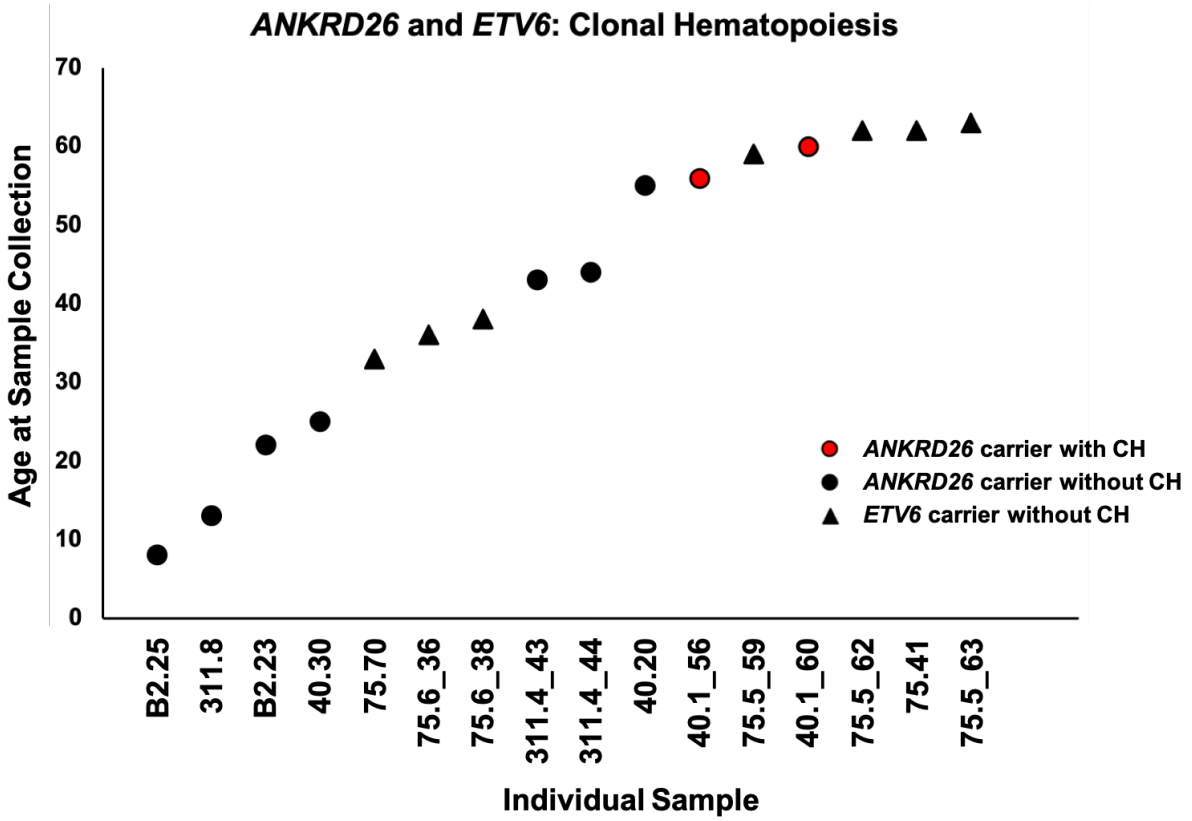
The age of individuals with germline *ANKRD26* or *ETV6* mutations at the age of sample collection is shown in **Figure 3.3**. The median age at time of sample collection was 43 years for unaffected *ANKRD26* germline mutation carriers and 59 years for unaffected *ETV6* germline mutation carriers. I collected serial samples from two individuals with germline *ETV6* mutations and from one individual with a germline

ANKRD26 mutation.

Among the seven unaffected individuals with germline *ANKRD26* mutations who provided samples, the overall prevalence of CH was 14.3%. I detected the presence of CH in one individual with a germline *ANKRD26* mutation (5' UTR, c.-118C>T, NM_014915.2). This clone was driven by a somatic mutation in *SF3B1* (p.Lys700Glu) that was present at a VAF of 9.4% at age 56 and at a VAF of 17.4% at age 60. Mutations in *SF3B1* are a well-established mechanism of leukemogenesis, and 7.7% of COSMIC hematopoietic and lymphoid (H&L) malignancies carry LP/P somatic *SF3B1* mutations. The *SF3B1* p.Lys700Glu mutation is a recognized somatic hotspot that is observed in 2.1% (n = 525) of COSMIC H&L cases (n = 25028). This variant has a pathogenic FATHMM score of 0.99 (Tate et al., 2019). No germline *ETV6* mutation carriers (n=4) had detectable evidence of CH at the limits of detection of the NGS assay.

Unlike individuals with germline mutations in *RUNX1*, no individuals under the age of 50 with germline mutations in *ANKRD26* or *ETV6* had evidence of detectable CH despite nearly half of the unaffected samples being collected from individuals in this age group (**Figure 3.4**). As was the case for *DDX41* germline mutation carriers (described in Chapter II), the only patient in the *ANKRD26* germline mutation cohort with evidence of CH was also the oldest individual in the cohort. This individual had CH present at ages 56 and 60, both of which represent ages in the general population at which the prevalence of CH begins to increase (Genovese et al., 2014; Jaiswal et al., 2014). No individuals in the germline *ETV6* mutation carrier cohort had detectable CH even though more than half of the samples from this cohort were collected from individuals over the age of 50.

Figure 3.4. Clonal hematopoiesis in *ANKRD26* or *ETV6* germline mutation carriers. One individual with a germline *ANKRD26* mutation (5' UTR, c.-118C>T, NM_014915.2) had a CH clone driven by *SF3B1* p.Lys700Glu in two samples collected at the ages of 56 and 60.



The only patient with a germline *ANKRD26* mutation and an HM in this cohort developed AML with 23% blasts at the age of 48. I included this patient's peripheral blood samples in my genomic analyses. Intriguingly, this patient's leukemic clone carried a large variety of detectable driver mutations: *CUX1* (p. Phe472GlnfsX105, VAF 92.4%), *RUNX1* (p.Arg320X, VAF 55.8%), *TET2* (p.Phe1309LeufsX54, VAF 40.0%), *FLT3* (p.Asp835His, second tyrosine kinase domain (TKD), VAF 18.3%), *MUC4* (p.Ser1669ArgfsX25, VAF 17.1%), and *SAMD9* (p.Val798GlyfsX7, VAF 13.3%) (**Table 3.3**).

In total, the observation time for the *ANKRD26* cohort was 275 years. The incidence of CH was 4.5×10^{-3} CH cases/observation year (4.5 CH cases per 1000

observation years) among *ANKRD26* germline mutation carriers. The incidence rate of HMs was $3.6x10^{-3}$ malignancies/observation year (3.6 malignancies per 1000 observation years) among *ANKRD26* germline mutation carriers. The observation time for the *ETV6* cohort was 258 years, with no diagnoses of an HM.

Table 3.3. Germline mutations and somatic driver mutations identified in each individual in *ANKRD26* and *ETV6*-mutated HT/HHM phenocopy cohort.

Unaffected Individuals						
Germline Gene	**Variant**	**Individual Sample ID**	**Age**	**CH present?**	**CH mutation**	**VAF**
ANKRD26	5' UTR c.-128G>A NM_014915.2	B2.25	8	N	NA	NA
ANKRD26	5' UTR c.-128G>A NM_014915.2	B2.23	22	N	NA	NA
ANKRD26	5' UTR c.-118C>T NM_014915.2	40.30	25	N	NA	NA
ANKRD26	5' UTR c.-118C>T NM_014915.2	40.20	55	N	NA	NA
ANKRD26	5' UTR c.-118C>T NM_014915.2	40.1_56	56	Y	*SF3B1* p.Lys700Glu	9.4%
ANKRD26	5' UTR c.-118C>T NM_014915.2	40.1_60	60	Y	*SF3B1* p.Lys700Glu	17.4%
ANKRD26	5' UTR c.-119C>G NM_014915.2	311.4_43	43	N	NA	NA
ANKRD26	5' UTR c.-119C>G NM_014915.2	311.4_44	44	N	NA	NA
ANKRD26	5' UTR c.-119C>G NM_014915.2	311.8_13	13	N	NA	NA
ETV6	p.R369Q NM_001987.4	75.70	33	N	NA	NA
ETV6	p.R369Q NM_001987.4	75.6_36	36	N	NA	NA
ETV6	p.R369Q NM_001987.4	75.6_38	38	N	NA	NA
ETV6	p.R369Q NM_001987.4	75.5_59	59	N	NA	NA
ETV6	p.R369Q NM_001987.4	75.5_62	62	N	NA	NA
ETV6	p.R369Q NM_001987.4	75.41	62	N	NA	NA
ETV6	p.R369Q NM_001987.4	75.5_63	63	N	NA	NA
Affected Individuals						
Germline Gene	**Variant**	**Individual Sample ID**	**Age**	**Malignancy**	**Somatic Driver Mutations**	**VAF**
ANKRD26	5' UTR c.-128G>A NM_014915.2	B2.24	48	AML (23% blasts)	*CUX1* p. Phe472GlnfsX105	92.4%
					RUNX1 p.Arg320X	55.8%
					TET2 p.Phe1309LeufsX54	40.0%
					FLT3 p.Asp835His	18.3%
					MUC4 p.Ser1669ArgfsX25	17.1%
					SAMD9 p.Val798GlyfsX7	13.3%

The observation time for the *ANKRD26* and *ETV6* HT/HHM phenocopy combined cohorts was 533 years. The incidence of CH was 2.1 x 10^{-3} CH cases/observation year (2.1 CH cases per 1000 observation years) in this cohort. With one diagnosis of an HM, the incidence rate for malignancy was $1.9x10^{-3}$ malignancies/observation year (1.9 malignancies per 1000 observation years).

Discussion

Among the known HHMs, three are phenocopies that share a phenotype of lifelong thrombocytopenia (low platelet count), bruising, bleeding, and an increased lifetime risk of leukemogenesis (Roloff GW, 2021.). Among these phenocopies, the majority of investigations have focused on *RUNX1*-driven HHMs. This bias has likely occurred for two reasons. First, *RUNX1*-driven HHMs were the first HHM identified, with the seminal publication from Song *et al*. published in 1999 (Song et al., 1999). This publication occurred 12 years before the first publication linking *ANKRD26* germline mutations to the HT/HHM phenotype and 16 years before the first publications linking *ETV6* germline mutations to the HT/HHM phenotype (Noetzli et al., 2015; Noris et al., 2011; Topka et al., 2015; Zhang et al., 2015). Second, *RUNX1*-driven HHMs have the highest penetrance for HMs among the HT/HHM phenocopies, with approximately 44% of mutation carriers developing blood cancers during the course of their lives. This penetrance is higher than the lifetime risk for blood cancers in *ETV6* germline mutation carriers (33%) and *ANKRD26* germline mutation carriers (8%) (Table 3.1). As is often the case, the most "severe" hereditary syndromes are the first to be recognized clinically, with more subtle syndromes and lower penetrance phenotypes being recognized later.

Therefore, it is not surprising that prior work in the HHM field to date that has focused on high risk preleukemic states in the HT/HHM phenocopies has been limited to the much larger cohorts of germline *RUNX1* carriers. The original estimate for CH prevalence in germline *RUNX1* carriers was 66%, with the vast majority of these individuals developing early-onset CH before the age of 50 (Churpek et al., 2015). However, in Chapter II of this dissertation, I expanded the sample size of this cohort by

an order of magnitude. I also expanded the cohort to include *DDX41* mutation carriers (the most common HHM) and *GATA2* mutation carriers (one of the most highly penetrant HHMs) in addition to germline *RUNX1* mutation carriers. This analysis allowed us to reduce the estimated CH prevalence in *RUNX1* mutation carriers (35.8% overall) and the estimated prevalence of CH among *RUNX1* mutation carriers under the age of 50 (27.9%). I also demonstrated that CH among *RUNX1* mutation carriers is overwhelmingly driven by somatic mutations in *BCOR*, a unique mutational pattern that appears to be largely limited to *RUNX1* mutation carriers, but which is rare in the general population.

Prior to my most recent work, it was unknown if germline *ANKRD26* or *ETV6* carriers shared a similar predisposition toward early-onset CH driven by somatic *BCOR* mutations as “early” events and second hits in *ANKRD26* or *ETV6* as “late” events.

To address this knowledge gap, I have performed the first study focused on examining pre-leukemic states in the remaining two HT/HHM phenocopies driven by germline *ANKRD26* or *ETV6* mutations. This represents a natural extension of my work in Chapter II. I hypothesized that both of these HT/HHM syndromes would be notable for a higher prevalence of CH relative to the general population. I also hypothesized that CH in these populations would be predominantly driven by *BCOR* mutations, as is observed in germline *RUNX1* mutations. Ultimately, both of these hypotheses were incorrect. CH was not observed in any of the *ETV6* mutation carriers in my study, and CH was only observed in the oldest *ANKRD26* germline mutation carrier in my cohort. This CH was driven by a somatic mutation in the spliceosome-related gene *SF3B1*, and the VAF of this clone increased over time as the individual aged. To date, no individuals in the unaffected

pre-leukemic cohort have been diagnosed with an HM, including the patient diagnosed with CH.

It is possible, however, that the unexpected findings above stemmed from the limited number of germline variants and families who were eligible for study. This patient cohort was small and drawn from only four families, with three different *ANKRD26* variants and one *ETV6* variant represented. These variants may not be representative of the most prevalent leukemogenic phenomena that exist in HT/HHMs driven by *ANKRD26* or *ETV6* germline variants. Ultimately, larger numbers of unaffected and affected germline *ANKRD26* or *ETV6* mutation carriers will be needed to better determine the pre-leukemic genetic milieu that exists in these syndromes.

I also profiled a 48-year-old patient with a germline *ANKRD26* mutation who had been diagnosed with AML. This analysis demonstrated that the patient had a number of driver mutations identified with higher-than-expected VAFs. This finding is similar to that of germline *RUNX1* mutation carriers who developed blood cancers in Chapter II of this thesis. This finding suggests that the clonal dynamics of blood cancers in *ANKRD26* or *RUNX1* germline mutation carriers may be similar, which possibly reflects the close interaction between *ANKRD26* expression and the RUNX1 TF, which regulates *ANKRD26* expression via interactions in the *ANKRD26* 5' UTR (Bluteau et al., 2014). Additionally, the spectrum of driver mutations identified in the patient with a germline *ANKRD26* mutation and AML was notable for both typical and atypical drivers. For example, *FLT3* mutations are a common somatic driver that are detected in about 19% of H&L malignancies in the COSMIC database (Tate et al., 2019). The specific *FLT3* mutation in this patient's leukemia clone, a p.Asp835His missense mutation that is located

in the second TKD domain of FLT3, is also a well-described driver mutation that is present in 7% of *de novo* AML cases (Thiede et al., 2002). Similarly, the patient had a driver mutation identified in *TET2*, and *TET2* mutations are observed in 12.7% of H&L malignancies in COSMIC. However, the particular p.Phe1309LeufsX54 mutation in this clone has not been described in COSMIC. In contrast, *RUNX1* is only mutated in 5.8%, *MUC4* in 4.61%, *CUX1* in 1.44%, and *SAMD9* in 0.55% of COSMIC's H&L malignancies, yet each of these genes were mutated in the AML sample from my patient (Tate et al., 2019). The particular driver mutations observed in the *ANKRD26* carrier with AML were also rare, with the *RUNX1* p.Arg320X mutation observed in 0.004% (1/23337) of COSMIC H&L cases, and the *MUC4*, *CUX1*, and *SAMD9* driver mutations were not previously described in COSMIC.

The somatic *RUNX1* mutation detected in my patient with AML was particularly intriguing, as it was present in the majority of the tumor sample (VAF 55.8%) and is rare in COSMIC H&L cases. This *RUNX1* mutation may effectively represent a second hit that serves as a late leukemogenic event in *ANKRD26* germline mutation carriers, particularly given the role of RUNX1 in regulating *ANKRD26* expression. This leukemogenic process would mirror the leukemogenic mechanisms observed in germline *RUNX1* mutation carriers, who are prone to developing second hits in *RUNX1* during leukemogenesis.

In conclusion, this analysis represents the first natural history study focused on identifying leukemogenic mechanisms in individuals with HT/HHM phenocopies driven by germline mutations in *ANKRD26* or *ETV6*. I identified *SF3B1*-driven CH in 14.3% of germline *ANKRD26* mutation carriers but did not detect CH in any *ETV6* germline mutation carriers. This rate of CH in *ANKRD26* mutation carriers is higher than that

observed in population controls, but it does not approach the high rates of CH that are observed in germline *RUNX1* mutation carriers (35.8%), many of whom are under the age of 50. Individuals with germline *RUNX1* mutations, which phenocopies *ANKRD26* or *ETV6* mutations, develop *BCOR*-driven, early onset CH before developing HMs most commonly driven by second hits in *RUNX1*. I did not observe a similar pattern of recurrent early or late events in the *ANKRD26* or *ETV6* mutation carrier cohorts. I did, however, identify a somatic *RUNX1* mutation in a leukemia clone that developed in a germline *ANKRD26* mutation carrier. This *RUNX1* mutation is rare in patients with AML and may effectively represent a second hit event given the role of RUNX1 in regulating *ANKRD26* expression. Given the relatively small sample size of my cohort as well as the limited number of pedigrees with germline *ANKRD26* mutations worldwide, it is unclear if CH clones driven by *SF3B1* mutations are a recurring mechanism of disease in germline *ANKRD26* mutation carriers. Future studies focused on the HT/HHM phenocopies driven by *ANKRD26* or *ETV6* germline mutations will benefit from the collection of samples from additional families, which may be expedited by expanding upon the bioinformatics framework established during the creation of the RUNX1 database, which facilitates the collection and curation of genomics data from the HHM syndromes (Homan et al., 2021).

Chapter IV

Germline *ANKRD26* mutations are associated with constitutive activation of the MAPK/ERK signaling pathway in patient-derived induced pluripotent stem cells and induced pluripotent stem cell-derived hematopoietic stem and progenitor cells.

In this study, I designed and performed the experiments, analyzed the data, and will write the manuscript.

The data shown in this chapter also includes data adapted from the following manuscript: Borst S, Nations CC, Klein JG, Pavani G, Maguire JA, Camire RM, **Drazer MW**, Godley LA, French DL, Poncz M, and Gadue P. (2021) Study of inherited thrombocytopenia resulting from mutations in *ETV6* or *RUNX1* using a human pluripotent stem cell model. *Stem Cell Reports*. 16(6):1458-67. PMID: 34019812.

In this latter study, I prepared patient samples for iPSC generation and assisted with editing/revising the manuscript.

Introduction

One limitation of studying HT/HHMs is that large numbers of viable patient samples are difficult to obtain and therefore are available in limited numbers. Additionally, particular cellular populations of interest, such as hematopoietic stem and progenitor cells (HSPCs), megakaryocytes, and undifferentiated stem cells, may be difficult to isolate in quantities suitable for translational research. Many patients with HT/HHM syndromes live at substantial distances from our university and other academic centers, which limits my ability to collect serial viable samples.

Furthermore, traditional model systems, in particular murine model systems, do not fully recapitulate the phenotype observed in many patients with HT/HHMs. For example, Hock *et al.* generated a conditional *Etv6* knockout murine model in which the DNA-binding portion of the protein was excised. Unlike humans, in which the HT/HHM phenotype is associated with an autosomal dominant inheritance pattern, heterozygous conditional knockout (KO) mice have normal peripheral blood counts, including platelets, throughout their lifespans. In contrast, homozygous conditional KO animals have a 50% reduction in platelet counts, with an increased number of megakaryocytes in the bone marrow (Hock et al., 2004). Attempts to utilize *Runx1* haploinsufficient mice to model the HT/HHM phenotype have encountered similar challenges, as mice that are heterozygous for *Runx1* mutations display only a very mild decrease in platelet counts (Sun & Downing, 2004). A subsequent study from Matheny *et al.* showed that mice heterozygous for *Runx1* mutations seen in human patients did not develop thrombocytopenia (Matheny et al., 2007). No studies investigating the effects of *Ankrd26* 5' UTR mutations in a murine model have been published to date. To date, no murine model has effectively recapitulated the

leukemogenic phenotype of germline *ANKRD26*, *ETV6*, or *RUNX1* heterozygous mutations in humans.

Ultimately, these experiences demonstrate that murine model systems may require a more severe genetic insult, such as complete deficiency of a gene in question, in order to recapitulate the hematopoietic phenotypes of the heterozygous HT/HHM mutations that are observed in humans. This inability to effectively model HT/HHMs using murine models has ultimately slowed progress in the HT/HHM field, particularly in regard to the development of novel therapeutic approaches for this class of syndromes.

Therefore, in order to address this gap in the field, I developed a series of patient-derived iPSC models from patients with HT/HHM-related germline mutations as well as other hereditary cancer syndromes. This panel of iPSCs is detailed in **Table 4.1**. This panel contains iPSCs from a patient with a germline mutation in each of the HT/HHM phenocopy-related genes (*ANKRD26*, *ETV6*, or *RUNX1*).

Once the iPSCs were established, I focused my efforts on studying the effects of *ANKRD26* germline mutations on hematopoiesis for three reasons. First, although *RUNX1* germline mutations were the first HT/HHM phenocopy to be described and are known to have the highest leukemic penetrance of the HT/HHM syndromes, recent studies have demonstrated that germline *ANKRD26* mutations are the most common of the HT/HHM phenocopies. In contrast, germline mutations in *RUNX1* and *ETV6* are the second and third most common HT/HHMs, respectively (Downes et al., 2019). This finding is in keeping with the original manuscript from Noris *et al.*, which demonstrated that germline *ANKRD26* mutations were responsible for 10% of all cases of autosomal dominant hereditary thrombocytopenia in their patient series (Noris et al., 2011). Despite

this prevalence, *ANKRD26* has been under-investigated relative to *ETV6* and *RUNX1*. For example, as of August 2021, only 94 manuscripts related to *ANKRD26* have been published via PubMed, as compared to 1,491 *ETV6*-related publications and 4,358 *RUNX1*-related publications. Furthermore, no patient-derived iPSC lines with germline *ANKRD26* mutations have been characterized to date.

Table 4.1. Panel of induced pluripotent stem cells generated from patients with HT/HHM related germline mutations and/or other germline mutations related to hereditary cancer syndromes.

Name	Background	Allele 1	Allele 2
ANKRD26$^{+/mut}$	Patient	WT	c.-118C>T (5' UTR) (pathogenic)
ETV6$^{+/mut}$	Patient	WT	R264C (VUS)
ETV6$^{+/mut}$	Patient	WT	R369Q (pathogenic)
RUNX1$^{+/mut}$	Patient	WT	S388X (hypermorphic variant) (pathogenic)
ALL_discovery	Patient	WT	*GATA3, ETV4, PTK2B* (*VUSs*)
RUNX2$^{+/mut}$	Patient	WT	*RUNX2* p.Glu72GLnfs*102 (pathogenic)
TUBB1$^{+/mut}$	Patient	WT	p.Gln134X (pathogenic)
TUBB1$^{+/mut}$	Patient	WT	p.L130V (pathogenic)
BRCA1$^{+/mut}$	Patient	WT	Exon 23 deletion (pathogenic)
CHEK2$^{+/mut}$	Patient	WT	p.I157T (pathogenic)
TTLL6$^{+/mut}$	Patient	WT	p.Glu488X (pathogenic, in development)

Second, despite the relatively common prevalence of germline *ANKRD26* mutations in cohorts of patients with a hereditary thrombocytopenia phenotype, very few patient derived models have been developed from *ANKRD26* germline mutation carriers. In contrast, I was part of the effort to develop patient-derived iPSC models from individuals with germline *ETV6* or *RUNX1* mutations. This work led to the first systematic comparison of these *ETV6* and *RUNX1*-driven syndromes using the iPSC model system (Borst et al., 2021).

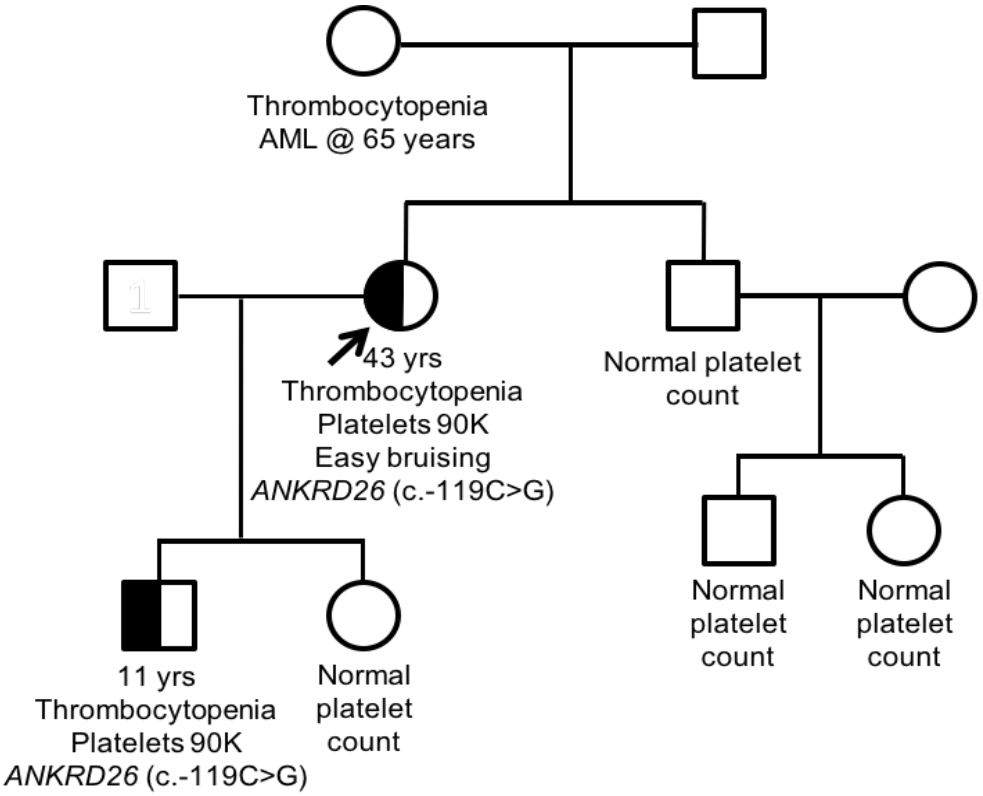
Finally, the link between germline *ANKRD26* mutations and hereditary leukemogenesis remains poorly understood. Bluteau *et al*. previously demonstrated that MAPK/ERK signaling is hyperactivated in the megakaryocytes and platelets of individuals with germline *ANKRD26* mutations, but this hyperactivation can only explain the thrombocytopenia observed in these individuals. MAPK/ERK hyperactivation is a leukemogenic mechanism, and Bluteau *et al*. hypothesized that MAPK/ERK hyperactivation, if present in stem cells and/or hematopoietic stem and progenitor cells (HSPCs), could provide the link between *ANKRD26* germline mutations and the full HT/HHM phenotype (Bluteau et al., 2014). However, no study to date has investigated if MAPK/ERK hyperactivation is present in HSPCs from patients with germline *ANKRD26* mutations.

I also hypothesized that MAPK/ERK hyperactivation is present in individuals with germline *ANKRD26* mutations. In order to test this hypothesis, I generated an iPSC line from a patient with a germline mutation in the 5' UTR of *ANKRD26* (c.-119C>G). I then performed quality control assays on this iPSC line to ensure it behaved as a *bona fide* stem cell line. I then differentiated this iPSC into iPSC-derived HSPCs using a protocol developed by my collaborators, Drs. Paul Gadue and Deborah French. I performed bulk RNA sequencing during this HSPC differentiation so as to determine if significant differences existed between the *ANKRD26*-mutated iPSCs and iPSC-derived HSPCs as compared to iPSCs and iPSC-HSPCs.

Methods

Study participants

Figure 4.1. Pedigree of an individual with a germline *ANKRD26* 5' UTR mutation from whom induced pluripotent stem cells (iPSCs) were generated.



I identified a patient with a germline *ANKRD26* 5' UTR mutation (c.-119C>G) who had enrolled on an IRB-approved research protocol at the University of Chicago. This protocol allows for the optional creation of permanent, immortalized cell lines, including iPSCs. The research subject volunteered to allow iPSCs to be generated from their donated tissue. The pedigree of the patient's family is shown in **Figure 4.1**. Their pedigree was notable for thrombocytopenia (approximately 90 K/uL) in both the proband and her son. The proband's mother also had a personal history notable for lifelong thrombocytopenia and had developed AML at the age of 65 years.

iPSC generation and maintenance

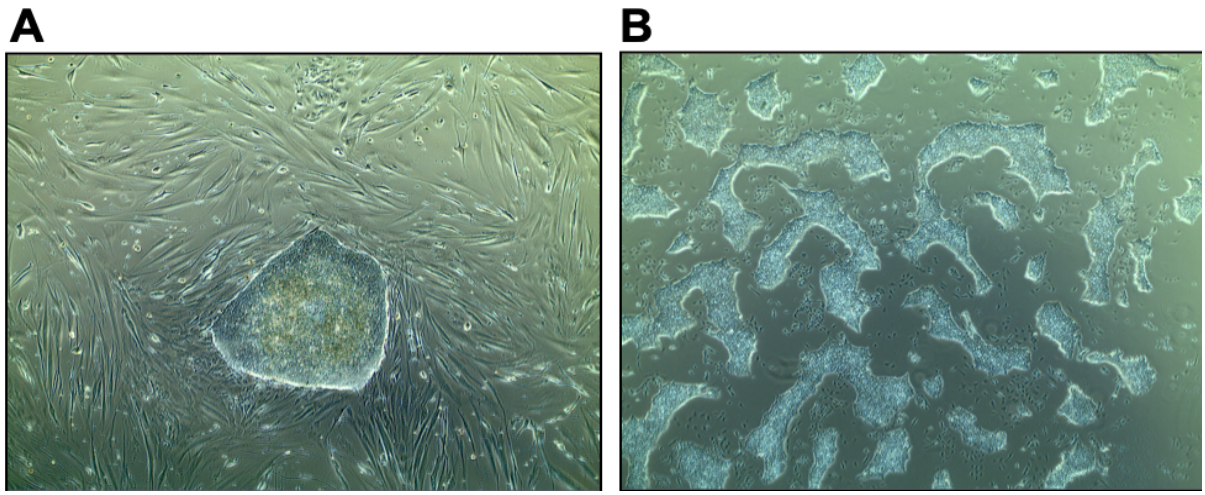
I collected cultured skin fibroblasts from the proband in the pedigree shown in

Figure 4.1. I cultured these fibroblasts in 5% CO_2/atmospheric O_2 using Amniomax complete media with added growth supplement and 1% Pen/Strep (Thermo Fisher) until they became 70% confluent, and I then split the fibroblasts using 0.05% trypsin. The fibroblasts were passaged seven times prior to transfection with Yamanaka factors. I then utilized a protocol described previously by the Gilad group to generate iPSCs (Burrows et al., 2016). I removed a trypsinization step from this protocol in order to improve the yield of colonies post-transfection.

I then transfected the skin fibroblasts with episomal plasmids that contained *KLF4*, *L-MYC*, *LIN28*, *OCT3/4*, *SOX2*, and an shRNA against *p53* (Addgene plasmids 27077, 27078, 27080, and 27082). I then maintained the colonies on a murine embryonic fibroblast (MEF)-based feeder system for approximately 30 days until individual colonies appeared. I manually selected 6-12 colonies and transferred each colony to one well of a six well plate coated with MEFs. I then maintained these colonies in hESC media (DMEM/F12, Corning) supplemented with 20% KOSR (LifeTechnologies), 0.1 mM non-essential amino acids (NEAA), 2 mM GlutaMax, 1% Pen/Strep, 0.1% 2-mercaptoethanol (LifeTechnologies), and 100 ng/mL human basic fibroblast growth factor. I then manually cut and passaged each colony for at least 10 passages on MEFs. After 10 passages of growth on MEFs I transitioned each colony to feeder free conditions on 0.01 mg/cm2 hESC-grade Matrigel (BD Sciences) and mTeSR media (StemCell) supplemented with 1% Pen/Strep. I maintained the iPSCs in feeder-free conditions for at least five passages before performing quality control analyses. Passaging in feeder-free conditions was performed using DPBS containing 0.5 mM EDTA. The gross morphology of both the feeder-based and feeder-free colonies was typical for iPSCs in both growth conditions

(**Figure 4.2**).

Figure 4.2. Microphotographs showing the typical morphology of iPSC colonies in a murine embryonic fibroblast (MEF)-based feeder system (panel A) and in a feeder free-based system (panel B).



iPSC characterization

I then characterized one clone from the iPSC line after it had been passaged at least five times in feeder-free conditions so as to ensure that this line behaved as *bona fide* stem cells. First, the line was karyotyped in the Department of Pathology at the University of Chicago and was found to carry a normal 46, XX karyotype (**Figure 4.3**). Next, I confirmed the pluripotency of the line using two separate methods. First, I generated embryoid bodies (EBs) using a technique described previously in the Gilad group (Burrows et al., 2016). I then stained the ensuing EBs for each germ layer (endoderm, mesoderm, and ectoderm) via immunohistochemistry. The *ANKRD26*$^{+/mut}$ EBs showed appropriate expression of alpha fetoprotein (AFP, 1:200, SC-130302, Santa Cruz Biotech) and FOXA2 (1:200, SC-6554, Santa Cruz Biotech, endoderm), smooth muscle actin (SMA, 1:1500, CBL171, Millipore; mesoderm), and MAP2 (1:200, sc-20172 and sc-74420, Santa Cruz Biotech; ectoderm). In addition to this, I also performed a

directed differentiation using the Human Pluripotent Stem Cell Functional Identification Kit (R&D, SC027B) which successfully reproduced the results of the EB-focused pluripotency assay (**Figure 4.4**).

Figure 4.3. Normal 46, XX karyotype of iPSCs generated from cultured skin fibroblasts collected from a female patient with a germline *ANKRD26* 5' UTR mutation.

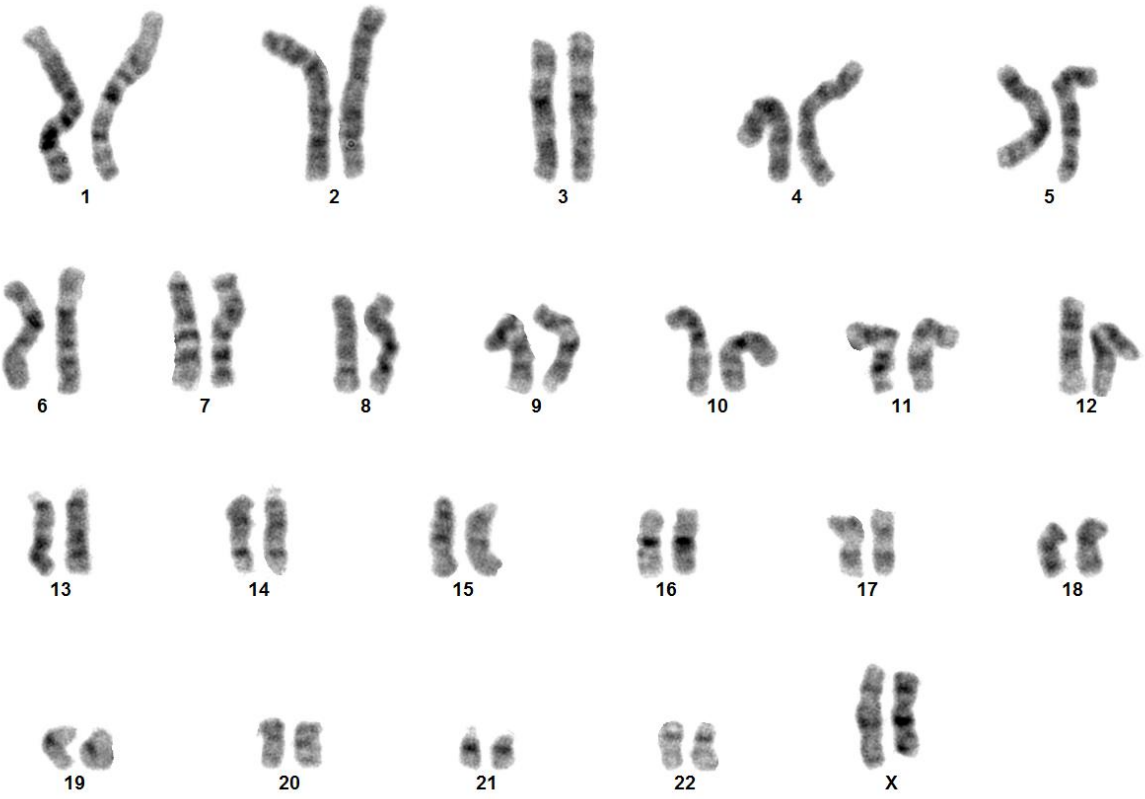
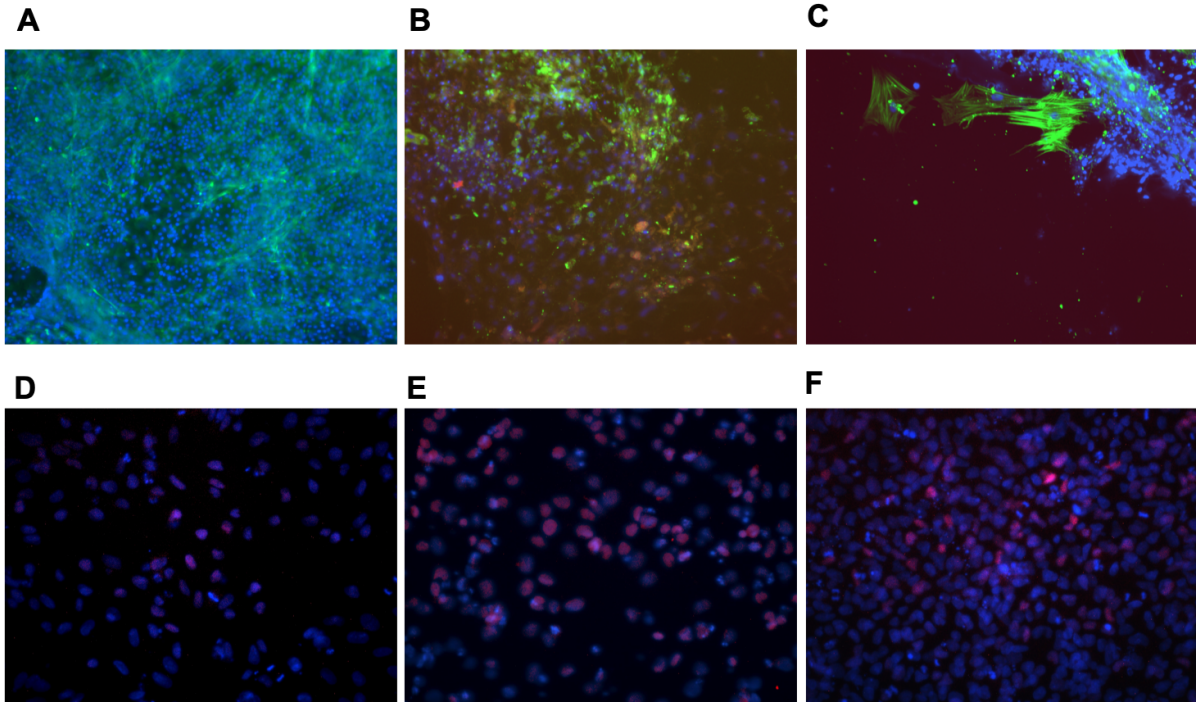


Figure 4.4. Pluripotency assays performed using the *ANKRD26*-mutated iPSC line. The iPSC line was capable of forming each germ cell layer: ectoderm (panels A, D), endoderm (B, E), and mesoderm (C, F) using both an established pluripotency assay with embryoid bodies (panels A-C) and a commercially available pluripotency kit (panels D-F). Blue: DAPI nuclei counterstain. Panel D was stained with goat anti-human Otx2; Panel E was stained with goat anti-human SOX17; Panel F was stained with goat anti-human brachyury. All cells in panels D-F were then stained with the Northern Lights 557-conjugated donkey anti-goat IgG secondary antibody (R&D, #NL001).



Next, I performed a series of experiments to determine if genomic integration and/or retention of the episomal plasmids had occurred in the feeder free iPSCs. I passaged the iPSCs beyond passage 15 in feeder free conditions, then extracted DNA from the feeder free cells using the DNeasy Blood/Tissue Kit (Qiagen). I then performed a series of PCR assays using at least 100 ng of genomic DNA and a series of primers directed at amplifying *EBNA-1* or *PXCLE*, both of which were present in the backbone of all reprogramming plasmids. These results demonstrated that genomic integration and/or retention of the episomal reprogramming vectors had not occurred. The primers utilized

for the amplification of *EBNA-1* and *PXCLE* are shown in **Table 4.2**.

Table 4.2. Primers utilized for PCR and qPCR assays in order to assess the feeder free iPSCs for genomic integration and/or retention of episomal reprogramming factors (*EBNA-1* and *PCXLE*) as well as for the expression of pluripotency transcription factors (*OCT3/4, SOX2*, and *NANOG*). qPCR data were normalized to *GAPDH* expression.

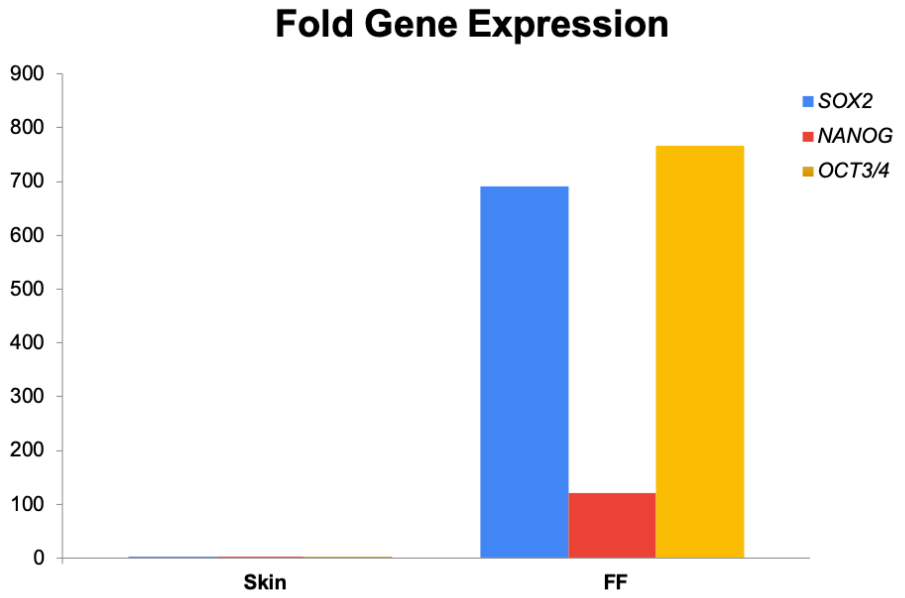
Name	Forward	Reverse	Purpose
GAPDH	ACCAACTGCTTAGCACCCCTGG	ATCACGCCACAGTTTCCCGGAG	Normalizing gene
OCT3/4(CDS)	CCCCAGGGCCCCATTTTGGTACC	ACCTCAGTTTGAATGCATGGGAGAGC	Endogenous qPCR
SOX2 (CDS)	TTCACATGTCCCAGCACTACCAGA	TCACATGTGTGAGAGGGGCAGTGTGC	Endogenous qPCR
NANOG_CDR	CAGAAGGCCTCAGCACCTAC	ATTGTTCCAGGTCTGGTTGC	Endogenous qPCR
EBNA1_End1	ATC ATC ATC CGG GTC TCC ACC G	ATT GCA GGT AGG AGC GGG CTT TG	Genomic: Plasmid and EBV integration
PCXLE_Common_F	TTTGCAAGCAGCAGATTACGCGCAG	TTTGCTCACCCAGAAACGCTGGTGAAAG	Genomic: Plasmid

Next, I performed a series of experiments to determine if my iPSCs continued to express endogenous pluripotency factors in the absence of the episomal reprogramming plasmids and/or genomic integration of the reprogramming plasmids. I extracted RNA from feeder free iPSCs and cultured skin fibroblasts using RNAzol (Sigma Aldrich). I then compared the expression of these factors to cultured skin fibroblasts. As would be expected, each TF was expressed at higher levels in the feeder free iPSCs than in cultured skin fibroblasts (**Figure 4.5**). The primers utilized for the amplification of *OCT3/4, SOX2*, and *NANOG* are shown in **Table 4.2**. I normalized all expression levels to *GAPDH* as an internal control. Primers and primer sequences were provided courtesy of Jonathan Burnett and Bryan Pavlovic.

iPSC to Hematopoietic Stem and Progenitor Cell (HSPC) differentiation

After confirming the pluripotency of the feeder free iPSCs as detailed above, I performed a directed differentiation to hematopoietic stem and progenitor cells (HSPCs) using a protocol designed by my colleagues Drs. Paul Gadue and Deborah

Figure 4.5. qPCR assays for pluripotency transcription factors. Each transcription factor was expressed at significantly higher levels in feeder free iPSCs than in cultured skin fibroblasts, as would be expected with the successful generation of feeder free iPSCs.



French at the University of Pennsylvania (Mills et al., 2014). For this protocol, I first cultured the iPSCs in feeder-free conditions until they reached 70% confluency. I then collected iPSCs for characterization via flow cytometry to ensure that they co-expressed markers of pluripotency (SSEA3/SSEA4). I also collected cells for bulk RNA sequencing starting at day 0. I then introduced a series of hematopoietic growth factors (**Table 4.3**) to differentiate the iPSCs into HSPCs. This differentiation protocol is notable for the sequential production of an initial mesodermal stromal monolayer that co-expresses KDR and CD31. This stromal layer then begins to develop “islands” of HSPCs around day 6 of the differentiation protocol. These HSPC islands increase in density before forming dense clusters of HSPCs on day 9 of the differentiation. These HSPCs co-express CD41 and CD235. Ultimately, the HSPCs are released directly from the stromal monolayer and are capable of forming erythroid, myeloid, and megakaryocytic precursors (Mills et al., 2014).

Table 4.3. Growth factors for directed iPSC to HSPC differentiations. All values are in ng/mL. BMP4: bone morphogenic protein-4; VEGF: vascular endothelial growth factor; Wnt3a: wingless-type MMTV integration site-3a; bFGF: basic fibroblast growth factor; SCF: stem cell factor; Flt3L: fms-like tyrosine kinase 3 ligand; TPO: thrombopoietin; IL-6: interleukin 6.

Days	Medium	BMP4	VEGF	Wnt3a	bFGF	SCF	Flt3L	TPO	IL-6
0-1	RPMI	5	50	25					
2	RPMI	5	50		25				
3	SP34	5	50		25				
4-5	SP34		15		5				
6	SFD		50		50	50	5		
7-10	SFD		50		50	50	5	50	10

I monitored the cells during this differentiation both via gross morphology, which changes in a distinct pattern during each stage of the differentiation, as well as via flow cytometry on days 7 and 10 of the differentiation to ensure that the HSPCs were expressing CD41 and CD235. I also collected RNA for bulk RNA sequencing on days 8 and 10 of the differentiation after ensuring that CD41 and CD235 were co-expressed starting at day 7. For each differentiation, I differentiated my patient-derived *ANKRD26*$^{+/mut}$ iPSC line in tandem with a sex-matched control iPSC line derived from a healthy volunteer, which was provided by the Gilad group (line 24280). This control line has been characterized in prior publications from the Gilad group and has been confirmed to behave as a *bona fide* iPSC line (Burrows et al., 2016; Pavlovic et al., 2018; Thomas et al., 2015).

I performed three independent differentiations of each iPSC line, but differentiated each line in tandem with the appropriate controls and utilized the same reagents for each pair of control/mutant differentiations so as to minimize batch effects secondary to reagents.

Bulk RNA sequencing (RNA-seq) and library preparation

I extracted RNA using RNAzol (Sigma-Aldrich) adhering to the instructions from the manufacturer. Of note, the iPSC-HSPC differentiation generates three distinct cellular populations as determined by gross morphology: a stromal layer, HSPC islands, and HSPCs in solution. For these RNA experiments, I collected all cellular populations from the tissue culture well in question, then performed RNA extractions on the total cell population. I submitted 1 μg of each sample to the University of Chicago Functional Genomics Facility for library generation, including 1:100 diluted ERCC spike in controls (Fisher Scientific, KK8580). I then performed paired end sequencing (100 bp) using an Illumina NovaSeq 6000. For each differentiation time point, I collected three independent samples for each line (i.e., 3 independent WT samples and 3 *ANKRD26*$^{+/mut}$ samples on day 0, for 18 total samples over the course of three time points: days 0, 8, and 10).

RNA-seq alignment and quantification

I aligned raw reads in the fastq format to the hg19 reference genome using HISAT2 and the GENCODE annotation (GRCh37.p13) (Frankish et al., 2019; Kim et al., 2019). I quantified and compared gene expression using the Cufflinks package (Trapnell et al., 2012). I used igvtools count to create a tiled data file (.tdf) of the read distribution visualization, then used peak calling results (.bed files) to indicate the location of peaks. I then utilized Metascape to further analyze my list of differentially expressed (DE) genes for each experimental condition, including gene annotation, membership searches, and pathway enrichment analyses (Zhou et al., 2019). Additionally, I analyzed RNA-seq data from each pair of iPSC lines at each time point for statistically enriched pathways and

processes (GO, KEGG, canonical pathways, and hallmark gene sets), then performed unsupervised hierarchical clustering of any significant terms based on kappa-statistical similarities, using a 0.3 kappa score threshold. I compared each pathway and process enrichment analysis to all genes in the genome. In order to be included in the analysis, a term was required to have a p-value < 0.01, an enrichment factor > 1.5 (the ratio between observed counts:counts expected by chance), and a minimum count of 3. All terms were grouped based on membership similarities. The p-values were calculated via a cumulative hypergeometric distribution (Zar, 1999). The q-values were calculated with the Benjamini-Hochberg correction for multiple testing (Hochberg & Benjamini, 1990).

CRISPR/Cas9 gene editing

In order to generate isogenic controls for future work, I also utilized a CRISPR/Cas9 protocol previously published by the Gadue and French groups that has been optimized for use in iPSCs (Maguire et al., 2019). I generated guide RNAs (gRNAs) to target within ±20 bp of the target region so as to revert the *ANKRD26* mutation in my patient-derived iPSC line (c.-119C>G) to the WT allele (*ANKRD26*$^{+/corrected}$). I also generated gRNAs to introduce the mutant allele into a control line (*ANKRD26*$^{+/intro}$). This protocol utilizes a pair of single stranded oligonucleotides (ssODNs). One ssODN contained the desired mutation of interest and the second ssODN contained the WT sequence of interest. This approach reduces indel production at the second, non-targeted allele. These ssODNs flank the gRNA of interest and include blocking mutations introduced into the protospacer adjacent motif (PAM) site so as to reduce subsequent cleavage after recombination. I then utilized the aforementioned protocol and gRNAs to

introduce the *ANKRD26* c.-119C>G mutation into a sex matched control line. The gRNAs, ssODN, and PCR primers that amplify the region of interest are shown in **Table 4.4**.

Table 4.4. The gRNAs, single stranded oligonucleotides containing the WT variant of interest as well as the WT sequence of interest, and PCR primers designed to amplify the genomic target region of interest are shown below.

ANKRD26_Introduce_Design		
Name	**Sequence**	**60-mer oligos for Phusion**
ANKRD26 gRNA_1F_intro	GCCTCCCAAAGGAAACTCCGCGG	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCTCCCAAAGGAAACTCCG
ANKRD26 gRNA_1R_intro	CCGCGGAGTTTCCTTTGGGAGGC	GACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAACCGGAGTTTCCTTTGGGAGGC
Genotyping Primers	Product: 527 bp	
ANKRD26_5_UTR F:	CATGGAGCACACTTGACCAC	
ANKRD26_5_UTR R:	TACTCCAGTGGCACTCAGTC	
Mutant Oligonucleotide	GTGTCTAGCGGGATCGCTTGCTTGGTAACCCCGAGGGAGAGATTGGAAAGCGCGGAGTTTCCTTTGGGAGGCTGCGGCCAGCCGGGGCTGACTTGTTATG	
WT Oligonuleotide	GTGTCTAGCGGGATCGCTTGCTTGGTAACCCCGAGGGAGAGATTGGAAACCGCGCAGTTCCCTTTGGGAGGCTGCGGCCAGCCGGGGCTGACTTGTTATG	

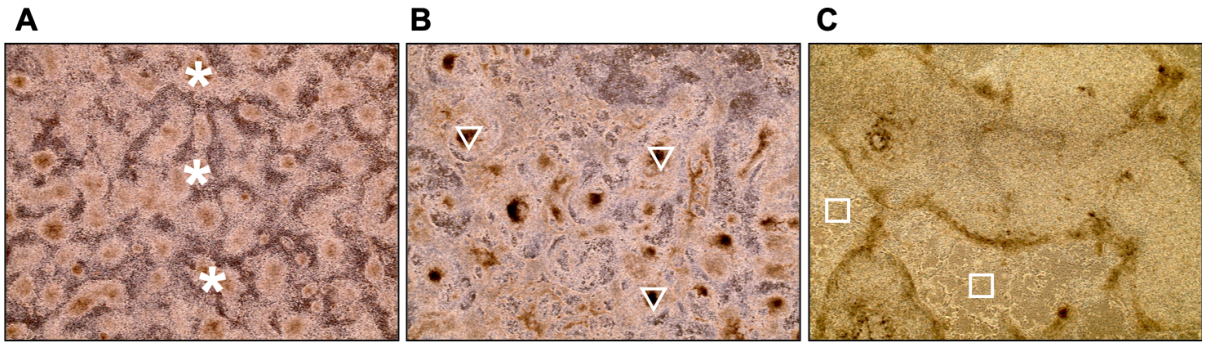
Results

ANKRD26-mutated iPSCs generate iPSC-derived HSPCs as anticipated

First, I performed directed HSPC differentiations to generate iPSC-derived HSPCs (iPSC-HSPCs) using my *ANKRD26*$^{+/mut}$ iPSC line in tandem with a sex-matched control line (24280). This demonstrated that there was no appreciable difference between the gross morphology of the *ANKRD26*$^{+/mut}$ iPSC line and the two control lines during the generation of iPSC-HSPCs (**Figure 4.6**). All three lines generated HSPCs that were visible via 4x microscopy in the HSPC islands. These HPSCs mobilized with gentle pipetting.

The *ANKRD26*$^{+/mut}$ iPSC line was also notable for the co-expression of the pluripotency markers SSEA3 and SSEA4 at day 0 of the differentiation, but then co-

Figure 4.6. Morphologic changes during iPSC to HSPC differentiation. A) stromal cells on day 4 (depicted with asterisks, *); B) HSPC "islands" on day 6 (depicted with triangles, ∇); C) prominent HSPC islands on day 9 (depicted with squares, □).



expressed CD41 and CD235 at day 10 of the HSPC differentiation (**Figure 4.7**). Each WT line also showed a similar pattern of co-expression of SSEA3/SSEA4 at day 0 and of CD41 and CD235 at day 10 (data not shown). An example of the gating strategy for flow cytometry studies is shown in **Figure 4.8**.

Figure 4.7. A) Flow cytometric analysis of undifferentiated iPSCs that co-express pluripotency markers (SSEA3, SSEA4). B) Analysis of iPSC-derived HSPCs, which co-express CD41 and CD235.

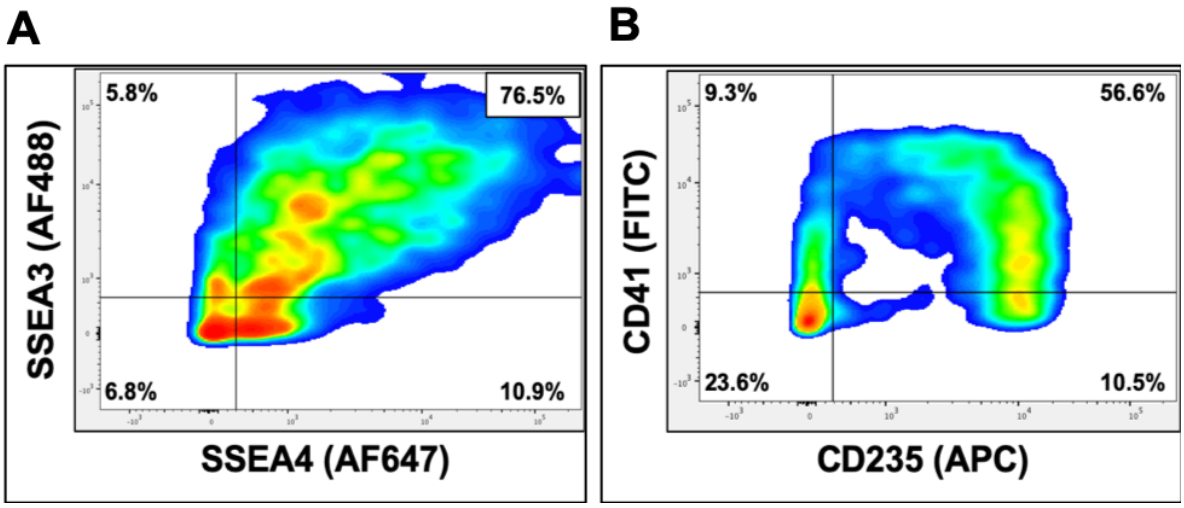
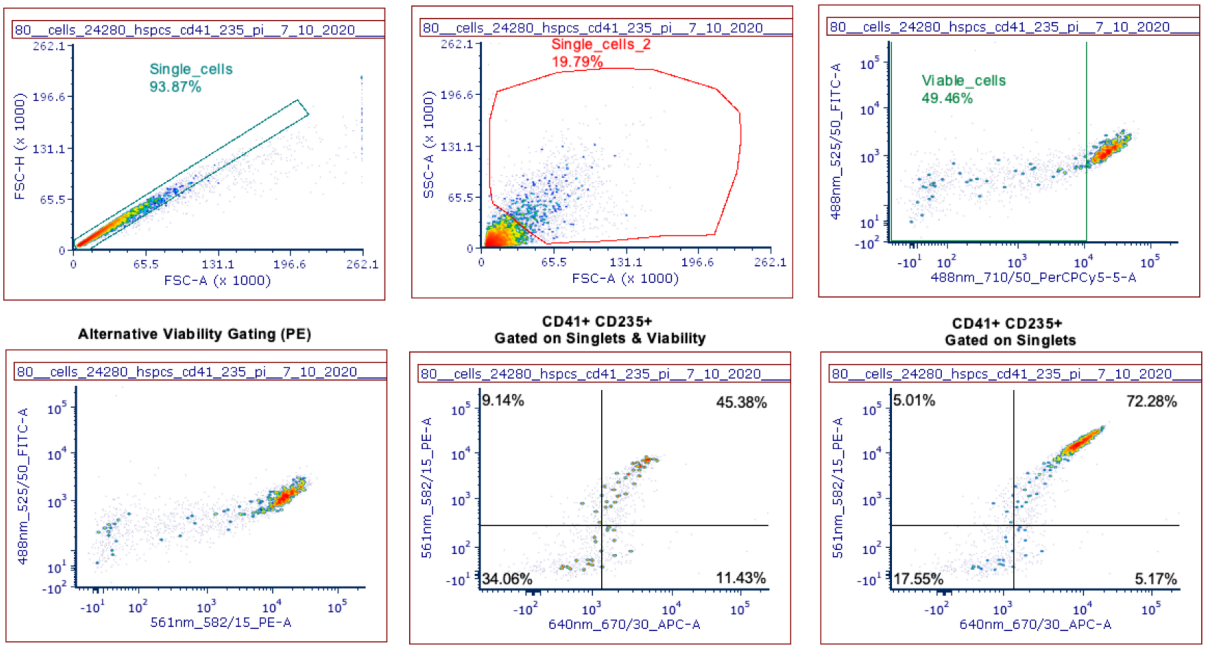


Figure 4.8. Gating strategy for flow cytometry experiments.



Germline ANKRD26 5' UTR-mutated and control iPSCs show expected hematopoiesis-related gene expression profiles during iPSC-HSPC differentiation

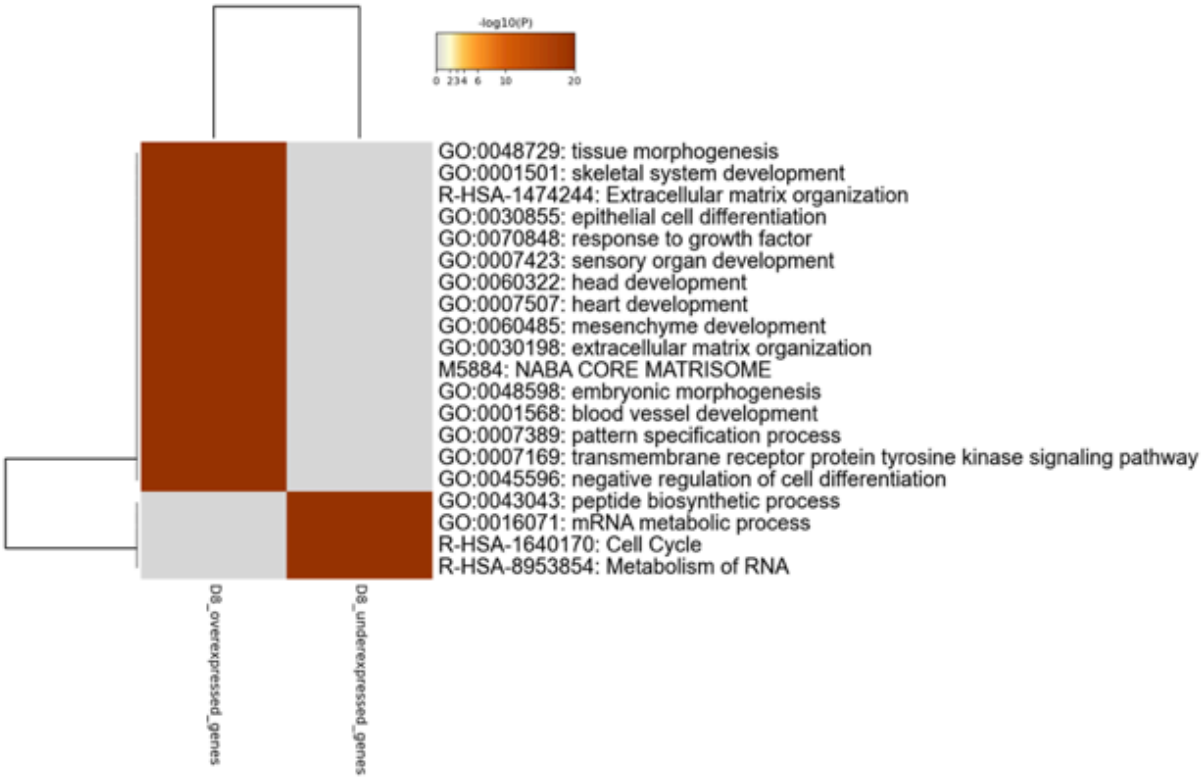
After confirming that my *ANKRD26*-mutated and control iPSCs underwent expected morphologic and flow cytometric changes during my directed iPSC-HSPC differentiation, I next analyzed the RNA expression profiles of each line in order to determine each line expressed a series of genes expected to be overexpressed in D8 HSPCs relative to D0 iPSCs during the directed iPSC-HSPC differentiation. This analysis demonstrated that 2939 and 2656 unique genes were over- and under-expressed, respectively, in the control HSPCs at day 8 of the directed differentiation relative to control iPSCs. I then performed a pathway and process enrichment analysis on these differentially expressed genes, which correlated with biological processes that would be expected during hematopoiesis, including embryonic morphogenesis (GO:0048598),

tissue morphogenesis (GO:0048729), heart development (GO:0007507), blood vessel development (GO:0001568), mesenchymal development (GO:0060485), and others (**Table 4.5**). A heatmap of an unbiased hierarchical clustering of these processes is shown in **Figure 4.9**.

Table 4.5. Ranking of the most significantly enriched molecular pathways in the control line from day 0 (iPSCs) through day 8 of iPSC-HSPC differentiation.

GO	Category	Description	Count	%	Log10(P)	Log10(q)
R-HSA-1474244	Reactome Gene Sets	Extracellular matrix organization	143	4.95	-60.76	-56.40
GO:0001568	GO Biological Processes	blood vessel development	238	8.24	-57.07	-53.01
GO:0048729	GO Biological Processes	tissue morphogenesis	198	6.86	-50.15	-46.27
GO:0009611	GO Biological Processes	response to wounding	180	6.23	-46.79	-43.12
GO:0034330	GO Biological Processes	cell junction organization	206	7.13	-46.19	-42.61
GO:0007169	GO Biological Processes	transmembrane receptor protein tyrosine kinase signaling pathway	191	6.61	-44.49	-41.04
GO:0048598	GO Biological Processes	embryonic morphogenesis	184	6.37	-44.10	-40.70
GO:0070848	GO Biological Processes	response to growth factor	205	7.10	-43.00	-39.67
M5885	Canonical Pathways	NABA MATRISOME ASSOCIATED	140	9.92	-41.87	-38.29
GO:0006935	GO Biological Processes	chemotaxis	182	6.30	-41.54	-38.38
GO:0007507	GO Biological Processes	heart development	177	6.13	-39.88	-36.77
GO:0003013	GO Biological Processes	circulatory system process	177	6.13	-39.55	-36.47
GO:0000904	GO Biological Processes	cell morphogenesis involved in differentiation	194	6.72	-39.41	-36.37
M5884	Canonical Pathways	NABA CORE MATRISOME	111	3.84	-38.75	-35.79
GO:0001501	GO Biological Processes	skeletal system development	158	5.47	-38.20	-35.29
GO:0030155	GO Biological Processes	regulation of cell adhesion	199	6.89	-36.79	-33.91
R-HSA-109582	Reactome Gene Sets	Hemostasis	175	6.06	-36.04	-33.19
GO:0000165	GO Biological Processes	MAPK cascade	201	6.96	-35.72	-32.90
R-HSA-9006934	Reactome Gene Sets	Signaling by Receptor Tyrosine Kinases	156	5.40	-35.52	-32.72
GO:0001667	GO Biological Processes	ameboidal-type cell migration	147	5.09	-34.93	-32.14

Figure 4.9. Heatmap of statistically enriched molecular pathways in the control line during days 0 to 8 of the iPSC-HSPC differentiation. This analysis demonstrated that pathways known to be involved in hematopoiesis were active during the iPSC-HSPC differentiation in the control line.



Next, I performed a similar series of analyses using the RNA-seq data generated from the iPSC-HSPC directed differentiation that I performed with the *ANKRD26*$^{+/mut}$ line. This analysis demonstrated that 680 and 575 unique genes were over- and under-expressed, respectively, between day 0 (iPSC) and day 8 (HSPC) of the iPSC-HSPC differentiation. The most significantly enriched molecular pathways in the *ANKRD26*$^{+/mut}$ line are shown in **Table 4.6**, and a heatmap of the unbiased hierarchical clustering from this line is shown in **Figure 4.10**. This analysis demonstrated that a very similar set of pathways and processes were over-expressed during hematopoiesis in the *ANKRD26*$^{+/mut}$ line as compared to the control line. Notably, the MAPK pathways were

not dysregulated during hematopoieisis in either the control or *ANKRD26*$^{+/mut}$ lines during hematopoiesis.

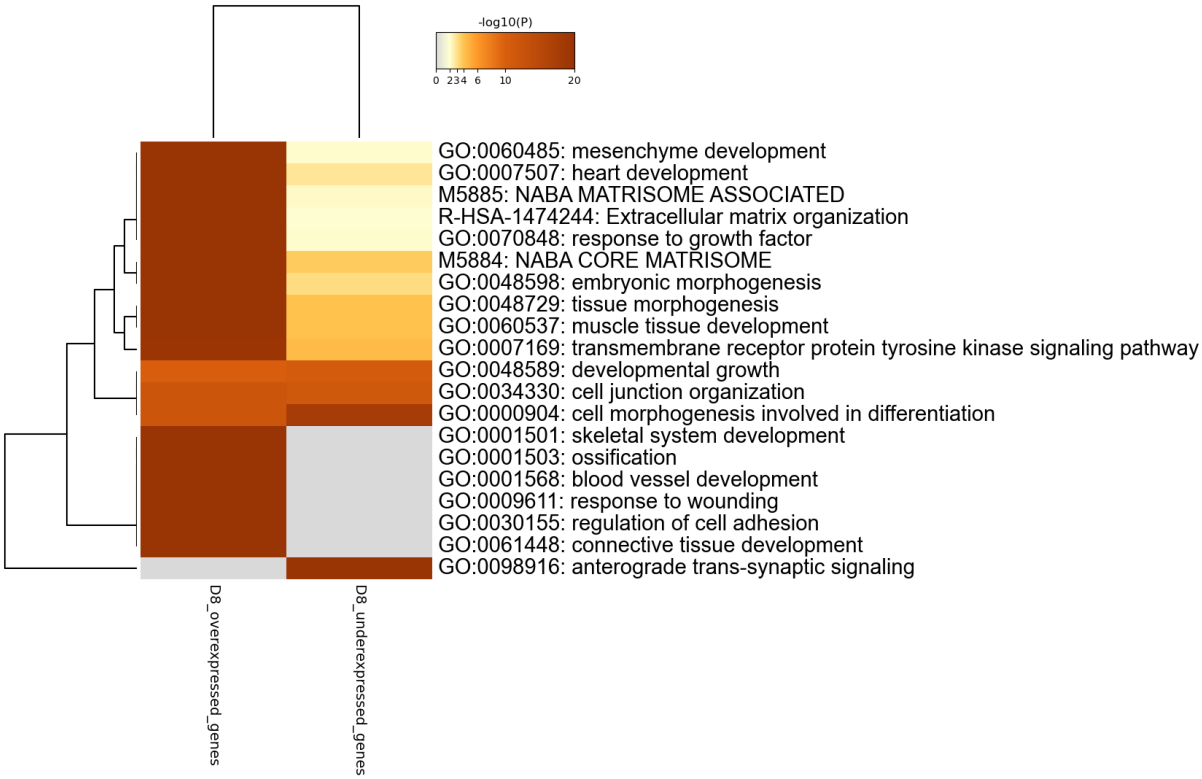
Table 4.6. Ranking of the most significantly enriched molecular pathways in the *ANKRD26*$^{+/mut}$ line from day 0 (iPSCs) through day 8 of iPSC-HSPC differentiation.

GO	Category	Description	Count	%	Log10(P)	Log10(q)
GO:0001568	GO Biological Processes	blood vessel development	114	16.84	-55.76	-51.40
GO:0007507	GO Biological Processes	heart development	85	12.56	-40.01	-36.25
GO:0001501	GO Biological Processes	skeletal system development	74	10.93	-35.34	-31.76
GO:0048729	GO Biological Processes	tissue morphogenesis	107	8.58	-33.68	-30.02
R-HSA-1474244	Reactome Gene Sets	Extracellular matrix organization	56	8.27	-32.43	-28.97
GO:0000904	GO Biological Processes	cell morphogenesis involved in differentiation	108	8.66	-29.81	-26.28
M5884	Canonical Pathways	NABA CORE MATRISOME	66	5.29	-29.36	-25.95
M5885	Canonical Pathways	NABA MATRISOME ASSOCIATED	81	11.96	-29.27	-25.95
GO:0070848	GO Biological Processes	response to growth factor	79	11.67	-29.04	-25.80
GO:0098916	GO Biological Processes	anterograde trans-synaptic signaling	71	12.46	-27.84	-24.71
GO:0060537	GO Biological Processes	muscle tissue development	75	6.01	-25.68	-22.66
GO:0009611	GO Biological Processes	response to wounding	65	9.60	-25.60	-22.54
GO:0048598	GO Biological Processes	embryonic morphogenesis	67	9.90	-25.25	-22.22
GO:0001503	GO Biological Processes	ossification	55	8.12	-24.40	-21.42
GO:0034330	GO Biological Processes	cell junction organization	98	7.86	-23.70	-20.77
GO:0060485	GO Biological Processes	mesenchyme development	46	6.79	-23.48	-20.56
GO:0030155	GO Biological Processes	regulation of cell adhesion	72	10.64	-22.90	-20.04
GO:0061448	GO Biological Processes	connective tissue development	42	6.20	-22.24	-19.46
GO:0007169	GO Biological Processes	transmembrane receptor protein tyrosine kinase signaling pathway	90	7.22	-22.23	-19.31
GO:0048589	GO Biological Processes	developmental growth	87	6.98	-20.59	-17.84

I then performed a similar set of analyses comparing the gene expression profiles of HSPCs collected from the control or *ANKRD26*$^{+/mut}$ lines on days 8 and 10 of the iPSC-HSPC differentiation protocol. This analysis demonstrated that, on a transcriptomic level, very little genetic differential expression existed between HSPCs harvested on days 8 and 10 of the iPSC-HSPC differentiation protocol. For example, in the control line, only 29 and 11 unique genes were over- and under-expressed, respectively, between days 8

and 10. In the *ANKRD26*$^{+/mut}$ line, only 39 and 22 unique genes were significantly over- and under-expressed, respectively, between days 8 and 10. Both of these findings are consistent with the gross morphologic changes that were observed during the differentiation protocol, as the majority of the stromal cells and HSPC “islands” were well formed by day 8 of the differentiation protocol. Morphologically, the primary changes that occurred between days 8 and 10 were the release of HSPCs from the islands embedded in the stroma into the media itself (**Figure 4.6**). In transcriptomic analyses of both the control and *ANKRD26*$^{+/mut}$ lines, the pathways that were differentially expressed between days 8 and 10 were typical of hematopoiesis, including ECM proteoglycans (R-HSA-3000178), blood vessel development (GO:0001568), and regulation of cell growth (GO:0001558) (data not shown). Therefore, these transcriptomic signatures were considered to be typical hematopoiesis-related events.

Figure 4.10. Heatmap of statistically enriched molecular pathways in the *ANKRD26*$^{+/mut}$ line during days 0 through 8 of the iPSC-HSPC differentiation. This analysis demonstrated that pathways known to be involved in hematopoiesis were active during the iPSC-HSPC differentiation in the *ANKRD26*$^{+/mut}$ line.



Germline ANKRD26 5' UTR mutations lead to over-expression of the 5' UTR targeted transcription factors RUNX1 and FLI1 in iPSCs and HSPCs

Next, I investigated the effects of germline 5' UTR *ANKRD26* mutations on the expression of *RUNX1* and *FLI1* during hematopoiesis. As shown in **Figure 3.1**, the 5' UTR of *ANKRD26* contains binding sites for both RUNX1 and FLI1 (Tijssen et al., 2011). *ANKRD26* expression is negatively regulated by the binding of RUNX1 and/or FLI1. This makes *ANKRD26* unique among the HT/HHMs, as well as the larger group of HHMs, in that it is the only HHM-related gene in which LP/P mutations in ClinVar with adequate supporting evidence are located exclusively in a genomic region that leads to

dysregulation of the gene in question. However, it is not known if *ANKRD26* 5' UTR mutations in patients or patient-derived models result in a corresponding disruption in the regulation of *RUNX1* or *FLI1* expression.

Table 4.7. Expression data for specific genes relevant to HT-HHMs (*ANKRD26*, *RUNX1*), HT (*FLI1*, *TUBB1*), and HHM (*TP53*) syndromes.

Gene	Control	*ANKRD26*-mutated	log2(fold_change)	p-value	q-value	Significant?
Day 0 iPSCs						
ANKRD26	3.06	2.86	-0.09	0.88	0.97	No
FLI1	0.08	0.15	0.92	1	1	No
RUNX1	0.11	0.18	0.65	1	1	No
TP53	20.55	90.83	2.14	0.00	0.01	Yes
TUBB1	0.12	0.05	1.14	1	1	No
Day 8 HSPCs						
ANKRD26	8.57	2.11	-2.11	0.00	0.02	Yes
FLI1	3.89	16.26	2.06	0.00	0.02	Yes
RUNX1	0.81	3.08	1.93	0.00	0.03	Yes
TP53	12.09	60.85	2.33	0.00	0.01	Yes
TUBB1	0.05	0.19	1.99	1	1	No
Day 10 HSPCs						
ANKRD26	9.10	8.57	-0.09	0.80	0.95	No
FLI1	5.13	15.86	1.63	0.00	0.00	Yes
RUNX1	1.38	3.47	1.33	0.00	0.02	Yes
TP53	13.05	30.58	1.23	0.00	0.00	Yes
TUBB1	0.02	0.29	3.96	1	1	No

I addressed this question by analyzing expression data for *FLI1* and *RUNX1* in the control and *ANKRD26*$^{+/mut}$ lines during a directed iPSC-HSPC differentiation. There were no significant differences observed in the expression of either gene in the pluripotent state. However, both *FLI1* and *RUNX1* were over-expressed in the *ANKRD26*$^{+/mut}$ line on days 8 and 10 of the HSPC differentiation relative to controls (**Table 4.7**). This finding suggests that germline 5' UTR *ANKRD26* mutations not only abrogate RUNX1 and FLI1 binding, but they may also result in a disruption of the global transcription profile of both TFs.

Germline ANKRD26 5' UTR mutations do not lead to over-expression of ANKRD26 in iPSCs and HSPCs

Prior work from Bluteau *et al.* demonstrated that RUNX1 normally downregulates *ANKRD26* expression during normal megakaryopoiesis, with *ANKRD26* expression decreasing significantly megakaryocytes form proplatelets. In contrast, shRNA-mediated knockdown of *RUNX1* led to persistent over-expression of *ANKRD26* throughout megakaryopoiesis, which is thought to partly drive the decreased platelet maturation and thrombocytopenia that is observed in individuals with germline *ANKRD26* mutations. Similarly, germline *RUNX1* mutations frequently disrupt the binding of RUNX1 to the 5' UTR of *ANKRD26*, leading to persistent *ANKRD26* expression during megakaryopoiesis (Bluteau et al., 2014). Importantly, 5' UTR mutations in *ANKRD26*, including at the c.-119 site, disrupt both the RUNX1 and FLI1 binding sites. Mutations at the c.-119 site have been shown to drive perpetual overexpression of *ANKRD26* throughout the process of megakaryopoiesis (Bluteau et al., 2014).

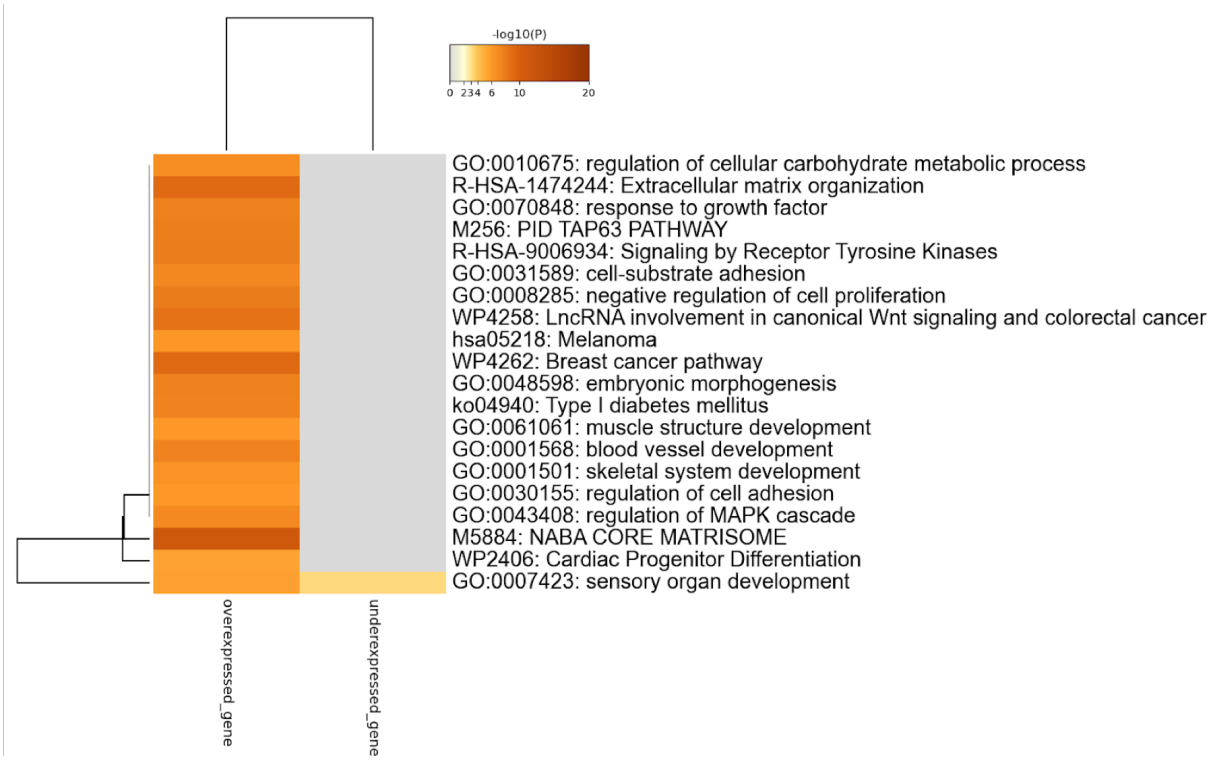
However, it is unknown if *ANKRD26* is similarly over-expressed in either the pluripotent state or in HSPCs. To address this gap, I analyzed the expression data from *ANKRD26*$^{+/mut}$ and control iPSCs undergoing a directed differentiation to HSPCs. This analysis showed that no significant differences in *ANKRD26* expression were observed between these two lines in the pluripotent state or in "late" HSPCs on day 10 of the differentiation. However, *ANKRD26* expression was decreased in the day 8 HSPCs generated from the *ANKRD26*$^{+/mut}$ line relative to controls. These findings suggest that *ANKRD26* 5' UTR mutations do not lead to increased *ANKRD26* expression in iPSCs or HSPCs, unlike what has been observed in *ANKRD26*$^{+/mut}$ megakaryocytes and platelets.

Germline ANKRD26 5' UTR mutations are associated with constitutive overexpression of the MAPK/ERK pathways in the pluripotent state

I then analyzed the differentially expressed (DE) genes during the process of HSPC differentiation in each line in a more unbiased fashion. First, I analyzed the DE genes of the undifferentiated iPSCs (day 0 of the iPSC differentiation) based on genotype, comparing the control line (24280) to the *ANKRD26*$^{+/mut}$ iPSC line.

This D0 analysis demonstrated that a total of 337 unique genes were over-expressed and 76 genes were significantly under-expressed in the *ANKRD26*$^{+/mut}$ line as compared to the WT line. I then performed unbiased hierarchical clustering to identify the 20 most statistically enriched pathways (GO biological processes, KEGG, Reactome gene sets, PANTHER pathways, canonical pathways, and hallmark gene sets) in the *ANKRD26*$^{+/mut}$ line during the pluripotent state (D0). The ensuing heatmap from this analysis is shown in **Figure 4.11.**

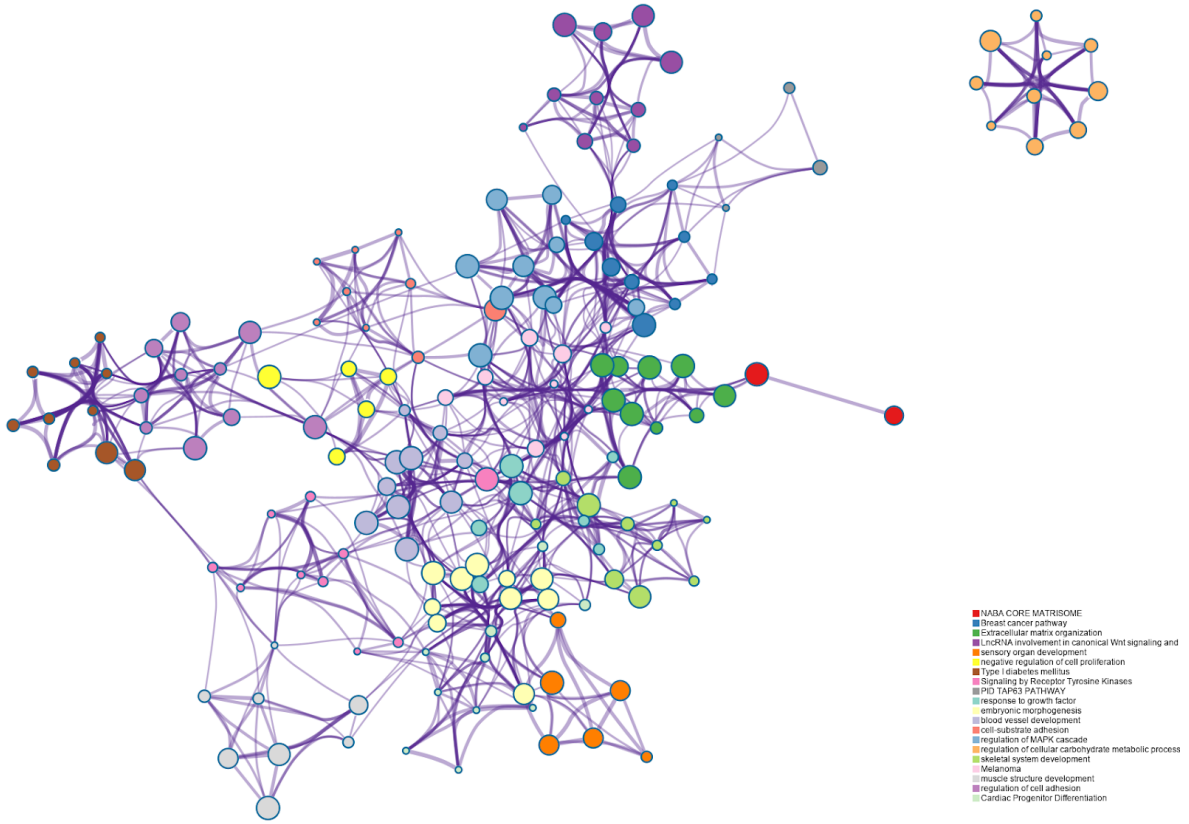
Figure 4.11. Heatmap of statistically enriched molecular pathways in the *ANKRD26*$^{+/mut}$ line relative to a WT control line in the pluripotent state (D0 of HSPC differentiation, prior to the addition of HSPC-related cytokines.



Intriguingly, one of the top 20 molecular pathways that was overexpressed on D0 was GO:0043408: regulation of MAPK cascade. I then converted these pathways into a network layout, with each pathway term represented by a node whose size corresponded to the number of genes falling within the term. This analysis, shown in **Figure 4.12**, demonstrated that a number of molecular pathways, such as regulation of the MAPK cascade, embryonic morphogenesis, signaling by receptor tyrosine kinases, and blood vessel development, were densely connected within the larger pathway analysis. I then applied a molecular complex detection (MCODE) algorithm to these PPI networks in order to identify the most closely related complexes as well as to identify the terms in each complex with the most significant p-values (**Figure 4.13**) (Bader & Hogue, 2003). This analysis demonstrated that a number of pathways known to be relevant to the HT-HHMs,

such as platelet degranulation and response to elevated platelet cytosolic calcium, were enriched in the *ANKRD26*$^{+/mut}$ line during the pluripotent state (D0).

Figure 4.12. Network analysis of statistically enriched molecular pathways in the *ANKRD26*$^{+/mut}$ line relative to a WT control line in the pluripotent state (D0 of HSPC differentiation, prior to the addition of HSPC-related cytokines. Regulation of MAPK cascades, embryonic morphogenesis, tyrosine receptor kinase signaling, and angiogenesis were densely connected with the larger pathway analysis.

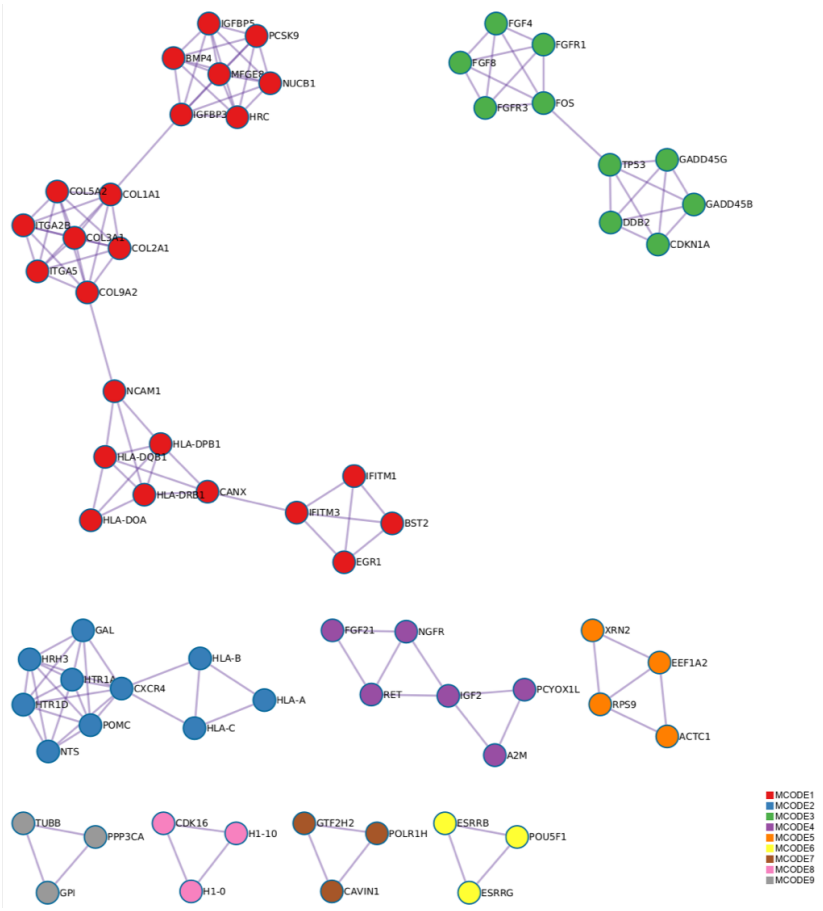


Germline *ANKRD26* 5' UTR mutations are associated with constitutive overexpression of the MAPK/ERK pathways during days 8 and 10 of iPSC to HSPC differentiation

Next, I performed a similar series of analyses using RNA-seq data from days 8 and 10 of the iPSC-HSPC differentiation. This analysis identified 1003 and 613 unique genes that were significantly over- and under-expressed, respectively, in the *ANKRD26*$^{+/mut}$ line as compared to the WT line on day 8 of the differentiation. There were

1421 and 1135 unique genes over- and under-expressed, respectively, in the mutant line as compared to the WT line on day 10 of the differentiation. I then performed a pathway analysis comparing *ANKRD26*$^{+/mut}$ HSPCs and control HSPCs produced after 10 days of differentiation. Once again, the MAPK cascade was a statistically enriched term in the *ANKRD26*$^{+/mut}$ HSPCs relative to the control HSPCs, suggesting that dysregulation of the MAPK pathways remain present during hematopoiesis in the iPSC-HSPC differentiation.

Figure 4.13. MCODE analysis of statistically enriched molecular pathways in the *ANKRD26*$^{+/mut}$ line in the pluripotent state (D0 of HSPC differentiation). The most significant terms contributing to each pathway are displayed. Pathways relevant to HT-HHMs, such as platelet degranulation and response to elevated platelet cytosolic calcium, were enriched in this output.



Germline ANKRD26 5' UTR mutations are associated with constitutive overexpression of the MAPK/ERK pathways in both pluripotent cells and HSPCs

Next, in order to identify genes and pathways that were uniformly disrupted in both the pluripotent state and HSPCs, I identified the genes that were differentially expressed in iPSCs (day 0 of the protocol), "early" HSPCs (day 8), and "mature" HSPCs (day 10). In total, there were 22 genes that were under-expressed in each time point ("constitutively under-expressed" genes in the *ANKRD26*$^{+/mut}$ line) and there were 84 genes that were over-expressed in each time point ("constitutively over-expressed" genes in the *ANKRD26*$^{+/mut}$ line) (**Figure 4.14**).

Figure 4.14. The total number of significantly under- and over-expressed genes at days 0 (iPSCs), 8 ("early" HSPCs), and 10 ("mature" HSPCs) are shown. In total, 22 unique genes were constitutively under-expressed in both the iPSCs and HSPCs. In contrast, 84 genes were constitutively over-expressed in iPSCs and HSPCs.

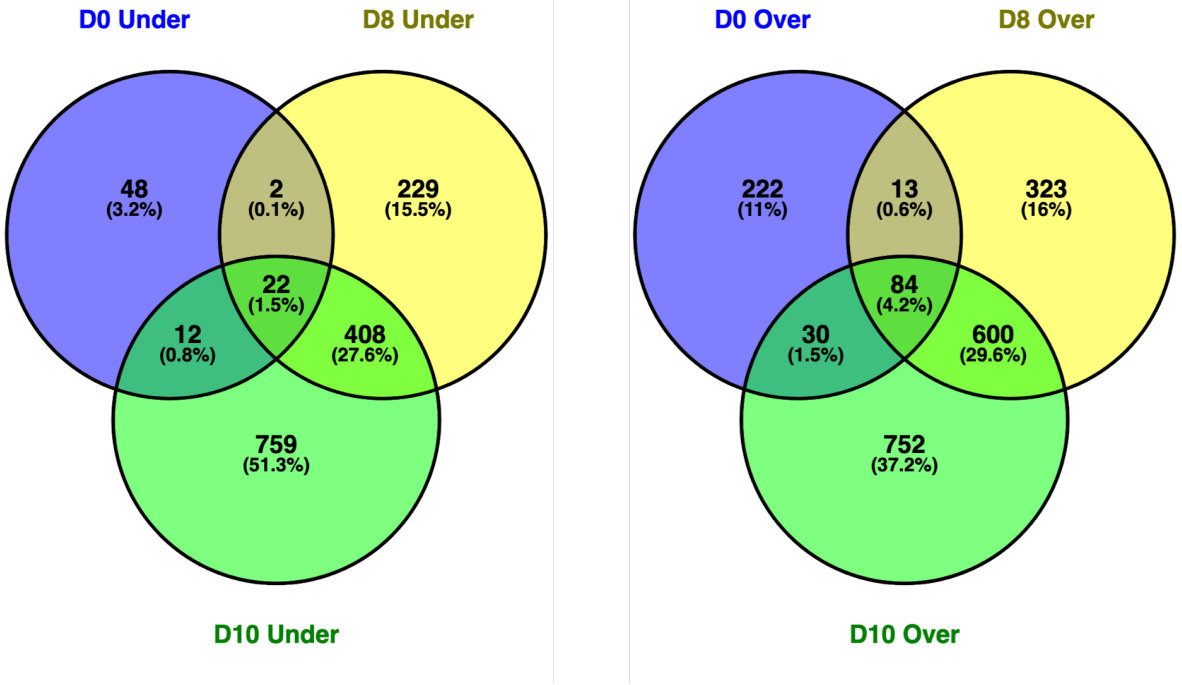
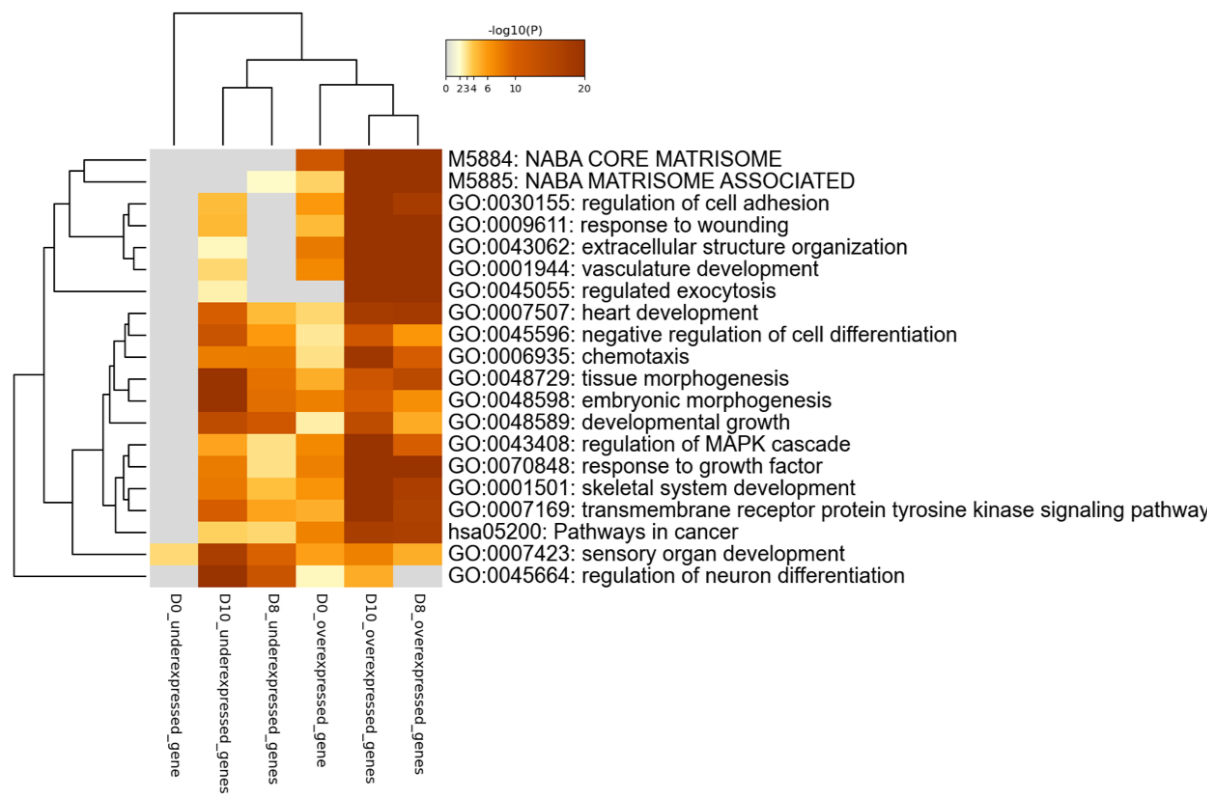


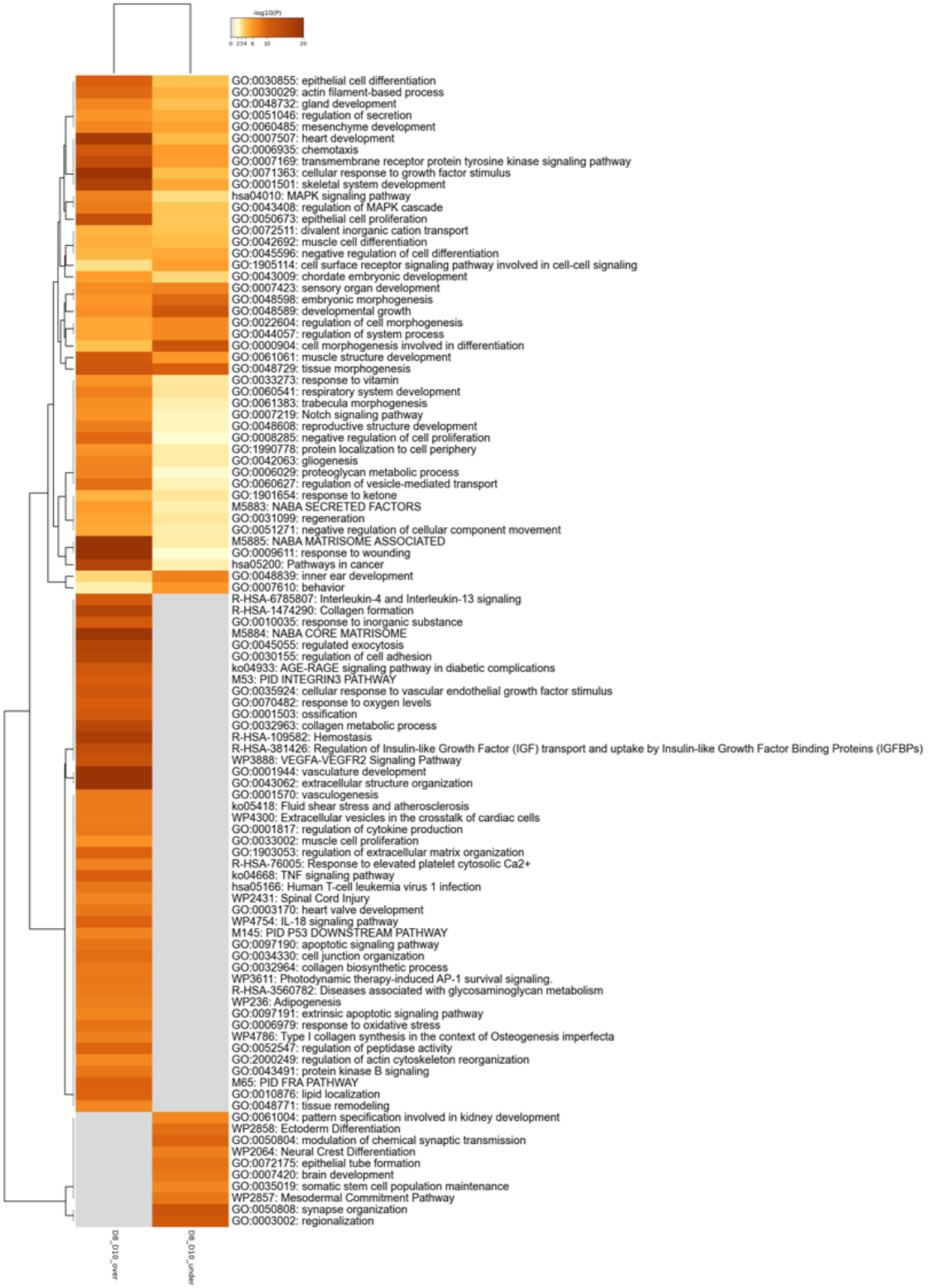
Figure 4.15. Heatmap of the 20 most significantly, constitutively, enriched molecular pathways in the *ANKRD26*$^{+/mut}$ line relative to a control line in the pluripotent state (day 0, iPSCs) through days 8 and 10 of iPSC-HSPC differentiation. This analysis demonstrated that the MAPK cascade was dysregulated pathways in the *ANKRD26*$^{+/mut}$ line at each stage of the iPSC-HSPC differentiation.



I then performed an unbiased hierarchical clustering using these constitutively under- and over-expressed genes so as to determine the most enriched pathways in the *ANKRD26*$^{+/mut}$ line relative to the control line during each step of the iPSC-HSPC differentiation. The heatmap from this analysis representing the 20 most significantly enriched pathways is shown in **Figure 4.15**. This heatmap is notable for persistently dysregulated expression of the MAPK cascade in addition to a number of other pathways that were also dysregulated in individual time points (days 0, 8, and 10). I also performed a pathway and process analysis of the 100 most significantly enriched pathways that were

constitutively dysregulated in the *ANKRD26*$^{+/mut}$ line relative to the control line during iPSC-HSPC differentiation; this heat map is displayed in **Figure 4.16**.

Figure 4.16. Heatmap of the 100 most significantly, constitutively, enriched molecular pathways in the *ANKRD26*$^{+/mut}$ line relative to a control line in the pluripotent state (day 0, iPSCs) through days 8 and 10 of iPSC-HSPC differentiation.



When ranked by statistical significance, the MAPK cascade pathway (GO:0000165) was the 18th most significant pathway among those that were constitutively over- or under-expressed in the *ANKRD26*$^{+/mut}$ line during iPSC-HSPC differentiation (**Table 4.8**). Additionally, a number of terms relevant to hematopoiesis, such as extracellular structure organization (R-HAS-1474244), blood vessel development (GO:0001568), response to wounding (GO:0009611), and hemostasis (R-HAS-109582), among others, were constitutively differentially expressed in the *ANKRD26*$^{+/mut}$ line during iPSC-HSPC differentiation.

Table 4.8. Ranking of the most significantly, constitutively, enriched molecular pathways in the *ANKRD26*$^{+/mut}$ line relative to a control line in the pluripotent state (day 0, iPSCs) through days 8 and 10 of iPSC-HSPC differentiation.

GO	Category	Description	Count	%	Log10(P)	Log10(q)
R-HSA-1474244	Reactome Gene Sets	Extracellular matrix organization	143	4.95	-60.76	-56.40
GO:0001568	GO Biological Processes	blood vessel development	238	8.24	-57.07	-53.01
GO:0048729	GO Biological Processes	tissue morphogenesis	198	6.86	-50.15	-46.27
GO:0009611	GO Biological Processes	response to wounding	180	6.23	-46.79	-43.12
GO:0034330	GO Biological Processes	cell junction organization	206	7.13	-46.19	-42.61
GO:0007169	GO Biological Processes	transmembrane receptor protein tyrosine kinase signaling pathway	191	6.61	-44.49	-41.04
GO:0048598	GO Biological Processes	embryonic morphogenesis	184	6.37	-44.10	-40.70
GO:0070848	GO Biological Processes	response to growth factor	205	7.10	-43.00	-39.67
M5885	Canonical Pathways	NABA MATRISOME ASSOCIATED	140	9.92	-41.87	-38.29
GO:0006935	GO Biological Processes	chemotaxis	182	6.30	-41.54	-38.38
GO:0007507	GO Biological Processes	heart development	177	6.13	-39.88	-36.77
GO:0003013	GO Biological Processes	circulatory system process	177	6.13	-39.55	-36.47
GO:0000904	GO Biological Processes	cell morphogenesis involved in differentiation	194	6.72	-39.41	-36.37
M5884	Canonical Pathways	NABA CORE MATRISOME	111	3.84	-38.75	-35.79
GO:0001501	GO Biological Processes	skeletal system development	158	5.47	-38.20	-35.29
GO:0030155	GO Biological Processes	regulation of cell adhesion	199	6.89	-36.79	-33.91
R-HSA-109582	Reactome Gene Sets	Hemostasis	175	6.06	-36.04	-33.19
GO:0000165	GO Biological Processes	MAPK cascade	201	6.96	-35.72	-32.90
R-HSA-9006934	Reactome Gene Sets	Signaling by Receptor Tyrosine Kinases	156	5.40	-35.52	-32.72
GO:0001667	GO Biological Processes	ameboidal-type cell migration	147	5.09	-34.93	-32.14

Germline ANKRD26 5' UTR mutations are associated with constitutive over-expression of TP53 in iPSCs and iPSC-HSPCs

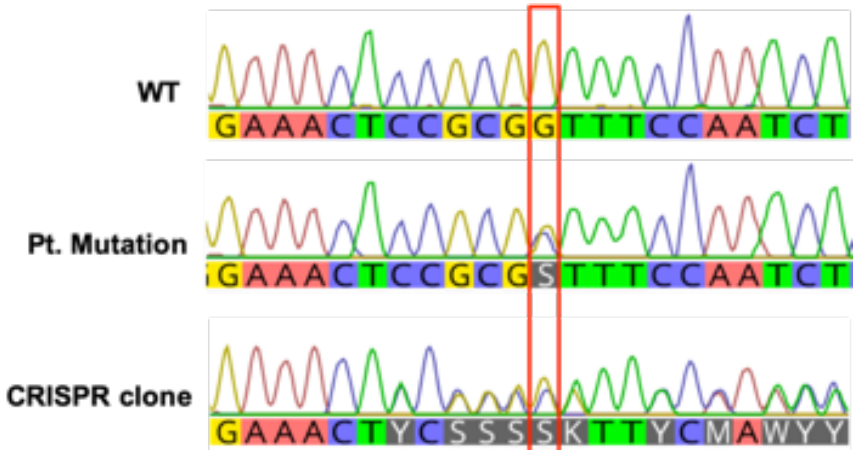
Prior work from Cai *et al.* demonstrated that *TP53* mRNA levels in *Runx1* deficient mice were similar to WT mice, but that p53 protein levels were lower in *Runx1* deficient mice (Cai et al., 2015). Given the similar phenotype between *RUNX1* and *ANKRD26* mutations in humans, I specifically analyzed expression data for *TP53*. This demonstrated that *TP53* mRNA is constitutively over-expressed in the *ANKRD26*$^{+/mut}$ line during all stages of the iPSC-HSPC differentiation (**Table 4.7**). Among the 84 genes that were constitutively over-expressed during all stages of the iPSC-HSPC differentiation, I identified a number of known targets of p53 previously identified by Andrysik *et al.*, including *CDKN1A* and *FDXR* (Andrysik et al., 2017).

Generation of an isogenic ANKRD26+/mut line for future experimental studies

In addition to the analyses above, I also performed a series of CRISPR/Cas9 gene editing experiments in order to generate an isogenic control line for future experimental studies using a protocol published previously by my collaborators, Drs. Deborah French and Paul Gadue (Maguire et al., 2019). Ultimately, I generated gRNAs to introduce the *ANKRD26* mutation that was present in my patient-derived iPSC line (c.-119C>G) into the control iPSC line (*ANKRD26*$^{+/intro}$). This process was ultimately successful and generated an *ANKRD26*$^{+/intro}$ iPSC line in which a frameshift mutation disrupted the 5' UTR of *ANKRD26* with 4% efficiency (**Figure 4.17**). However, shortly after generating this line I made the difficult decision to pause this project secondary to the constraints of the

COVID19 pandemic. This *ANKRD26*$^{+/intro}$ line is now cryopreserved and ready for quality control analyses (karyotyping, qPCR assays, pluripotency assays) in preparation for the next stage of experiments for this project. These experiments will include colony forming assays, flow cytometric studies, and liquid expansion assays (megakaryocytes, myeloid precursors, and erythroid precursors) in addition to unbiased transcriptomic studies (bulk RNA sequencing of iPSCs, HSPCs, megakaryocytes, myeloid precursors, and erythroid precursors).

Figure 4.17. Chromatogram of *ANKRD26* 5' UTR. The highlighted region corresponds to c.-119. The CRISPR clone (bottom) carries the same c.-119C>G mutation that was identified in a patient with HT/HHM who carried this germline mutation. Of note, the CRISPR/Cas9 gene editing protocol in this experiment was 4% efficient and introduced a frameshift mutation at this residue that disrupted multiple residues within the *ANKRD26* 5' UTR.



Discussion

Prior work has demonstrated that the MAPK/ERK1/ERK2 pathways are activated during early phases of normal megakaryopoiesis, but then diminish in activity during later stages of megakaryopoiesis, when proplatelets are generated (Bluteau et al., 2014; Minamiguchi et al., 2001). In contrast, Bluteau *et al*. demonstrated that germline *ANKRD26* 5' UTR mutations lead to persistent MAPK/ERK1/ERK2 hyperactivation during

late stages of megakaryopoiesis. This hyperactivation can be reversed via MEK inhibition (PD98059), which targets the MAPK/ERK1/2 pathways. Of note, PD98059 is not a rational clinical agent secondary to insolubility (Wang et al., 2007). In their 2014 manuscript, Bluteau *et al.* noted that dysregulated MAPK/ERK1/ERK2 signaling has also been linked to HMs, which led them to speculate that a similar mechanism, if present in HSPCs, may lead to the increased leukemogenic risk that is observed in patients with germline *ANKRD26* 5' UTR mutations (Bluteau et al., 2014). However, they were unable to investigate if MAPK/ERK1/ERK2 dysregulation was present in HSPCs collected from patients with *ANKRD26* germline mutations. To date, this has remained an uninvestigated issue in the field of HT/HHMs.

To address this knowledge gap, I generated a patient-derived iPSC line carrying the c.-119C>G mutation in the 5' UTR of *ANKRD26* in collaboration with the group of Dr. Yoav Gilad at the University of Chicago. This is a known likely pathogenic mutation that is listed in ClinVar. This represents only the second known patient-derived line from a patient with a germline *ANKRD26* mutation (Tan et al., 2020). I then demonstrated that my *ANKRD26*$^{+/mut}$ line behaves as a *bona fide* pluripotent stem cell via karyotyping, qPCR analyses, pluripotency assays, and review of gross morphology of the line in both MEF feeder and feeder-free culture conditions.

Next, I performed a series of iPSC-HSPC differentiations utilizing a protocol I learned from my collaborators Drs. Deborah French and Paul Gadue during an external rotation at the University of Pennsylvania (Mills et al., 2014). My transcriptomic-level data from these differentiations demonstrated that the HSPCs expressed a transcriptomic program that would be expected for *bona fide* HSPCs. The HSPCs also co-expressed

CD41 and CD235 as would be expected for HSPCs.

Most importantly, my RNA-seq studies suggest that hyperactivation of the MAPK/ERK1/2 pathways is a constitutive event in *ANKRD26*-mutated iPSCs and HSPCs, similar to what Bluteau *et al.* observed in patient-derived megakaryocytes and platelets (Bluteau et al., 2014). This suggests that dysregulation of the MAPK pathways may potentially serve as the link in the thrombocytopenic, leukemogenic phenotype while also offering a potential path forward for future translational studies.

Previously, I had hypothesized that the HT/HHM phenocopies driven by *ANKRD26* or *RUNX1* ultimately stemmed from their molecular interaction, in which the expression of *ANKRD26* is directly regulated by RUNX1. The data provided in this chapter provide some evidence that, on a transcriptomic level, these two syndromes share a similar pattern of MAPK/ERK1/ERK2 activation. For example, although I did not investigate MAPK/ERK1/ERK2 signaling directly in a *RUNX1*-mutated iPSC line, prior work from Niimi *et al.* demonstrated that somatic *RUNX1* mutations in MDS result in an increase in MAPK-related signaling activation, particularly in the RTK-RAS portion of the MAPK pathways. Similarly, Niimi *et al.* demonstrated that *TP53* and *RUNX1* somatic mutations were mutually exclusive in MDS carriers, just as somatic *TP53* mutations were uncommon in my cohort of *RUNX1* germline mutation carriers described in Chapter II of this thesis (Niimi et al., 2006). No *TP53*-driven CH or malignancies have been discovered in my cohort of *ANKRD26* germline mutation carriers, although the size of my *ANKRD26* cohort is relatively small.

For the next steps in this project, I will be performing protein-level experiments to determine if the increased expression of genes involved in MAPK/ERK1/2 signaling is

manifested in increased protein translation, including of the MAPK-related proteins ERK, phospho-ERK, AKT, phospho-AKT, STAT5, and phospho-STAT5. I will also determine if p53 protein levels are increased in *ANKRD26*-mutated iPSCs relative to controls.

In the upcoming months, I will be performing additional experiments in order to better define the cellular phenotype of *ANKRD26*$^{+/mut}$ iPSCs, iPSC-HPSCs, iPSC-myeloid precursors, and iPSC-megakaryocytes. These studies were unfortunately halted secondary to the COVID19 pandemic, but I have previously generated iPSC-megakaryocytes during my external rotation at the University of Pennsylvania. These experiments, in addition to Methocult and Megacult-based colony forming assays, will allow me to better determine if *ANKRD26* 5' UTR mutations lead to downstream deficiencies in HSPC-derived cellular lineages. Ultimately, the presence of a cellular phenotype, either via liquid or semi-solid colony lineage assays, will allow me to more readily perform highly translational studies involving drug screens of established drug libraries as well as with selected MEK inhibitors that have been utilized in both pre-clinical and clinical studies (Braicu et al., 2019; Lee et al., 2020). If my hypothesis that increased MAPK signaling is a fundamental downstream effect of *ANKRD26* mutations that drives both the thrombocytopenia and leukemogenesis that is observed in mutation carriers, MAPK-inhibition may potentially serve a role in reversing the thrombocytopenic phenotype and reducing leukemogenic risk.

Chapter V

Germline *TUBB1* mutations are a potential novel hereditary thrombocytopenia and hereditary hematopoietic malignancy syndrome

Introduction

Currently, germline mutations in *ANKRD26*, *ETV6*, or *RUNX1* cause a distinct phenotype consisting of lifelong thrombocytopenia, platelet dysfunction, and an increased lifetime risk of hematopoietic malignancies. At least a dozen genes have been demonstrated to increase a mutation carrier's lifetime risk for HHMs such as MDS, AML, and other blood cancers (**Table 3.1**). However, Guidugli *et al.* demonstrated that the majority (79%) of individuals with an HT/HHM phenotype who undergo clinical grade genetic testing will not ultimately have a germline genetic driver identified (Guidugli et al., 2017). These data suggest that additional germline HT/HHM-related mutations remain to be discovered.

Therefore, I performed a pedigree review of all families enrolled on the 11-0014 protocol at the University of Chicago. I then identified pedigrees notable for a family history of both thrombocytopenia and HMs. I reviewed clinical-grade sequencing data generated by Clinical Laboratory Improvement Amendments (CLIA)-certified laboratories when these data were available. Any HT/HHM phenotype families who had been assessed via CLIA-grade testing and were negative for known germline driver mutations were then moved to a research-based sequencing pipeline. I then performed unbiased whole exome and/or whole genome sequencing (WES/WGS) on samples from these families to determine if I could identify novel germline genetic drivers of HT/HHMs. This

approach was in keeping with the approach described in **Figure 1.5** of this thesis. I hypothesized that this approach would identify candidate germline variants that may represent novel HT/HHM syndromes.

Methods

Patient tissue acquisition

All patients from the University of Chicago (**Figure 5.1**) provided informed consent for the 11-0014 protocol. This protocol allows for germline sequencing as well as the establishment of permanent cell lines such as iPSCs. Each patient provided a saliva sample using the Oragene Sample kit or a peripheral blood sample. I extracted DNA from peripheral blood samples using the QIAamp DNA Blood Mini Kit (Qiagen) per the manufacturer's instructions. I then utilized a Qubit fluorometer and reagents (Life Technologies) to perform DNA quantitation.

Whole exome sequencing

Genomic DNA was submitted to the Human Genetics Research Laboratory at the University of Chicago for processing and sequencing. Exome sequencing was performed with the Agilent SureSelect Clinical Research Exome kit (Agilent Technologies), and sequencing was performed with the Illumina NextSeq and 150 bp paired-end reads (Illumina). Sequencing data were stored on a protected high-performance computing system at the University of Chicago that exceeds requirements for the Health Insurance Portability and Accountability Act. Variants were then called and analyzed via a GATK-based bioinformatics pipeline at the University of Chicago (McKenna et al., 2010). I

assessed the data for quality-related metrics, including minimum coverage (30x or deeper for at least 90% of targeted regions), with a mean depth of coverage of greater than 150x for each sample. I then confirmed the presence of any potentially pathogenic variants via Sanger sequencing.

Lymphoblastoid cell line establishment

I utilized a protocol from the group of Yoav Gilad to establish LCLs via transformation with EBV. After extracting lymphocytes from peripheral blood samples that were collected in tubes containing acid citrate dextrose, I then added previously thawed EBV and phytohemagglutinin to the lymphocytes in a 4.5 mL suspension of RPMI 1640 media. I left these cells undisturbed for 72-96 hours in a 37° degree Celsius, 5% CO2 incubator. I subsequently passaged the cells as needed based on media color changes, performing each passage when the RPMI media turned yellow (which represented an acidic pH).

Induced pluripotent stem cell line establishment

In order to establish iPSCs from my LCLs, I utilized a protocol from the group of Yoav Gilad. Two days prior to nucleofection, I split my LCLs into two T25 flasks. Each T25 contained 5-10 mLs of RPMI and LCLs at a density of 0.35 x 10^6/mL. On the day of the nucleofection, I performed cell counts and viability assays to ensure that the LCLs had reached a concentration between 0.8 and 1.2 x 10^6/ml and a viability of at least 65%. I then performed nucleofection using episomal plasmids that contained *KLF4*, *L-MYC*, *LIN28*, *OCT3/4*, *SOX2*, and an shRNA against *p53* (Addgene plasmids 27077, 27078,

27080, and 27082). I maintained the ensuing colonies on a murine embryonic fibroblast (MEF)-based feeder system for approximately 30 days until individual colonies appeared. I transferred 6-12 colonies to one well of a six well plate coated with MEFs. I maintained these colonies in hESC media (DMEM/F12, Corning) supplemented with 20% KOSR (LifeTechnologies), 0.1 mM non-essential amino acids (NEAA), 2 mM GlutaMax, 1% Pen/Strep, 0.1% 2-mercaptoethanol (LifeTechnologies), and 100 ng/mL human basic fibroblast growth factor. I then manually cut and passaged each colony for at least 10 passages on MEFs. After 10 passages of growth on MEFs I transitioned each colony to feeder free conditions on 0.01 mg/cm2 hESC-grade Matrigel (BD Sciences) and mTeSR media (StemCell) supplemented with 1% Pen/Strep. I maintained the iPSCs in feeder-free conditions for at least five passages before performing quality control analyses. Passaging in feeder-free conditions was performed using DPBS containing 0.5 mM EDTA. The gross morphology of both the feeder-based and feeder-free colonies was typical for iPSCs in both growth conditions.

Statistical analyses

I calculated the number of individuals with predicted loss of function *TUBB1* mutations from two cohorts of patients at the University of Chicago and the University of Helsinki who had whole exome sequencing performed for the purposes of syndrome discovery (n = 558). I obtained ExAC data for the European (non-Finnish) and Ashkenazi Jewish populations, then I performed two-sided exact binomial tests to compare the frequencies of germline *TUBB1* mutations in my patient cohort as compared to the noncancer ExAC European (non-Finnish) population. I also performed a sensitivity

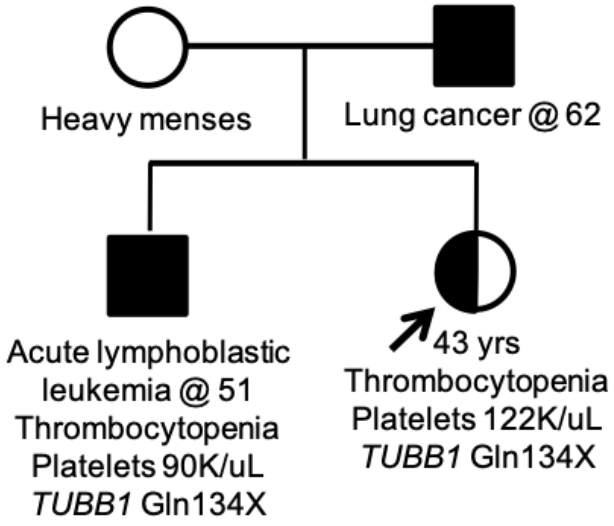
analysis using the ExAC Ashkenazi Jewish population so as to account for the population structure within individuals with Jewish ancestry. I considered *P* values < 0.05 to be significant.

Results

Identification of a proband with a personal history of thrombocytopenia, a family history of acute lymphoblastic leukemia, and a germline truncating TUBB1 *variant*

I identified a rare (gnomAD frequency 2.9 x 10^{-4}) truncating *TUBB1* variant in an Ashkenazi Jewish family (**Figure 5.1**). This variant is more common in individuals with Ashkenazi Jewish ancestry (gnomAD frequency, Ashkenazi Jewish: 6.4 x 10^{-3}). I first identified the variant in the proband, who had a personal history of thrombocytopenia, with a platelet count of 122 000/μL, as well as a family history of a brother who developed ALL at the age of 51. I sequenced germline tissue (cultured skin fibroblasts) from the brother, who had been treated previously with a matched unrelated donor allogeneic hematopoietic stem and progenitor cell transplant. My sequencing work subsequently confirmed that the *TUBB1* p.Gln134X variant was also present in the brother.

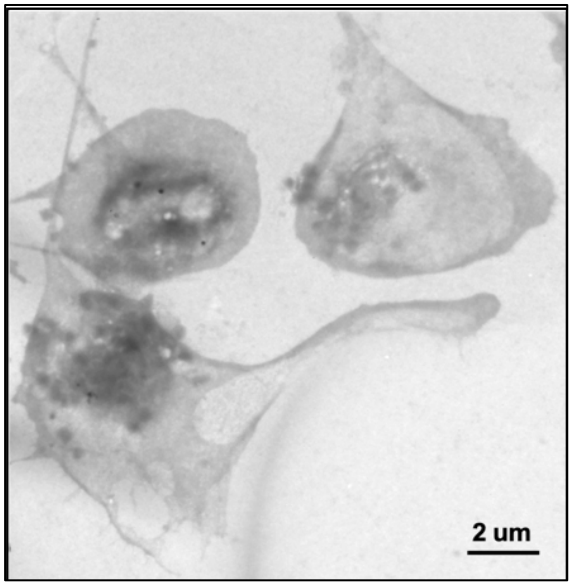
Figure 5.1. Pedigree from UChicago family in which a truncating *TUBB1* mutation (p.Gln134X) was identified. The variant was present in both the proband, who had thrombocytopenia, as well as her brother, who had acute lymphoblastic leukemia.



Platelet aggregation and microscopy studies

I then performed a series of coagulation and microscopy studies on samples taken from the proband in **Figure 5.1** in collaboration with Dr. Geoffrey Wool from the Department of Pathology at the University of Chicago. These studies included platelet lumi-aggregation, peripheral blood cell smear review, and platelet transmission electron microscopy. The platelet lumi-aggregation studies demonstrated that there were no overt abnormalities of platelet aggregation or secretion, albeit in the setting of mild thrombocytopenia which may slightly reduce the sensitivity of this assay. The platelet morphology was notable for anisocytosis, including the presence of giant platelets. The platelet transmission electron microscopy studies were notable for normal appearing platelets that were of a normal size and contained a normal quantity of dense granules per platelet. **Figure 5.2** contains representative images from the platelet transmission electron microscopy studies.

Figure 5.2. Platelet transmission electron microscopy peformed on platelets collected from the proband in Figure 5.1, who carried a truncating *TUBB1* mutation (p.Gln134X). The platelets were normal in size and contained a normal number of dense granules per platelet. The dense granules were normal in size.



Generation of lymphoblastoid cell lines and patient-derived iPSCs containing the TUBB1 *p.Gln134X variant*

I then generated LCLs from peripheral blood lymphocytes that I collected from the proband in **Figure 5.1**. Following the successful generation of LCLs, I then re-programmed these LCLs into iPSCs using the approach detailed in the Methods section of this Chapter. Unfortunately, I was unable to perform quality control analyses using this iPSC line prior to the COVID-related shutdowns that occurred in the spring of 2020.

The University of Chicago and University of Helsinki patient hematopoietic malignancy and thrombocytopenia cohorts are not enriched for nonsense variants in TUBB1

The thrombocytopenia and blood cancer syndrome discovery cohorts at the University of Chicago and Helsinki were not enriched for *TUBB1* nonsense mutations as

compared to the noncancer ExAC European (non-Finnish) population (OR, 6.2; 95% CI, 0.9 – 44.6; *P* = 0.07). This variant was also not enriched when compared to the noncancer ExAC Ashkenazi Jewish population (OR, 0.3; 0.04 – 2.0; *P* = 0.2) These data are shown in **Table 5.1**.

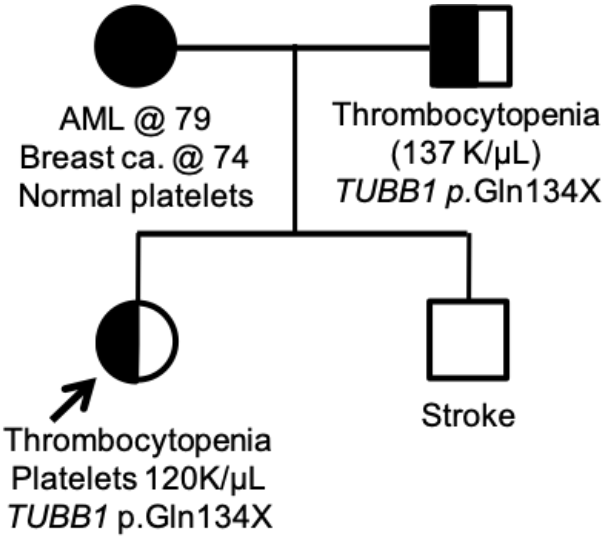
Table 5.1. Germline *TUBB1* mutation frequencies in patients with hereditary hematopoietic malignancies and/or hereditary thrombocytopenia as compared to a noncancer population and a noncancer Ashkenazi Jewish population.

	University of Chicago and University of Helsinki Patient Cohorts (n = 558)		ExAC NonCancer Population (All)			ExAC Noncancer Population (Ashkenazi Jewish)						
	Number of mutated alleles	Proportion of Individuals with a Mutation	Number of mutated alleles	Mean number of individuals sequenced	Proportion with a mutation	Number of mutated alleles	Mean number of individuals sequenced	Proportion with a mutation	Universities of Chicago/Helsinki versus ExAC population, OR (95% CI)	*P*	Universities of Chicago/Helsinki versus ExAC Ashkenazi Jewish population, OR (95% CI)	*P*
TUBB1	1	0.003584229	82	282846	0.0002899	66	10362	0.006369427	6.2 (0.9 - 44.6)	0.07	0.3 (0.04 - 2.0)	0.2

Identification of a second proband with a personal history of thrombocytopenia, a family history of acute lymphoblastic leukemia, and a germline truncating TUBB1 *variant*

After I identified the family shown in Figure 5.1, I reviewed the sequencing data previously generated from patients enrolled to the 11-0014 protocol and identified an additional pedigree in which the p.Gln134X variant was present. This pedigree was notable for a proband with mild thrombocytopenia (120 K/μL) as well as a family history of thrombocytopenia in her father (137 K/μL) and AML and breast cancer in her mother, diagnosed at 79 and 74 years of age, respectively. The proband's mother had normal platelet counts. After performing segregation studies in the family, I confirmed that the father with thrombocytopenia carried the *TUBB1* p.Gln134X variant, but the mother with AML and breast cancer did not carry this *TUBB1* variant.

Figure 5.3. Pedigree from UChicago family in which a truncating *TUBB1* mutation (p.Gln134X) was identified. The variant was present in both the proband, who had thrombocytopenia, as well as her father, who had thrombocytopenia.



Discussion

This is the second report of a pedigree in which a carrier of a germline *TUBB1* predicted deleterious mutation subsequently developed both HT and HMs. The first report, from Matsumura *et al*., detailed a *TUBB1* p.T149P germline variant in a pedigree in which the variant of interest segregated with both thrombocytopenia and myeloid malignancies (Matsumura et al., 2019). Notably, the p.Thr149Pro germline variant is not listed in gnomAD, ExAC, or ClinVar and was 100% penetrant for myeloid malignancies in the Matusumura report. However, the p.Gln134X variant was only 25% penetrant for HMs in the pedigrees I described from the University of Chicago. The p.Gln134X variant was 100% penetrant for thrombocytopenia in the mutation carriers identified in the University of Chicago pedigrees. Both the p.T149P and p.Gln134X variants are located in exon 4 of the TUBB1 protein, which contains residues 93 through 452 and is the mutational hotspot

for all but two known LP/P germline ClinVar variants in *TUBB1*. Exon 4 is the mutational hotspot for all known LP/P germline missense variants in ClinVar.

Furthermore, the p.Gln134X mutation was not associated with HMs of any kind as compared to normal, non-cancer ExAC controls or normal, non-cancer Ashkenazi Jewish ExAC controls. As a sensitivity analysis, I would have needed to identify one additional mutation carrier with an HM in order to achieve statistical significance as compared to the normal, non-cancer Ashkenazi Jewish population. I would have needed to identify two additional mutation carriers with an HM in order to achieve statistical significance as compared to the normal, non-cancer ExAC population.

At this time, it appears that the *TUBB1* germline p.Gln134X mutation is associated with hereditary thrombocytopenia, but should not be considered a mutation that is associated with HT/HHMs. Future directions with this project will involve utilizing the patient-derived *TUBB1* iPSCs as a contemporary cellular model for the study of hereditary thrombocytopenia syndromes and tubulin disorders.

Chapter VI

Germline *TTLL6* mutations are associated with an increased risk for hereditary hematopoietic malignancies

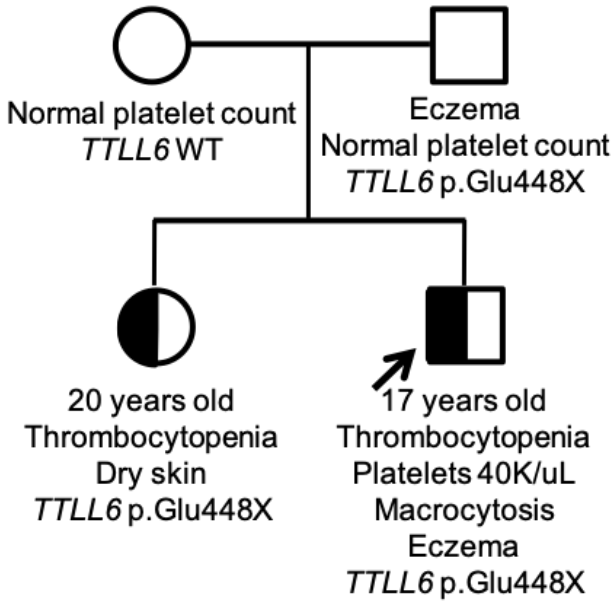
Introduction

Currently, germline mutations in at least a dozen genes have been demonstrated to increase a mutation carrier's lifetime risk for HHMs such as MDS, AML, and other blood cancers. However, the majority (79%) of individuals with an HHM-type phenotype will not have a germline driver mutation identified when contemporary next-generation sequencing (NGS)-based sequencing approaches are used to sequence the known HHM-related genes. This problem suggests that additional HHM-related genetic drivers remain to be identified. However, most single research groups do not have access to a sufficient number of families with HHMs to power a sufficiently sized study for the identification of HHMs at a single institution using patient samples alone.

Therefore, a number of HHMs remain to be discovered. In order to address this knowledge gap in the field, I reviewed all of the pedigrees enrolled on the IRB-approved 11-0014 research protocol at the University of Chicago in order to identify pedigrees that were notable for a phenotype of hereditary thrombocytopenia (HT), HHM, or an overlapping HT/HHM phenotype. This protocol is one of the largest of its kind in the world, with more than 3,000 patients enrolled to date. This sample size addresses the limitation of small sample sizes that limits many research groups performing syndrome-based discovery. I then reviewed the CLIA-certified laboratory data that was available for these families. Since I have previously demonstrated that many contemporary NGS panels are

insufficiently designed for the accurate diagnosis of HHMs, I then reviewed the quality of clinical-grade sequencing that was performed for each family (Roloff et al., 2020). Any individuals who were sequenced with inadequate NGS-based panels were re-sequenced using an NGS panel that was deemed appropriate for their phenotype. Any patients with an HT or HT/HHM phenotype were then moved to a research-based sequencing pipeline and underwent whole exome and/or whole genome sequencing (WES/WGS). This approach was in keeping with the approach described in **Figure 1.5** of this thesis.

Figure 6.1. Pedigree from UChicago family in which a truncating *TTLL6* mutation (p.Glu488X) was identified. The variant segregated with an eczema-like phenotype and was present in both pediatric individuals with thrombocytopenia.



Using this approach, I identified a family (**Figure 6.1**) that had been diagnosed with early onset hereditary thrombocytopenia in two teenage children. Interestingly, the family history also was notable for eczema in both children and the father. Eczema has been shown to be present in many individuals with HT/HHMs driven by germline *RUNX1* mutations (Bellissimo et al., 2020; Sorrell et al., 2012). However, this family did not carry

any germline variants in known HT/HHM-related genes. Therefore, I collected germline DNA from each individual in the pedigree and performed WES to evaluate this family for the presence of a known driver mutation. Ultimately, this analysis demonstrated the presence of a rare heterozygous truncating variant in *TTLL6* (gnomAD minor allele frequency $2.2x10^{-5}$) that segregated with the eczema-like phenotype in the father and both children. No candidate variants adhering to an autosomal recessive inheritance pattern were identified.

Table 6.1. Clinical features of individuals in study families. AML: acute myeloid leukemia.

Patient	Family	Relationship	Sex	Hematopoietic Diagnosis	Karyotype	Somatic Mutations	Age at Diagnosis	Other Tumors	Family history	Germline Variant	gnomAD frequency	gnomAD Finnish frequency
H_1.1	H_1	Father	M	AML	Hyperdiploid, trisomy 8, inv(16)	NA	66	Melanoma (58 years)	AML (son)	p.Gln94X	$5.25x10^{-5}$	$4.9x10^{-4}$
H_1.2	H_1	Son	M	AML	Normal	*DNMT3A, IDH2, NPM1*	34	NA	AML (father)	p.Gln94X	$5.25x10^{-5}$	$4.9x10^{-4}$
H_2.1	H_2	Proband	F	Multiple myeloma	del(17p)	NA	61	NA	Rhabdomyosarcoma (grandchild)	p.Gln94X	$5.25x10^{-5}$	$4.9x10^{-4}$
UC_1.1	UC_1	Brother	M	Thrombocytopenia, macrocytosis	NA	NA	17	NA	Thrombocytopena (sister)	p.Glu488X	$2.2x10^{-5}$	Absent
UC_1.2	UC_1	Sister	F	Thrombocytopenia	NA	NA	20	NA	Thrombocytopena (brother)	p.Glu488X	$2.2x10^{-5}$	Absent

After identifying this *TTLL6* mutation in this family, I then communicated with my colleague Dr. Ulla Wartiovaara-Kautto at the University of Helsinki, who has sequenced a large number of families with HT/HHM phenotypes. Intriguingly, Dr. Wartiovaara-Kautto had also identified three individuals in two unrelated pedigrees with HMs who carry a rare germline *TTLL6* truncating mutation (p.Gln94X, gnomAD minor allele frequency of $5.25x10^{-5}$, gnomAD Finnish MAF of $4.9x10^{-4}$) segregated with the diagnosis of AML (**Table 6.1**). These pedigrees were notable for a father and son, both carriers of the *TTLL6* variant, who both developed AML at the ages of 66 and 34 years, respectively. A third individual with the same *TTLL6* truncating variant developed multiple myeloma at the age of 61.

The tubulin tyrosine ligase-like (TTLL) enzymes are involved in glutamylation, a process by which glutamate chains are added to the target proteins at glutamate residues, as well as the maintenance of microtubule mass (Garnham et al., 2015; Rogowski et al., 2010). *TTLL6* encodes for TTLL6, a tubulin polyglutamylase that is involved in the polyglutamylation of C-terminal tails in tubulin. This process leads to microtubule severing and the maintenance of microtubule mass. The TTLL class of enzymes are counteracted by a class of enzymes, the cytosolic carboxypeptidases (CCPs), which act to remove the glutamates that are added by TTLL enzymes (Lacroix et al., 2010; Wu & Li, 2020; Ye et al., 2014). Although *Ttll6*-deficient mouse models have not been characterized previously, a homozygous *Ccp6* knockout mouse model develops a threefold increase in megakaryocytes, a threefold increase in platelets, splenomegaly, dysfunctional platelets, and defective hemostasis (Ye et al., 2014). This phenotype is on the opposite end of the hematopoietic “extreme” of the family I identified in the initial pedigree, which was notable for thrombocytopenia (low platelet count). Given the antagonistic roles of CCP6 and TTLL6 in megakaryopoiesis, this further raised my suspicion that germline *TTLL6* mutations may lead to defective megakaryopoiesis.

My initial suspicion for a link between germline *TTLL6* mutations and HT and/or HHM syndromes was also strengthened by the known role of a number of other HT or HT/HHM-related genes in cytoskeletal function. For example, *ANKRD26* germline mutations are the most common cause of the HT/HHM phenotype, and the ANKRD26 protein has an established role in cytoskeletal function, specifically centrosome maintenance and amplification (Downes et al., 2019; Evans et al., 2021). Similarly, the tubulin, beta 1 (TUBB1) protein is one of two core proteins that heterodimerize to form

microtubulin, and germline *TUBB1* mutations are a long-recognized cause of HT (Balduini & Savoia, 2012; Kunishima et al., 2009). Intriguingly, there is growing evidence that germline *TUBB1* mutations may also drive the HT/HHM phenotype, further strengthening the link between disrupted cytoskeletal maintenance and the HT/HHM phenotype (Matsumura et al., 2019).

In addition to its role in microtubule mass maintenance, *TTLL6* has also been identified as a member of the STING pathway, which has increasingly been implicated in the pathogenesis of HHMs (Roloff GW, 2021.). The STING pathway is involved in interferon production, viral immune response, and innate immunity. The cGAS molecule is partly responsible for modulating the immune response within the STING pathway. cGAS is a recognized target of TTLL6, which glutamylates cGAS, reducing the ability of cGAS to bind DNA. A CCP-family enzyme, CCP6, reverses this process by removing the glutamate chain from cGAS and restoring its ability to bind DNA (Xia et al., 2016; Ye et al., 2014). As detailed above, homozygous *Ccp6* knockout mice develop an essential thrombocythemia-like blood malignancy, with elevated megakaryocyte counts, elevated platelet counts, and splenomegaly (Ye et al., 2014).

Similarly, the most common HHM is driven by germline mutations in *DDX41*, which encodes for the DEAD-box helicase 41 (DDX41), a member of the STING pathway. DDX41, like cGAS, is involved in the sensing of cytosolic double stranded DNA from viral and bacterial pathogens. DDX41 knockdown results in decreased activation of TANK-binding kinase 1 (TBK1, a member of the MAPK pathways), diminished production of type 1 interferon, and decreased cytokine responses to viral pathogens(Liu & Wang, 2016; Zhang et al., 2011). Germline *DDX41* mutations are responsible for 1% of all seemingly

"sporadic" cases of AML (Lewinsohn et al., 2015; Polprasert et al., 2015). The population prevalence of HHM-associated germline mutations in *DDX41* (gnomAD frequency ranging from absent up to 1.0 x 10^{-4}) is similar to that of the *TTLL6* mutations observed in my patients (2.2x10^{-5}) and those of my colleague Dr. Wartiovaara-Kautto (1.1x10^{-4}). Although still rare, the frequency of these variants is relatively high compared to other HHM syndromes, in which the variants are nearly uniformly absent from population databases. The median age of leukemogenesis in *DDX41* mutation carriers is in the seventh decade of life (Lewinsohn et al., 2015; Polprasert et al., 2015). The relatively increased population frequency of *DDX41* germline mutations is likely a product of the relatively late onset of the *DDX41* phenotype and the minimal impact of late-onset cancer on the reproductive fitness of mutation carriers. Interestingly, no pediatric cases of malignancies in germline *TTLL6* mutation carriers have been diagnosed to date despite Dr. Wartiovaara-Kautto's patient registry containing a large number of pediatric patients.

Given its important role in cytoskeletal function, tubulin maintenance, and innate immunity via the STING pathway, I proceeded to examine the potential role for germline *TTLL6* mutations in HT and HHMs further. I hypothesized that germline *TTLL6* mutations would be an HT/HHM phenocopy leading to decreased platelet counts and increased leukemogenic risk in a constitutional murine knockout model.

Methods

Patient tissue acquisition

All patients from the University of Chicago (**Figure 5.1**) provided informed consent for the 11-0014 protocol allowing for germline sequencing. I collected saliva samples from each of the individuals in the pedigree via the Oragene sample kit, then extracted DNA using the QIAamp DNA Blood Mini Kit (Qiagen) per the manufacturer's instructions. I then utilized a Qubit fluorometer and reagents (Life Technologies) to perform DNA quantitation.

Whole exome sequencing

Genomic DNA was submitted to the Human Genetics Research Laboratory at the University of Chicago for processing and sequencing. Exome sequencing was performed with the Agilent SureSelect Clinical Research Exome kit (Agilent Technologies), and sequencing was performed with the Illumina NextSeq and 150 bp paired-end reads (Illumina). Sequencing data were stored on a protected high-performance computing system at the University of Chicago that exceeds requirements for the Health Insurance Portability and Accountability Act. Variants were then called and analyzed via a GATK-based bioinformatics pipeline at the University of Chicago (McKenna et al., 2010). I then assessed the data for quality-related metrics, including minimum coverage (30x or deeper for at least 90% of targeted regions), with a mean depth of coverage of greater than 150x for each sample. I then confirmed the presence of any potentially pathogenic variants via Sanger sequencing.

Analysis of UK Biobank Data

I utilized the genebass (gene-biobank association summary statistics) browser to query the UK Biobank and to perform gene- and variant-based association tests between 281,850 exomes and 3,700 phenotypic traits. Using genebass, I analyzed rare variant associations between predicted loss of function (pLoF) and missense variants in *TTLL6* and the full range of phenotypic traits assessed in the UK Biobank dataset (Karczewski et al., 2021). In addition, I queried pLoF and missense variants in known HT/HHM-related genes (*ETV6*, *RUNX1*) as well as syndromes with well-established roles in hereditary thrombocytopenia but without known leukemogenic risk (*GP1BA*, *TUBB1*) as positive controls for hematopoietic-related phenotypes in UK Biobank data. I utilized a genome-wide p-value threshold of $2.5x10^{-8}$ for SKAT-O tests, $8x10^{-9}$ for single variant tests, and $6.7x10^{-7}$ for burden tests, in keeping with parameters established by the genebass authors.

Murine model of Ttll6 deficiency

In order to model the effects of *Ttll6* deficiency in a murine model, I obtained a pair of heterozygous constitutional *Ttll6* knockout (KO) mice generated on a C57BL/6 background from the NIH Knockout Mouse Project (KOMP) (MGI ID 5605489, $Ttll6^{tm1.1(KOMP)Vlcg}$). This strain was created at Regeneron Pharmaceuticals, Inc. from the ES cell clone 17697A-A5 and obtained by the KOMP repository. This murine strain was generated via the insertion of a Velocigene definitive null cassette containing a LacZ reporter and a loxP-flanked neomycin cassette prior to flp-mediated excision of both the LacZ and neo from the locus in order to create a 20097 bp deletion in *Ttll6* between

chr11:96000174-96020270, genome build 37 (Valenzuela et al., 2003). I then confirmed that targeted *Ttll6* deletion occurred vis PCR primers that were specific to the WT allele (F: TATCCTGCACATCTGGCTTG; R: TGGGCAGTGCTCTCTATGTTC) or the post-Cre Regeneron cassette (F: ACTTGCTTTAAAAAACCTCCCACA; R: ACTCTGAAAGGGTGCTCCTTATGGG).

I then utilized an anti-TTLL6 antibody (Atlas Antibodies, product number HPA052397, 1:1000 dilution) for immunohistochemical (IHC) assays using dissected, formalin-fixed spleen tissue collected from WT and homozygous knockout (Hom KO) animals. The tissues were processed and stained at the Human Tissue Resource Center at the University of Chicago. Images were taken at 4x, 10x, and 20x magnification and analyzed via the Leica ImageScope software.

Expansion and monitoring of Ttll6 deficient murine colony

After confirming the presence of the post-Cre Regeneron cassette via PCR for genotyping purposes, I then expanded my murine colony of *Ttll6*-deficient mice via serial crossing of *Ttll6*$^{+/-}$ mice. This breeding approach produced heterozygous knockout (Het KO) and homozygous knockout mice (Hom KO) in expected Mendelian ratios. I housed all mice in the University of Chicago's pathogen-free barrier facility under a protocol approved by the University of Chicago Institutional Animal Care and Use Committee. This facility is accredited by the Association for Assessment and Accreditation of Laboratory Animal Care.

The mice were monitored daily for the development of peripheral and abdominal tumors via palpation and for mediastinal tumors via scruffing of the neck. The cadence of

these assessments, however, decreased during COVID19-related work restrictions beginning in the spring of 2020. I assessed the mice for symptoms suggestive of systemic disease via monitoring for hunched posture and/or ruffled fur.

A cohort of mice had serial bleeds performed via submandibular bleed with the plan to perform monthly bleeds; however, this bleed schedule was also disrupted nearly entirely by the COVID19 pandemic. All peripheral blood samples were used to measure complete blood counts (CBCs). I performed a full necropsy for any mouse with evidence of tumorigenesis and/or systemic illness. Each necropsy included a terminal bleed (for CBCs and peripheral smears), gross examination of all major organs (spleen, liver, lymph nodes, thymus), and histologic/morphologic examination of all major organs in addition to sternal bone marrow. I biobanked live cells from bone marrow tissue, spleen, liver, and tumor tissue (when applicable). I performed histopathology and flow cytometry (using single cell tissues) on all tissues collected via necropsy. I performed all CBC measurements using a Hemavet 950 (CDC Technologies) and also made Wright-Giemsa stained peripheral blood smears for visual assessment. In collaboration with Stephen Arnovitz from the Godley group, we also transplanted malignant cells collected from CD45.2 donor mice into sub-lethally irradiated wild-type recipients (CD45.1). Stephen Arnovitz monitored these recipients for engraftment using flow cytometric analyses and assessments of peripheral blood counts. Histopathologic specimens collected during necropsy were fixed in 10% neutral-buffered formalin, embedded in paraffin, and then cut in 3 to 4 μm sections. I made touch preps from all examined organs which I stained with Wright-Giemsa stain prior to examination using an Olympus BX40 microscope (Olympus).

All CBCs were compared using the Student's *t*-test, and all survival statistics were plotted using survminer R data package, which employs the Kaplan-Meier method with statistical analyses performed using the Cox Proportional Hazards Model. Outcomes of interest for the Cox model were time to death, with the analysis stratified by genotype (*Ttll6*$^{+/+}$, *Ttll6*$^{+/-}$, or *Ttll6*$^{-/-}$). Statistical significance was considered to be a p-value < 0.05 for each analysis.

Statistical analyses

I calculated the number of individuals with nonsense *TTLL6* mutations from among two cohorts of patients at the University of Chicago and the University of Helsinki who had whole exome sequencing performed for the purposes of syndrome discovery (n = 558). I obtained ExAC data for the European (non-Finnish) and European (Finnish) populations, then I performed two-sided exact binomial tests to compare the frequencies of germline *TTLL6* mutations in my patient cohort as compared to the noncancer ExAC European (non-Finnish) population. I also performed a sensitivity analysis using the ExAC European (Finnish) population so as to account for the population structure within Finland. I considered p-values < 0.05 to be significant.

Results

The University of Chicago and University of Helsinki patient hematopoietic malignancy and thrombocytopenia cohorts are enriched for nonsense variants in TTLL6

The thrombocytopenia and blood cancer syndrome discovery cohorts at the University of Chicago and Helsinki were enriched for *TTLL6* nonsense mutations as

compared to the noncancer ExAC European (non-Finnish) population (OR, 155.6; 95% CI, 47.4 - 511.5; $P < 0.0001$). This enrichment remained statistically significant when compared to the noncancer ExAC European (Finnish) population (OR, 20.6; 7.0 – 60.6; $P < 0.0001$) These data are shown in **Table 6.2**.

Table 6.2. Germline *TTLL6* mutation frequencies in patients with hereditary hematopoietic malignancies and/or hereditary thrombocytopenia as compared to a noncancer population.

	University of Chicago and University of Helsinki Patient Cohorts (n = 558)		ExAC NonCancer Population (All)			ExAC Noncancer Population (Finnish)						
	Number of mutated alleles	Proportion of Individuals with a Mutation	Number of mutated alleles	Mean number of individuals sequenced	Proportion with a mutation	Number of mutated alleles	Mean number of individuals sequenced	Proportion with a mutation	Universities of Chicago/Helsinki versus ExAC population, OR (95% CI)	*P*	Universities of Chicago/Helsinki versus ExAC population, OR (95% CI)	*P*
TTLL6	5	0.008960573	6	103293	0.0000581	10	22839	0.000437848	155.6 (47.4 - 511.5)	< 0.0001	20.6 (7.0 - 60.6)	<0.0001

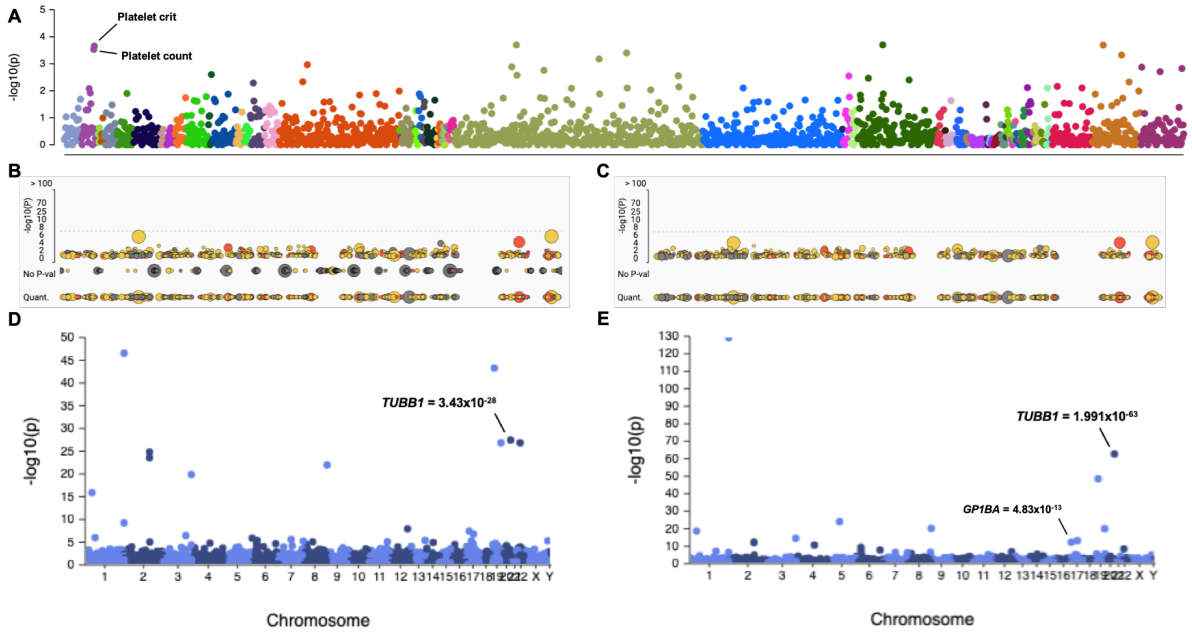
Predicted loss of function variants in TTLL6 are not associated with altered hematopoiesis in UK Biobank data

Germline pLoF and missense variants in *TTLL6* did not show any associations within the UK Biobank (**Figure 6.2**) using a genome-wide p-value threshold of 2.5×10^{-8} for SKAT-O tests, 8×10^{-9} for single variant tests, and 6.7×10^{-7} for burden tests. Of note, the platelet crit ($p = 2.21 \times 10^{-4}$) and platelet count ($p = 2.95 \times 10^{-4}$) traits were the 4th and 5th most highly ranked phenotypic traits, respectively, for *TTLL6* pLoF variants, but neither of these associations met genome-wide significance.

I then compared *TTLL6* data in the UK Biobank with genes with known roles in hereditary thrombocytopenia, *TUBB1* and *GP1BA* (**Figure 6.2**). This analysis demonstrated that *TUBB1* pLoF variants were associated with platelet crit ($p = 3.43 \times 10^{-28}$) and platelet count ($p = 1.991 \times 10^{-63}$) traits. *GP1BA* variants were associated with the platelet count trait ($p = 4.83 \times 10^{-13}$). I performed two additional analyses of HT/HHM syndromes driven by germline pLoF variants in *ETV6* or *RUNX1*. These analyses

demonstrated that *ETV6*- and *RUNX1*-driven HT/HHM syndromes do not show an association with hematopoiesis-related traits in the UK Biobank, falling well short of the p-value threshold of 2.5×10^{-8} for genome-wide significance. For example, *ETV6* pLoF variants have p-values of 1.64×10^{-3} for platelet count and 1.83×10^{-3} for platelet crit, whereas *RUNX1* pLoF variants have p-values of 2.45×10^{-2} and 3.26×10^{-2}. Neither *ETV6* nor *RUNX1* showed an association with hereditary cancer syndrome traits. Additionally, I performed analyses for the most common HHM-related gene, *DDX41*, and one of the most highly penetrant HHM-related syndromes, caused by *GATA2* mutations. Gene burden associations for both *DDX41* and *GATA2* did not detect any significant phenotypic associations involving hematopoietic traits and/or hereditary cancer traits.

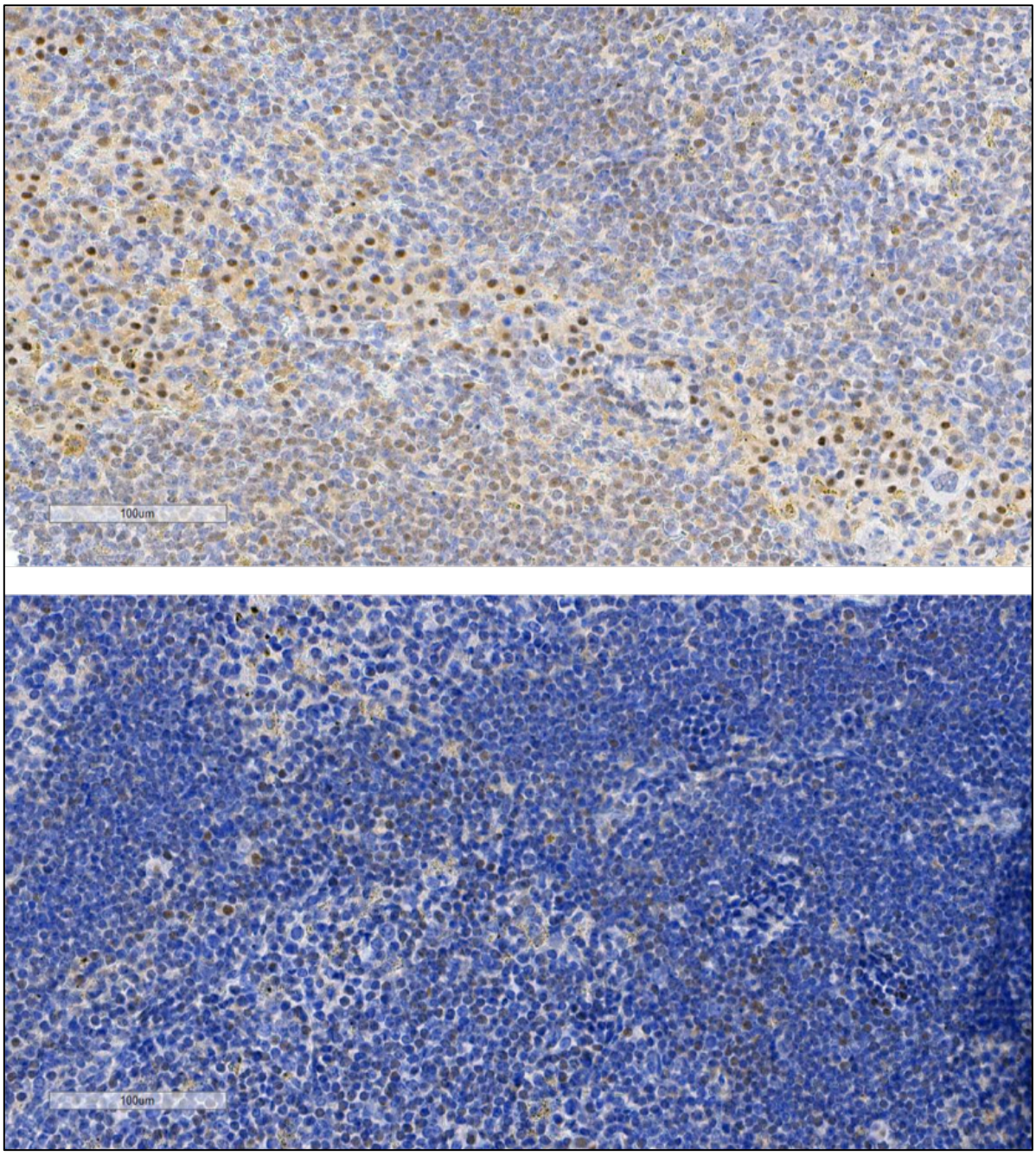
Figure 6.2. Panel A: 3,699 pLoF gene burden associations with *TTLL6*. Platelet crit and platelet count non-significant associations are highlighted. Panel B: *TTLL6* single variant associations with platelet crit; no individual variant met genome-wide significance (dashed horizontal line) for an association with platelet crit. Panel C: *TTLL6* single variant associations with platelet count; no individual variant met genome-wide significance (dashed horizontal line) for platelet count. Panel D: pLoF gene burden associations with platelet crit, with significant *TUBB1* association highlighted ($p = 3.43 \times 10^{-28}$) Panel E: pLoF gene burden associations with platelet count, with significant *TUBB1* ($p = 1.991 \times 10^{-63}$) and *GP1BA* ($p = 4.83 \times 10^{-13}$) associations highlighted.



TTLL6 deficiency leads to spontaneous tumorigenesis and decreased mortality in murine models but does not impact peripheral blood counts

I confirmed that mice with the post-Cre Regeneron cassette were deficient in TTLL6 protein expression via IHC (**Figure 6.3**).

Figure 6.3. Anti-TTLL6 staining of murine spleen tissue collected from a *Ttll6*$^{+/+}$ mouse (top) and a *Ttll6*$^{-/-}$ mouse (bottom) demonstrated absent TTLL6 expression in *Ttll6* deficient mice. 20x magnification.



There were no genotype-specific differences in peripheral blood counts at any age **(Figure 6.4**). Similarly, no sex-specific differences in peripheral blood counts were present at any age. However, my initial research plan to perform serial bleeds of the mice

on a monthly basis was disrupted by the COVID19 pandemic, resulting in a dramatically decreased number of bleeds over time per animal.

Despite the similar peripheral blood counts between each genotype, the *Ttll6*$^{-/-}$ animals spontaneously developed abdominal tumors beginning at 230 days of age, with the oldest animal being 685 days of age at the time of tumorigenesis. The *Ttll6*$^{+/-}$ animals developed tumors between 506 and 1041 days of age. No *Ttll6*$^{+/+}$ animals developed tumors that were appreciable on physical exam. Ultimately, this tumorigenesis led to a difference in overall survival that was driven by genotype, with *Ttll6-/-* animals living shorter than *Ttll6*$^{+/-}$ and *Ttll6*$^{-/-}$ animals ($p = 0.024$) (**Figure 6.5**). The tumors were predominantly located in the mesenteric lymph nodes, and one animal developed cervical lymphadenopathy. No differences in the peripheral blood counts of the diseased animals were noted compared to healthy animals that were sacrificed at similar time points. There were no genotype-specific differences in the mass of the animals' spleens or livers (**Figure 6.6**). There was also no difference in the mass of the spleens or livers after adjusting for total body mass.

At the time of this thesis submission, IHC, flow cytometric, and transplant studies were in process in order to complete the hematopathologic diagnosis of the tumors that had developed in the *Ttll6*$^{-/-}$ and *Ttll6*$^{+/-}$ animals. Preliminary review, however, of the tumor tissue by Dr. Jason Cheng at the University of Chicago suggests that these mice developed CD3+, CD19+ tumors with myelomonocytic expansion.

Figure 6.4. Peripheral blood counts, stratified by genotype (*Ttll6*$^{+/+}$, *Ttll6*$^{+/-}$,*Ttll6*$^{-/-}$). No significant differences were observed. No sex-dependent differences were observed. Notation: n = x,y,z; n = # of WT, # of Het KO, # of Hom KO at each time point. Black squares: WT; Blue circles: Het KO; Red circles: Hom KO.

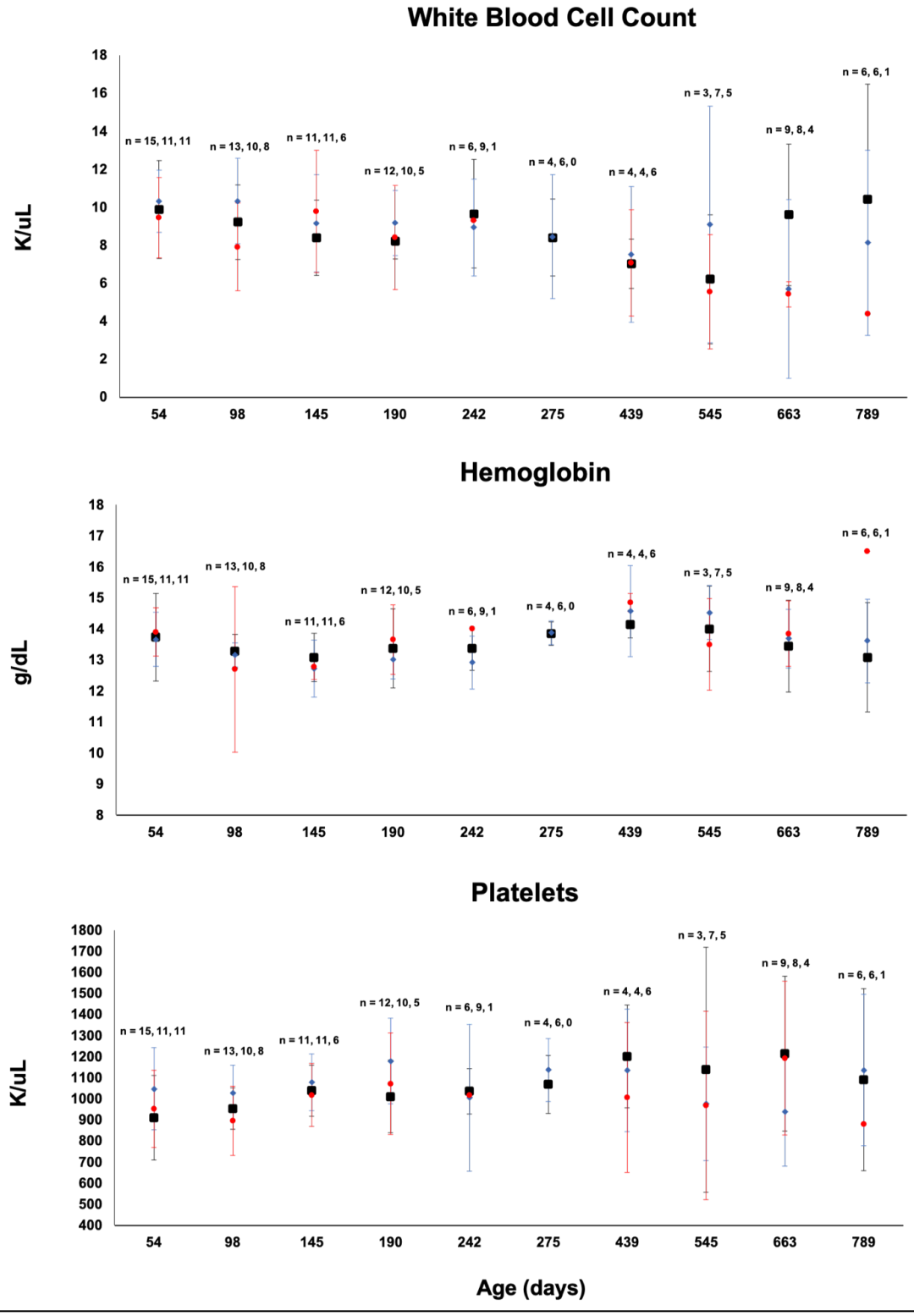


Figure 6.5. Kaplan Meier survival curve for *Ttll6*$^{+/+}$ (WT, n=3), *Ttll6*$^{+/-}$ (Het KO, n=11), and *Ttll6*$^{-/-}$ mice (Hom KO, n=4) mice showing a significantly shortened overall survival for the Hom KO cohort (p = 0.024).

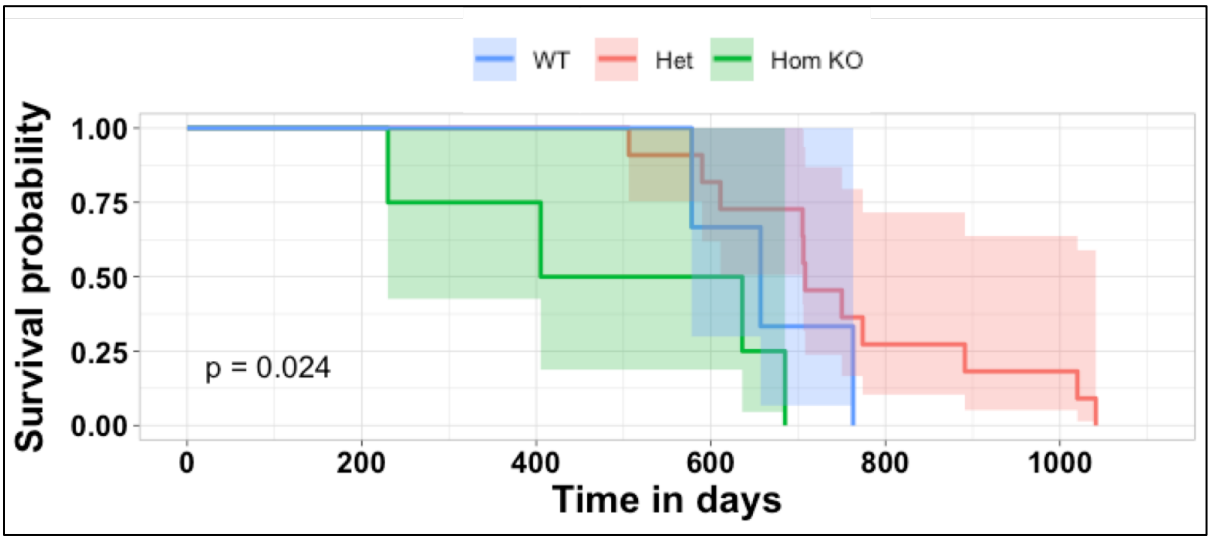
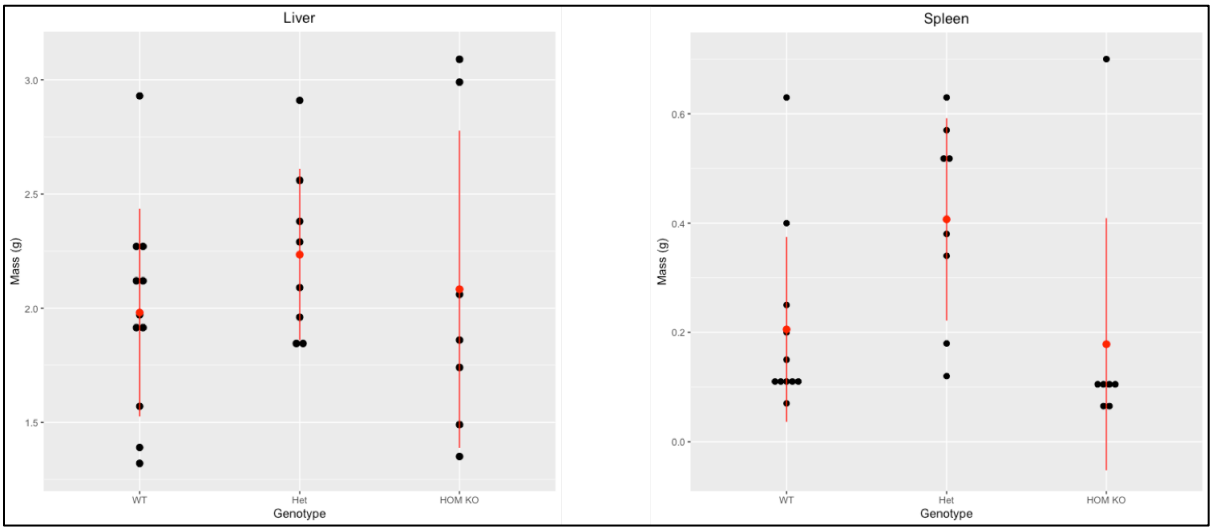


Figure 6.6. Organ weights in all animals based on genotype.



Discussion

In this study, I have presented the first evidence that ultra-rare germline mutations in *TTLL6* cause a phenotype that is notable for thrombocytopenia and increased leukemogenic risk, making this the fourth of the HT/HHM phenocopies (joining *ANKRD26*, *ETV6*, and *RUNX1* germline mutations). I first identified a pedigree with a hereditary

thrombocytopenia phenotype at the University of Chicago, and I then utilized whole exome sequencing to identify a p.Glu488X mutation in the two teenage children in this pedigree with early onset thrombocytopenia. The family was also notable for eczema, a phenotype that has been observed in other HT/HHM families, particularly those with germline *RUNX1* mutations. Eczema segregated with the p.Glu488X variant in the UChicago family. This phenotype suggests that dysregulated inflammatory response/autoimmunity is present in this novel syndrome driven by germline *TTLL6* mutations, as has been observed in germline *RUNX1* carriers (Bellissimo et al., 2020; Sorrell et al., 2012). This mechanism aligns with TTLL6's known role in suppressing the STING pathway, which itself is involved in the activation of the innate immune pathway (Xia et al., 2016). Ultimately, the p.Glu488X mutation was present in both children as well as a father who reportedly had a normal platelet count and as well as eczema. However, peripheral blood counts were not available from the father for review. Therefore, it is possible that his platelet count is "low-normal", which I have anecdotally seen in a variety of patients with HT/HHM-related mutations who had reportedly "normal" platelet counts. Alternatively, anticipation is a well-described phenomenon in the hereditary hematopoietic syndromes, and it is possible that germline *TTLL6* mutations are similarly affected by anticipation (Horwitz et al., 1996; Tegg et al., 2011).

My colleague at the University of Helsinki, Dr. Ulla Wartiovaara-Kautto, has also identified two additional pedigrees with the same ultra-rare *TTLL6* nonsense germline mutation (p.Gln94X) in her clinic. One of these pedigrees is notable for AML in a father and son, and the other is notable for multiple myeloma in a 61-year-old woman. Intriguingly, the original series of publications describing germline mutations in *DDX41*

were also notable for mutation carriers who developed multiple myeloma (Lewinsohn et al., 2016). This phenotype is striking given the role of both DDX41 and TTLL6 in the STING pathway.

Next, I demonstrated that constitutional *Ttll6* deficiency in murine models drives late-onset leukemogenesis, but not thrombocytopenia. In this model, the median age of tumorigenesis in *Ttll6*$^{-/-}$ was in excess of 400 days of life. The median age of tumorigenesis in *Ttll6*$^{+/-}$ animals was closer to 650 days of life, and no WT animals developed tumors. I am performing additional immunophenotyping studies to further refine the histopathologic subtype of these tumors, but preliminary histopathologic findings suggest these tumors are CD3+, CD19+ tumors with myelomonocytic expansion. Although this phenotype is discrepant from the more myeloid-type tumors observed in the human pedigrees from my patient cohorts at the University of Chicago and the University of Helsinki, *TTLL6* somatic variants in larger tumor databases appear to be more lymphoid-predominant. For example, 75.7% of somatic *TTLL6* variants in the COSMIC tumor database were identified in lymphoid malignancies, with only 24.3% identified in myeloid malignancies, a significant ($p < 0.00001$) skewing toward lymphoid malignancies (Tate et al., 2019). This skewing toward lymphoid malignancies in both a large tumor database as well as my murine model of *Ttll6* deficiency suggests that the more myeloid-heavy pedigrees that initially prompted this project may reflect the clinical selection bias by which I and my mentor, Dr. Godley, largely encounter our patients of interest – i.e., as clinical specialists in myeloid malignancies. Therefore, I am more likely to encounter families with hereditary myeloid malignancies.

However, it is also possible that murine models of HHMs are inherently more prone

to develop lymphoid malignancies even when the HHM in question is known to be more myeloid-skewed in humans. For example, the majority (66.7%) of malignancies in a murine model of *Brca1* deficiency were of lymphoid subtypes, whereas the majority (53.2%) of HMs with somatic *BRCA1* mutations in the COSMIC tumor database were of myeloid origin (Tate et al., 2019; Vasanthakumar et al., 2016). Similarly, a large case series of patients with known *BRCA1* germline mutations was similarly skewed toward myeloid malignancies (Iqbal et al., 2016).

In addition to my murine and family-based studies, I performed population-level analyses using data from the UK Biobank. On a population level, rare variant association analyses did not detect associations between germline *TTLL6* pLoF or missense variants and hematopoietic traits in the UK Biobank. However, caution is warranted as the UK Biobank remains relatively underpowered for the detection of the effects of "ultra-rare" variants, which are defined as those with population frequencies on the order of 10^{-4} or lower (Karczewski et al., 2021). The population frequency of the *TTLL6* variants in this study ranged from 2.2×10^{-5} (UChicago pedigree) up to 1.1×10^{-4} (University of Helsinki pedigrees). This frequency is likely sufficiently low so as to reduce the sensitivity of rare variant association analyses for the HT/HHMs in the current iteration of the UK Biobank. Germline mutations in *RUNX1* and *ETV6*, both of which are driven by ultra-rare germline mutations, similarly did not produce a signal for hematopoietic parameters, particularly platelet-related laboratory values, in my UK Biobank rare variant analyses. These analyses also did not show a signal for blood cancer-related phenotypes within the UK Biobank for individuals with *ETV6* or *RUNX1* pLoF variants.

However, the UK Biobank is sensitive to some hereditary hematopoietic

phenotypes, and I was able to replicate an association between established hereditary thrombocytopenia syndromes driven by ultra-rare germline variants in *TUBB1* and *GP1BA* (**Figure 6.2**). This finding suggests factors other than variant population frequency may drive population-level association analyses in the UK Biobank. For example, the most frequent known LP/P *TUBB1* variants are present in gnomAD at maximum population frequencies of $6.7x10^{-5}$, and the most frequent LP/P *GP1BA* variants have maximum population frequencies $1.61x10^{-4}$. *TUBB1*-driven hematopoietic syndromes, in particular, are as prevalent as *RUNX1* and *ETV6* germline mutations, as evidenced by a large phenotype-based case series focused on hereditary bleeding disorders that recently demonstrated that germline *TUBB1* mutations account for a similar proportion of hereditary thrombocytopenia cases as germline mutations in *RUNX1* and *ETV6* (Downes et al., 2019). Therefore, one would expect to see an association-based signal for *TUBB1*, *RUNX1*, and *ETV6* in the UK Biobank if variant population frequency was the only factor relevant for these associations.

The discrepant findings in the UK Biobank, therefore, between hereditary thrombocytopenia and HT/HHM syndromes cannot be explained entirely by the population frequency of the variants of interest. This discrepancy suggests that the cancer-specific factors of the HT/HHMs may reduce the sensitivity of burden analyses in the UK Biobank for these syndromes. One scenario may be that volunteers with strong family histories of blood cancers are hesitant to participate in genetics-based biobanking studies for fear of genetic discrimination (Caulfield & Murdoch, 2017). This apprehension may also explain why pLoF variants in the most common HHM-related gene, *DDX41* (maximum population frequency of $1.0x10^{-4}$) or one of the most penetrant HHMs (*GATA2*

germline mutations), similarly did not produce significant hematopoietic or cancer-related associations in population-level association studies using UK Biobank data. Alternatively, the hematopoietic phenotype for HT/HHM syndromes, as determined by laboratory-based parameters, may have lower than expected penetration in the larger population as compared to family-based analyses that have been published previously.

Even with these limitations of power and potential ascertainment bias, it is intriguing that two of the top phenotypic associations for *TTLL6* in the UK Biobank were for platelet-related traits (platelet count and platelet crit). This finding suggests that the thrombocytopenia phenotype observed in the UChicago pedigree may potentially be reproduced as the sample size in biobank-sized databases increases over time.

My murine model of *Ttll6* deficiency did not replicate the thrombocytopenia phenotype that I observed in the human pedigree, however murine models are notoriously poor models for reproducing the hematopoietic phenotypes of HT/HHMs that are observed in humans. This shortcoming is particularly noteworthy in regard to thrombocytopenia phenotypes. For example, conditional heterozygous *Etv6* knockout mice have normal platelet counts throughout their lifespans, in contrast to humans in which moderate-to-severe thrombocytopenia adheres to an autosomal dominant inheritance pattern (Hock et al., 2004; Moriyama et al., 2015; Noetzli et al., 2015; Topka et al., 2015; Zhang et al., 2015). Similarly, *Runx1* haploinsufficient mice had only a mild decrease in platelet counts in a 2004 study (Sun & Downing, 2004). In a later study, murine models with *Runx1* point mutations orthologous to those observed in human families showed normal platelet counts (Matheny et al., 2007). No murine models of *Ankrd26* 5’ UTR mutations exist currently. Additionally, no murine model to date has

effectively recapitulated the leukemogenic phenotype of germline *ANKRD26*, *ETV6*, or *RUNX1* heterozygous mutations in humans.

Finally, the late onset of tumorigenesis within my murine model, a median of 400 days for *Ttll6*$^{-/-}$ mice and of 650 days for *Ttll6*$^{+/-}$ animals, is intriguing given the role of TTLL6 in the STING pathway. At this time, the only other HHM-related gene in the STING pathway, *DDX41*, is notable for its late onset of leukemogenesis (62 years) in humans (Lewinsohn et al., 2015; Polprasert et al., 2015; Zhang et al., 2011). As detailed in Chapter II of this thesis, *DDX41*-driven HHMs are also notable for a relative paucity of CH prior to frank leukemogenesis. This suggests that HHMs driven by mutations in STING pathway-related genes have a particularly distinct leukemogenic mechanism that leads to "bland" pre-leukemic and leukemic states as compared to the other HT/HHMs and HHMs. My future research efforts will involve profiling the pre-leukemic milieu in my *Ttll6*-deficient murine models as well as developing methods by which leukemogenesis can be accelerated. This approach will improve the practical application of the *Ttll6*-deficient model while also offering potential insights into leukemogenesis in patients with germline mutations in *DDX41* or *TTLL6*, the two known STING-related HHMs. As TTLL6-mediated glutamylation of cGAS inhibits the activation of the STING pathway, one would hypothesize that deleterious germline *TTLL6* mutations lead to reduced inhibition of cGAS and chronic hyperactivation of the STING pathway. Potential mechanisms for increasing leukemogenesis in *Ttll6*-deficient mice, therefore, will likely involve exogenous stimulation of the STING pathway. Multiple methods of agonizing STING have been described, including the murine-specific STING agonist dimethyloxoxanthenyl acetic acid (DMXAA), the synthetic cyclic dinucleotide ML RR-S2 CDA (ADU-S100, Aduro Biotech), and

dimerized amidobenzimidazole. The success of STING agonism may be assessed by measuring products of the STING pathway, which include type 1 interferon, beta interferon (INF-β), and the chemokine ligands 9 and 10 (CXCL9 and CXCL10) (Li et al., 2019).

Chapter VII

Discussion, Future Directions, and Research Impact

Discussion

The goal of my doctoral training in Human Genetics was to address three major knowledge gaps in the field of HT/HHMs, with a particular emphasis on the HT/HHM syndromes that are marked by lifelong thrombocytopenia, platelet dysfunction, and increased lifetime leukemogenic risk.

First, individuals with HT and HHM syndromes experience variable lifetime risk for leukemogenesis, and it is currently not well known why some mutation carriers develop blood cancers and others do not experience leukemogenesis. Further, it is poorly understood how the pre-leukemic genetic milieu differs between the various HHM syndromes. I hypothesized that similar pre-leukemic states, as defined by somatic mutational spectra, would exist between the HT/HHM syndromes driven by *ANKRD26* and *RUNX1*, as the expression of *ANKRD26* is tightly controlled by the TF RUNX1 on a molecular level. Ultimately, this hypothesis proved to be partly incorrect, with *RUNX1*-driven HT/HHMs being notable for a preponderance of clones driven by early mutations in *BCOR*.

However, my hypothesis was also proven to be partly correct in that late second hit mutations in *RUNX1* are a defining characteristic of blood cancers that develop in *RUNX1* germline mutation carriers. Although I had access to a substantially larger number of unaffected and affected patients with *RUNX1* germline mutations as compared to *ANKRD26* germline mutations, it is remarkable that the only *ANKRD26* mutation carrier

in my study with a blood cancer also carried a *RUNX1* somatic mutation in their tumor clone. This suggests that somatic *BCOR* mutations are a relatively unique early driver in germline *RUNX1* mutation carriers, but somatic *RUNX1* mutations may be a shared leukemogenic mechanism among the HT/HHM syndromes driven by *ANKRD26* or *RUNX1* germline mutations. In contrast, *DDX41*-, *ETV6*-, and *GATA2*-related HHMs had a unique array of somatic mutations prior to leukemogenesis. *DDX41* mutation carriers rarely experienced CH, with late second hits in *DDX41* appearing to occur nearly simultaneously with leukemogenesis. In contrast, *GATA2* mutation carriers experienced a substantial risk of CH, but these mutations were located in the classic epigenetic regulators, such as *DNMT3A* and *KMT2C. ETV6* was notable for an absent of CH, but the sample size was relatively low.

The findings from this natural history study will enable me to address a major gap in the field, namely the lack of high-fidelity models of the pre-leukemic state in the HHMs. Based on my sequencing of patients in this study, it is apparent that single germline HHM-related genes are not sufficient to drive leukemogenesis in the vast majority of HHMs. However, no large natural history study of pre-leukemic states in the HHMs had been performed previously. My work has helped address this gap, and I will now be able to generate induced pluripotent stem cells with both the germline HHM-related mutation of interest in addition to high risk somatic mutations that I have identified in my sequencing studies. These modeling efforts will inform the field's understanding of both the molecular derangements that occur in these pre-leukemic states while also providing insights that may guide future therapeutic efforts to abrogate leukemogenesis in unaffected carriers of HHM-related germline mutations.

The second knowledge gap in the field of HHMs is that the link between the thrombocytopenic and leukemogenic phenotypes in HT/HHMs driven by germline *ANKRD26* mutations remains unknown. No murine models with relevant mutations in the 5' UTR of *Ankrd26* exist, and murine models frequently do not recapitulate the full phenotype observed in human patients with other germline HT/HHM-related mutations. Additionally, primary patient samples are difficult to obtain from *ANKRD26* mutation carriers in the clinic, as serial bone marrow biopsies are required to collect even small numbers of relevant hematopoietic tissue, such as hematopoietic stem and progenitor cells (HSPCs), megakaryocytes, and other progenitor cell lineages. There is understandably great hesitancy on the part of both patients and physicians to perform these invasive procedures solely for the purpose of research sample collection. Therefore, to address this knowledge gap and patient sample issue, I generated an iPSC line from cultured skin fibroblasts collected from a patient with a germline 5' UTR *ANKRD26* mutation. After confirming these iPSCs behaved as a *bona fide* stem cell line, I then differentiated this iPSC line into HSPCs (iPSC-HSPCs) and utilized RNA-seq to characterize the transcriptomic milieu in iPSCs and iPSC-HSPCs with germline *ANKRD26* mutations. This analysis demonstrated that the MAPK pathways are hyperactivated in both iPSCs and HSPCs with germline *ANKRD26* mutations. Prior studies have implicated disrupted MAPK/ERK1/ERK2 signaling as a mechanism that drives the thrombocytopenic phenotype in individuals with *ANKRD26* germline mutations, but these groups have only been able to speculate that this mechanism may also be present in HSPCs and more primitive lineages, which would potentially drive leukemogenesis in this patient population (Bluteau et al., 2014; Minamiguchi et al., 2001).

Finally, the demonstration that germline *TTLL6* mutations drive leukemogenesis in murine models and are significantly enriched in cohorts of patients with hereditary thrombocytopenia and HMs expands the known HT/HHM syndromes to include *ANKRD26*, *ETV6*, and *RUNX1.* This finding also further implicates the role of germline mutations in genes in the STING pathway, with both *TTLL6* and *DDX41* being involved in this molecular pathway.

Future Directions

Much of the work detailed in this thesis was performed in the midst of the COVID19 pandemic, which greatly hampered my ability to perform the laboratory-based assays that will constitute the next steps in many of my projects. For example, by demonstrating that the transcriptome in *ANKRD26*-mutated iPSCs and iPSC-HSPCs is notable for a persistent hyperactivation in MAPK-related pathways, I have potentially identified the link between leukemogenesis, platelet dysfunction, and thrombocytopenia in individuals with germline *ANKRD26* mutations. My next steps will involve validating that this transcriptome-level difference results in protein-level differences in the downstream targets within the MAPK pathways, which include phosphorylated STAT5, AKT, and ERK1/2. To perform these experiments, I will differentiate *ANKRD26*-mutated and control iPSCs into HSPCs. As before, I will utilize flow cytometry and an assessment of gross morphology of my cell lines in order to confirm the successful differentiation of my iPSCs into HSPCs. I will collect total cellular protein extracts at days 0, 8, and 10 of this differentiation. These time points correspond to the pluripotent state (day 0), early HSPCs (day 8), and mature HSPCs (day 10). If my hypothesis, informed by my transcriptomic

data, is correct, the *ANKRD26*-mutated iPSCs will have a significant increase in the level of phosphorylated STAT5, AKT, and ERK1/2, as has been observed in *ANKRD26*-mutated megakaryocytes and proplatelets (Bluteau et al., 2014). I will utilize Western blots to assess these protein levels in my iPSC lines. In addition to this work, I will differentiate both my *ANKRD26*-mutated and control iPSC-HSPCs into relevant cell lineages, including megakaryocytes, erythroid progenitors, and myeloid progenitors, via liquid expansion assays. I will also perform colony-forming assays using methylcellulose- and collagen-based semi-solid matrix systems. I will utilize my flow cytometry and colony forming assays to determine if the *ANKRD26*-mutated iPSCs demonstrate any discrepant capacity to form any of these more terminal cell lineages.

Should these protein-level studies confirm that MAPK hyperactivation is present in *ANKRD26*-mutated iPSCs and HSPCs, the natural extension of this project would be to determine if MAPK inhibition abrogates the phenotype that develops secondary to *ANKRD26* germline mutations. Numerous MAPK inhibitors have been developed and studied in clinical trials (Lee et al., 2020). The most pressing issue would be identifying the ideal experimental system for studying MAPK inhibition in *ANKRD26*-mutated cells. For example, should my iPSC-level studies identify a cellular phenotype that stems from *ANKRD26* mutations, such as the decreased production of megakaryocytes in liquid expansion assays, this phenotype could serve as a readout for low-throughput assays that examine the effect of increasing levels of MAPK inhibition on the phenotype of interest. For example, if my hypothesis that STAT5, AKT, and ERK1/2 phosphorylation is increased in *ANKRD26*-mutated cells is proven correct, this protein-level assay could serve as a valuable readout for these studies. Ultimately, I would assess the degree of

MAPK inhibition using Western blots, with the readout being normalization of phosphorylated STAT5, AKT, and ERK1/2. Candidate MAPK inhibitors that successfully normalize both this protein level readout as well as any cellular phenotype stemming from *ANKRD26* mutations would then be advanced into *in vivo* studies. It is worth noting that MAPK hyperactivation seems to be present in individuals with myeloid malignancies that carry somatic *RUNX1* mutations (Niimi et al., 2006). It is currently unknown if germline *RUNX1* mutations lead to MAPK hyperactivation. Therefore, I will perform the above experiments in both *ANKRD26*- and *RUNX1*-mutated iPSCs.

In addition to these studies in patient-derived iPSCs, I am collaborating with Dr. Linda Degenstein at the University of Chicago Transgenic Mouse/Embryonic Stem Cell Facility to utilize CRISPR/Cas9 to generate a murine model with 5' UTR *Ankrd26*-mutations. This model will be designed to be equivalent to the c.-119C>G mutation that is present in the iPSCs utilized in this thesis. If this process is successful, I will determine if the HT/HHM phenotype is recapitulated in this murine model. The successful generation of a 5' UTR *Ankrd26*-mutated murine model that recapitulates the thrombocytopenic and/or leukemogenic phenotype would provide a valuable model for testing any MAPK inhibitor candidates stemming from my work with the iPSC system. These studies could then ultimately inform clinical trials designed to determine if MAPK inhibition improves clinical outcomes in patients with HT/HHMs driven by germline *ANKRD26* mutations. These outcomes could include improvements in platelet counts, a reduction in leukemogenesis, and improved survival in patients with germline *ANKRD26* mutations who develop HMs.

In addition to my work studying MAPK hyperactivation and the cellular phenotype

of *ANKRD26*-mutated iPSCs, I will be performing a series of CRISPR/Cas9 experiments to introduce somatic mutations relevant to high-risk pre-leukemic states into my panel of patient-derived iPSCs. I have previously utilized CRISPR/Cas9 to disrupt the 5' UTR of my *ANKRD26*-mutated iPSCs using a protocol designed by my collaborators Drs. Paul Gadue and Deborah French at the University of Pennsylvania. I will implement this approach to introduce pathogenic *BCOR* mutations into my *RUNX1* iPSC line and *RUNX1* mutations into my *ANKRD26*-mutated and *RUNX1*-mutated iPSC lines. These efforts will build upon my natural profiling studies in the HT/HHMs and HHMs to produce the first patient-derived models of high risk preleukemic states in the HT/HHM syndromes. Ultimately, these models may be utilized in highly translational research studies that will be focused on the downstream effects stemming from these combinations of germline and somatic mutations. Additionally, the iPSC platform is well-suited to functional genomics studies, such as CRISPR screens and CRISPR interference, that may inform the next generation of therapies for individuals with HT/HHM syndromes.

The next steps involving *TTLL6* will largely focus on completing the histopathologic and transplant studies of tumors that spontaneously form in *Ttll6*$^{-/-}$ and *Ttll6*$^{+/-}$ animals so as to determine the spectrum of malignancies to which these individuals are susceptible. I am collaborating with Dr. Jason Cheng in the University of Chicago Department of Pathology to complete these studies. I will then proceed to use DMXAA to agonize the STING pathway with the goal of expediting tumorigenesis in *Ttll6* deficient animals. If this approach is successful, it will improve the practicality of this model for the study of *TTLL6*-mutated HT/HHMs. The STING pathway has been a focus of drug development and I ultimately anticipate that some of these agents, in particular STING antagonists, may be

repurposed for the care of individuals with germline mutations in *TTLL6*.

Additionally, I will extend my studies of *TTLL6* to include an additional focus on patient-level data so as to determine how this syndrome differs from the other HHMs. Anecdotally, the identification of novel HT/HHM or HHM syndromes leads to a rapid identification of additional pedigrees both within one's own institution and in cohorts collected by other groups. This expansion in the number of known carriers of germline *TTLL6* mutations will facilitate mutational profiling of both unaffected and affected mutation carriers. For example, I have not investigated the pre-leukemic milieu in unaffected germline *TTLL6* mutation carriers. Similarly, the mutational spectrum in tumors that develop in *TTLL6* mutation-driven HT/HHMs is unknown aside from the three patients sequenced at the University of Helsinki. These studies would represent a natural extension of the work detailed in this thesis, particularly the benefit of having multi-institutional groups contributing patient samples for analysis using a uniform bioinformatics pipeline. Given their mutual roles in the STING pathway, I would hypothesize that both *TTLL6* and *DDX41* mutation carriers would experience similar patterns of molecularly "bland" pre-leukemic states. I would also hypothesize that tumors that develop in these individuals would have a relative paucity of somatic mutations and normal karyotypes.

In addition to identifying additional families with germline *TTLL6* mutations, I have also collaborated with Dr. Daniela del Gaudio in the Department of Human Genetics at the University of Chicago to identify additional patients with germline variants in *ANKRD26*, *ETV6*, or *RUNX1* who have been sequenced in Dr. del Gaudio's laboratory since 2015. I am currently working with Dr. del Gaudio and her team to identify patients

with LP/P variants or “high risk” variants of undetermined significance (VUSs). Ultimately, I will expand my cohort of HT/HHM patients to include these patients, which will increase my power to estimate the prevalence of CH clones in HT/HHMs, will identify additional somatic drivers in HMs that develop in these patients, and will increase the number of patients with HT/HHMs enrolled on research protocols at the University of Chicago.

Research Impact

In summary, the research in this thesis has identified the unique leukemogenic mechanisms that exist in unaffected carriers of germline mutations in *ANKRD26*, *DDX41*, *ETV6*, *GATA2*, and *RUNX1*. In Chapters II and III of this thesis, I demonstrated that individuals with germline mutations in *GATA2* or *RUNX1* appear to be at high risk for developing CH prior to frank leukemogenesis. However, the somatic mutations present in these clones differ by syndrome, as *BCOR*-mutated clones predominantly occur in *RUNX1* mutation carriers and mutations in the classic epigenetic regulators *DNMT3A* or *KMT2C* drive clones in *GATA2* mutation carriers. Although *ANKRD26* germline mutation carriers have lower rates of CH relative to *RUNX1* germline mutation carriers, both *ANKRD26*- and *RUNX1*-driven syndromes are remarkable for late somatic *RUNX1* mutations that ultimately drive leukemogenesis. In contrast, germline *DDX41* mutation carriers experience extraordinarily low rates of CH that are equivalent to those in population controls. However, leukemogenesis in *DDX41* mutation carriers is remarkable for a preponderance of second hits in *DDX41* and otherwise “bland” tumors. Additionally, my work has offered additional insights into the role of MAPK hyperactivation in driving both the leukemogenic and the thrombocytopenic phenotype in germline *ANKRD26*

mutation carriers. Finally, I have identified germline mutations in *TTLL6* as a fourth syndrome that phenocopies the known HT/HHMs driven by germline mutations in *ANKRD26*, *ETV6*, or *RUNX1*.

One important ramification of this doctoral work is that it may will better guide the diagnosis and care of individuals at risk for HHMs. For example, during the course of my doctoral training, I have demonstrated that currently available diagnostic methods for HHMs must incorporate data produced by next generation sequencing (NGS) panels. These panels are designed to profile somatic mutations in a malignancy, but these assays frequently detect germline LP/P variants that are passed into the leukemic clone before being detected incidentally. In Chapter I of this thesis, I demonstrated this process occurs more frequently than was previously recognized (Drazer et al., 2018). Next, I have shown that the NGS assays that are most frequently used in the diagnosis of HHMs are frequently inadequate for their intended purpose. These assays frequently do not sequence all of the genes known to drive HHMs, and they also are frequently incapable of detecting the full spectrum of variant types that are responsible for HHMs (Roloff et al., 2020). Finally, in Chapter VI, I demonstrate that germline *TTLL6* mutations represent a novel HT/HHM syndrome, further expanding the number of genes responsible for HHMs and further implicating the STING pathway as an important contributor to hereditary cancer syndromes. Ultimately, my work involving the use of patient-derived iPSCs, detailed in Chapter IV of this thesis, may better inform the development of the first therapies that are tested specifically for the prevention and treatment of leukemogenesis in HHMs.

References

Abou Dalle, I., Kantarjian, H., Bannon, S. A., Kanagal-Shamanna, R., Routbort, M., Patel, K. P., . . . DiNardo, C. D. (2020). Successful lenalidomide treatment in high risk myelodysplastic syndrome with germline DDX41 mutation. *Am J Hematol*, *95*(2), 227-229. https://doi.org/10.1002/ajh.25610

Al-Kali, A., Quintás-Cardama, A., Luthra, R., Bueso-Ramos, C., Pierce, S., Kadia, T., . . . Garcia-Manero, G. (2013). Prognostic impact of RAS mutations in patients with myelodysplastic syndrome. *Am J Hematol*, *88*(5), 365-369. https://doi.org/10.1002/ajh.23410

Andrysik, Z., Galbraith, M. D., Guarnieri, A. L., Zaccara, S., Sullivan, K. D., Pandey, A., . . . Espinosa, J. M. (2017). Identification of a core TP53 transcriptional program with highly distributed tumor suppressive activity. *Genome Res*, *27*(10), 1645-1657. https://doi.org/10.1101/gr.220533.117

Bader, G. D., & Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, *4*, 2. https://doi.org/10.1186/1471-2105-4-2

Balduini, C. L., & Savoia, A. (2012). Genetics of familial forms of thrombocytopenia. *Hum Genet*, *131*(12), 1821-1832. https://doi.org/10.1007/s00439-012-1215-x

Bandipalliam, P. (2005). Syndrome of early onset colon cancers, hematologic malignancies & features of neurofibromatosis in HNPCC families with homozygous mismatch repair gene mutations. *Fam Cancer*, *4*(4), 323-333. https://doi.org/10.1007/s10689-005-8351-6

Barjesteh van Waalwijk van Doorn-Khosrovani, S., Spensberger, D., de Knegt, Y., Tang, M., Löwenberg, B., & Delwel, R. (2005). Somatic heterozygous mutations in ETV6 (TEL) and frequent absence of ETV6 protein in acute myeloid leukemia. *Oncogene*, *24*(25), 4129-4137. https://doi.org/10.1038/sj.onc.1208588

Bejar, R., Stevenson, K., Abdel-Wahab, O., Galili, N., Nilsson, B., Garcia-Manero, G., . . . Ebert, B. L. (2011). Clinical effect of point mutations in myelodysplastic syndromes. *N Engl J Med*, *364*(26), 2496-2506. https://doi.org/10.1056/NEJMoa1013343

Bellissimo, D. C., Chen, C. H., Zhu, Q., Bagga, S., Lee, C. T., He, B., . . . Speck, N. A. (2020). Runx1 negatively regulates inflammatory cytokine production by neutrophils in response to Toll-like receptor signaling. *Blood Adv*, *4*(6), 1145-1158. https://doi.org/10.1182/bloodadvances.2019000785

Bera, T. K., Liu, X. F., Yamada, M., Gavrilova, O., Mezey, E., Tessarollo, L., . . . Pastan, I. (2008). A model for obesity and gigantism due to disruption of the Ankrd26 gene. *Proc Natl Acad Sci U S A*, *105*(1), 270-275. https://doi.org/10.1073/pnas.0710978105

Biermer, A. (1861). In (552 ed., Vol. 20). Archives of pathology and anatomy.

Bluteau, D., Balduini, A., Balayn, N., Currao, M., Nurden, P., Deswarte, C., . . . Raslova, H. (2014). Thrombocytopenia-associated mutations in the ANKRD26 regulatory region induce MAPK hyperactivation. *J Clin Invest*, *124*(2), 580-591. https://doi.org/10.1172/JCI71861

Borst, S., Nations, C. C., Klein, J. G., Pavani, G., Maguire, J. A., Camire, R. M., . . . Gadue, P. (2021). Study of inherited thrombocytopenia resulting from mutations in

ETV6 or RUNX1 using a human pluripotent stem cell model. *Stem Cell Reports*, *16*(6), 1458-1467. https://doi.org/10.1016/j.stemcr.2021.04.013

Bougeard, G., Charbonnier, F., Moerman, A., Martin, C., Ruchoux, M. M., Drouot, N., & Frébourg, T. (2003). Early onset brain tumor and lymphoma in MSH2-deficient children. *Am J Hum Genet*, *72*(1), 213-216. https://doi.org/10.1086/345297

Bougeard, G., Renaux-Petel, M., Flaman, J. M., Charbonnier, C., Fermey, P., Belotti, M., . . . Frebourg, T. (2015). Revisiting Li-Fraumeni Syndrome From TP53 Mutation Carriers. *J Clin Oncol*, *33*(21), 2345-2352. https://doi.org/10.1200/JCO.2014.59.5728

Braicu, C., Buse, M., Busuioc, C., Drula, R., Gulei, D., Raduly, L., . . . Berindan-Neagoe, I. (2019). A Comprehensive Review on MAPK: A Promising Therapeutic Target in Cancer. *Cancers (Basel)*, *11*(10). https://doi.org/10.3390/cancers11101618

Burroughs, L. M., Shimamura, A., Talano, J. A., Domm, J. A., Baker, K. K., Delaney, C., . . . Woolfrey, A. E. (2017). Allogeneic Hematopoietic Cell Transplantation Using Treosulfan-Based Conditioning for Treatment of Marrow Failure Disorders. *Biol Blood Marrow Transplant*, *23*(10), 1669-1677. https://doi.org/10.1016/j.bbmt.2017.06.002

Burrows, C. K., Banovich, N. E., Pavlovic, B. J., Patterson, K., Gallego Romero, I., Pritchard, J. K., & Gilad, Y. (2016). Genetic Variation, Not Cell Type of Origin, Underlies the Majority of Identifiable Regulatory Differences in iPSCs. *PLoS Genet*, *12*(1), e1005793. https://doi.org/10.1371/journal.pgen.1005793

Cai, X., Gao, L., Teng, L., Ge, J., Oo, Z. M., Kumar, A. R., . . . Speck, N. A. (2015). Runx1 Deficiency Decreases Ribosome Biogenesis and Confers Stress Resistance to Hematopoietic Stem and Progenitor Cells. *Cell Stem Cell*, *17*(2), 165-177. https://doi.org/10.1016/j.stem.2015.06.002

Cantor, D. (2006). The frustrations of families: Henry Lynch, heredity, and cancer control, 1962-1975. *Med Hist*, *50*(3), 279-302.

Caulfield, T., & Murdoch, B. (2017). Genes, cells, and biobanks: Yes, there's still a consent problem. *PLoS Biol*, *15*(7), e2002654. https://doi.org/10.1371/journal.pbio.2002654

Chen, D. H., Below, J. E., Shimamura, A., Keel, S. B., Matsushita, M., Wolff, J., . . . Raskind, W. H. (2016). Ataxia-Pancytopenia Syndrome Is Caused by Missense Mutations in SAMD9L. *Am J Hum Genet*, *98*(6), 1146-1158. https://doi.org/10.1016/j.ajhg.2016.04.009

Churpek, J. E., Pyrtel, K., Kanchi, K. L., Shao, J., Koboldt, D., Miller, C. A., . . . Graubert, T. A. (2015). Genomic analysis of germline and somatic variants in familial myelodysplasia/acute myeloid leukemia. *Blood*. https://doi.org/10.1182/blood-2015-04-641100

Collin, M. (2017). I am SAMD9L: 7q regulator I am. *Blood*, *129*(16), 2210-2212. https://doi.org/10.1182/blood-2017-03-770198

Cuellar-Rodriguez, J., Gea-Banacloche, J., Freeman, A. F., Hsu, A. P., Zerbe, C. S., Calvo, K. R., . . . Hickstein, D. D. (2011). Successful allogeneic hematopoietic stem cell transplantation for GATA2 deficiency. *Blood*, *118*(13), 3715-3720. https://doi.org/10.1182/blood-2011-06-365049

Davidsson, J., Puschmann, A., Tedgård, U., Bryder, D., Nilsson, L., & Cammenga, J. (2018). SAMD9 and SAMD9L in inherited predisposition to ataxia, pancytopenia,

and myeloid malignancies. *Leukemia*, *32*(5), 1106-1115. https://doi.org/10.1038/s41375-018-0074-4

Desai, P., Mencia-Trinchant, N., Savenkov, O., Simon, M. S., Cheang, G., Lee, S., . . . Hassane, D. C. (2018). Somatic mutations precede acute myeloid leukemia years before diagnosis. *Nat Med*, *24*(7), 1015-1023. https://doi.org/10.1038/s41591-018-0081-z

Dickinson, R. E., Griffin, H., Bigley, V., Reynard, L. N., Hussain, R., Haniffa, M., . . . Collin, M. (2011). Exome sequencing identifies GATA-2 mutation as the cause of dendritic cell, monocyte, B and NK lymphoid deficiency. *Blood*, *118*(10), 2656-2658. https://doi.org/10.1182/blood-2011-06-360313

DiFilippo, E. C., Coltro, G., Carr, R. M., Mangaonkar, A. A., Binder, M., Khan, S. P., . . . Patnaik, M. M. (2020). Spectrum of abnormalities and clonal transformation in germline RUNX1 familial platelet disorder and a genomic comparative analysis with somatic RUNX1 mutations in MDS/MPN overlap neoplasms. *Leukemia*, *34*(9), 2519-2524. https://doi.org/10.1038/s41375-020-0752-x

Donadieu, J., Lamant, M., Fieschi, C., de Fontbrune, F. S., Caye, A., Ouachee, M., . . . group, F. G. s. (2018). Natural history of GATA2 deficiency in a survey of 79 French and Belgian patients. *Haematologica*, *103*(8), 1278-1287. https://doi.org/10.3324/haematol.2017.181909

Downes, K., Megy, K., Duarte, D., Vries, M., Gebhart, J., Hofer, S., . . . BioResource, N. (2019). Diagnostic high-throughput sequencing of 2396 patients with bleeding, thrombotic, and platelet disorders. *Blood*, *134*(23), 2082-2091. https://doi.org/10.1182/blood.2018891192

Drazer, M. W., Feurstein, S., West, A. H., Jones, M. F., Churpek, J. E., & Godley, L. A. (2016). How I diagnose and manage individuals at risk for inherited myeloid malignancies. *Blood*, *128*(14), 1800-1813. https://doi.org/10.1182/blood-2016-05-670240

Drazer, M. W., Kadri, S., Sukhanova, M., Patil, S. A., West, A. H., Feurstein, S., . . . Godley, L. A. (2018). Prognostic tumor sequencing panels frequently identify germ line variants associated with hereditary hematopoietic malignancies. *Blood Adv*, *2*(2), 146-150. https://doi.org/10.1182/bloodadvances.2017013037

Evans, L. T., Anglen, T., Scott, P., Lukasik, K., Loncarek, J., & Holland, A. J. (2021). ANKRD26 recruits PIDD1 to centriolar distal appendages to activate the PIDDosome following centrosome amplification. *EMBO J*, *40*(4), e105106. https://doi.org/10.15252/embj.2020105106

Feurstein, S., Churpek, J. E., Walsh, T., Keel, S., Hakkarainen, M., Schroeder, T., . . . Godley, L. A. (2021). Germline variants drive myelodysplastic syndrome in young adults. *Leukemia*. https://doi.org/10.1038/s41375-021-01137-0

Feurstein, S., Drazer, M. W., & Godley, L. A. (2016). Genetic predisposition to leukemia and other hematologic malignancies. *Semin Oncol*, *43*(5), 598-608. https://doi.org/10.1053/j.seminoncol.2016.10.003

Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., . . . Flicek, P. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res*, *47*(D1), D766-D773. https://doi.org/10.1093/nar/gky955

Friedman, L. S., Ostermeyer, E. A., Szabo, C. I., Dowd, P., Lynch, E. D., Rowell, S. E., & King, M. C. (1994). Confirmation of BRCA1 by analysis of germline mutations linked to breast and ovarian cancer in ten families. *Nat Genet*, *8*(4), 399-404. https://doi.org/10.1038/ng1294-399

Garnham, C. P., Vemu, A., Wilson-Kubalek, E. M., Yu, I., Szyk, A., Lander, G. C., . . . Roll-Mecak, A. (2015). Multivalent Microtubule Recognition by Tubulin Tyrosine Ligase-like Family Glutamylases. *Cell*, *161*(5), 1112-1123. https://doi.org/10.1016/j.cell.2015.04.003

Genovese, G., Kähler, A. K., Handsaker, R. E., Lindberg, J., Rose, S. A., Bakhoum, S. F., . . . McCarroll, S. A. (2014). Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N Engl J Med*, *371*(26), 2477-2487. https://doi.org/10.1056/NEJMoa1409405

Godley, L. A. (2014). Inherited predisposition to acute myeloid leukemia. *Semin Hematol*, *51*(4), 306-321. https://doi.org/10.1053/j.seminhematol.2014.08.001

Grossman, J., Cuellar-Rodriguez, J., Gea-Banacloche, J., Zerbe, C., Calvo, K., Hughes, T., . . . Hickstein, D. D. (2014). Nonmyeloablative allogeneic hematopoietic stem cell transplantation for GATA2 deficiency. *Biol Blood Marrow Transplant*, *20*(12), 1940-1948. https://doi.org/10.1016/j.bbmt.2014.08.004

Guidugli, L., Johnson, A. K., Alkorta-Aranburu, G., Nelakuditi, V., Arndt, K., Churpek, J. E., . . . Li, Z. (2017). Clinical utility of gene panel-based testing for hereditary myelodysplastic syndrome/acute leukemia predisposition syndromes. *Leukemia*, *31*(5), 1226-1229. https://doi.org/10.1038/leu.2017.28

Gunz, F. W., Gunz, J. P., Veale, A. M., Chapman, C. J., & Houston, I. B. (1975). Familial leukaemia: a study of 909 families. *Scand J Haematol*, *15*(2), 117-131.

Hahn, C. N., Chong, C. E., Carmichael, C. L., Wilkins, E. J., Brautigan, P. J., Li, X. C., . . . Scott, H. S. (2011). Heritable GATA2 mutations associated with familial myelodysplastic syndrome and acute myeloid leukemia. *Nat Genet*, *43*(10), 1012-1017. https://doi.org/10.1038/ng.913

Hall, J. M., Friedman, L., Guenther, C., Lee, M. K., Weber, J. L., Black, D. M., & King, M. C. (1992). Closing in on a breast cancer gene on chromosome 17q. *Am J Hum Genet*, *50*(6), 1235-1242.

Hall, J. M., Lee, M. K., Newman, B., Morrow, J. E., Anderson, L. A., Huey, B., & King, M. C. (1990). Linkage of early-onset familial breast cancer to chromosome 17q21. *Science*, *250*(4988), 1684-1689. https://doi.org/10.1126/science.2270482

Hernández, J. M., González, M. B., García, J. L., Ferro, M. T., Gutiérrez, N. C., Marynen, P., & San Miguel, J. F. (2000). Two cases of myeloid disorders and a t(8;12)(q12;p13). *Haematologica*, *85*(1), 31-34.

Hochberg, Y., & Benjamini, Y. (1990). More powerful procedures for multiple significance testing. *Stat Med*, *9*(7), 811-818. https://doi.org/10.1002/sim.4780090710

Hock, H., Meade, E., Medeiros, S., Schindler, J. W., Valk, P. J., Fujiwara, Y., & Orkin, S. H. (2004). Tel/Etv6 is an essential and selective regulator of adult hematopoietic stem cell survival. *Genes Dev*, *18*(19), 2336-2341. https://doi.org/10.1101/gad.1239604

Homan, C. C., King-Smith, S. L., Lawrence, D. M., Arts, P., Feng, J., Andrews, J., . . . Brown, A. L. (2021). The RUNX1 Database (RUNX1db): establishment of an expert curated RUNX1 registry and genomics database as a public resource for

familial platelet disorder with myeloid malignancy. *Haematologica*. https://doi.org/10.3324/haematol.2021.278762

Horwitz, M., Goode, E. L., & Jarvik, G. P. (1996). Anticipation in familial leukemia. *Am J Hum Genet*, *59*(5), 990-998.

Hwang, S., Kim, E., Lee, I., & Marcotte, E. M. (2015). Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Sci Rep*, *5*, 17875. https://doi.org/10.1038/srep17875

Iqbal, J., Nussenzweig, A., Lubinski, J., Byrski, T., Eisen, A., Bordeleau, L., . . . Group, H. B. C. R. (2016). The incidence of leukaemia in women with BRCA1 and BRCA2 mutations: an International Prospective Cohort Study. *Br J Cancer*, *114*(10), 1160-1164. https://doi.org/10.1038/bjc.2016.58

Jaiswal, S., Fontanillas, P., Flannick, J., Manning, A., Grauman, P. V., Mar, B. G., . . . Ebert, B. L. (2014). Age-related clonal hematopoiesis associated with adverse outcomes. *N Engl J Med*, *371*(26), 2488-2498. https://doi.org/10.1056/NEJMoa1408617

Kadri, S., Long, B. C., Mujacic, I., Zhen, C. J., Wurst, M. N., Sharma, S., . . . Segal, J. P. (2017). Clinical Validation of a Next-Generation Sequencing Genomic Oncology Panel via Cross-Platform Benchmarking against Established Amplicon Sequencing Assays. *J Mol Diagn*, *19*(1), 43-56. https://doi.org/10.1016/j.jmoldx.2016.07.012

Karastaneva, A., Nebral, K., Schlagenhauf, A., Baschin, M., Palankar, R., Juch, H., . . . Seidel, M. G. (2020). Novel phenotypes observed in patients with. *J Med Genet*, *57*(6), 427-433. https://doi.org/10.1136/jmedgenet-2019-106339

Karczewski, K. J., Solomonson, M., Chao, K. R., Goodrich, J. K., Tiao, G., Lu, W., . . . Neale, B. M. (2021). Systematic single-variant and gene-based association testing of 3,700 phenotypes in

281,850 UK Biobank exomes In. medRxiv.

Keel, S. B., Scott, A., Sanchez-Bonilla, M., Ho, P. A., Gulsuner, S., Pritchard, C. C., . . . Shimamura, A. (2016). Genetic features of myelodysplastic syndrome and aplastic anemia in pediatric and young adult patients. *Haematologica*. https://doi.org/10.3324/haematol.2016.149476

Kewan, T., Noss, R., Godley, L. A., Rogers, H. J., & Carraway, H. E. (2020). Inherited Thrombocytopenia Caused by Germline. *J Investig Med High Impact Case Rep*, *8*, 2324709620938941. https://doi.org/10.1177/2324709620938941

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*, *37*(8), 907-915. https://doi.org/10.1038/s41587-019-0201-4

Kunishima, S., Kobayashi, R., Itoh, T. J., Hamaguchi, M., & Saito, H. (2009). Mutation of the beta1-tubulin gene associated with congenital macrothrombocytopenia affecting microtubule assembly. *Blood*, *113*(2), 458-461. https://doi.org/10.1182/blood-2008-06-162610

Lacroix, B., van Dijk, J., Gold, N. D., Guizetti, J., Aldrian-Herrada, G., Rogowski, K., . . . Janke, C. (2010). Tubulin polyglutamylation stimulates spastin-mediated microtubule severing. *J Cell Biol*, *189*(6), 945-954. https://doi.org/10.1083/jcb.201001024

Lee, S., Rauch, J., & Kolch, W. (2020). Targeting MAPK Signaling in Cancer: Mechanisms of Drug Resistance and Sensitivity. *Int J Mol Sci*, *21*(3). https://doi.org/10.3390/ijms21031102

Lewinsohn, M., Brown, A. L., Weinel, L. M., Phung, C., Rafidi, G., Lee, M. K., . . . Scott, H. S. (2016). Novel germ line DDX41 mutations define families with a lower age of MDS/AML onset and lymphoid malignancies. *Blood*, *127*(8), 1017-1023. https://doi.org/10.1182/blood-2015-10-676098

Lewinsohn, M., Brown, A. L., Weinel, L. M., Phung, C., Rafidi, G., Lee, M. K., . . . Scott, H. S. (2015). Novel germline DDX41 mutations define families with a lower age of MDS/AML onset, and lymphoid malignancies. *Blood*. https://doi.org/10.1182/blood-2015-10-676098

Li, A., Yi, M., Qin, S., Song, Y., Chu, Q., & Wu, K. (2019). Activating cGAS-STING pathway for the optimal effect of cancer immunotherapy. *J Hematol Oncol*, *12*(1), 35. https://doi.org/10.1186/s13045-019-0721-x

Li, F. P., & Fraumeni, J. F. (1969). Soft-tissue sarcomas, breast cancer, and other neoplasms. A familial syndrome? *Ann Intern Med*, *71*(4), 747-752. https://doi.org/10.7326/0003-4819-71-4-747

Liu, X., & Wang, C. (2016). The emerging roles of the STING adaptor protein in immunity and diseases. *Immunology*, *147*(3), 285-291. https://doi.org/10.1111/imm.12561

Maguire, J. A., Cardenas-Diaz, F. L., Gadue, P., & French, D. L. (2019). Highly Efficient CRISPR-Cas9-Mediated Genome Editing in Human Pluripotent Stem Cells. *Curr Protoc Stem Cell Biol*, *48*(1), e64. https://doi.org/10.1002/cpsc.64

Marquez, R., Hantel, A., Lorenz, R., Neistadt, B., Wong, J., Churpek, J. E., . . . Godley, L. A. (2014). A new family with a germline ANKRD26 mutation and predisposition to myeloid malignancies. *Leuk Lymphoma*. https://doi.org/10.3109/10428194.2014.903476

Matheny, C. J., Speck, M. E., Cushing, P. R., Zhou, Y., Corpora, T., Regan, M., . . . Speck, N. A. (2007). Disease mutations in RUNX1 and RUNX2 create nonfunctional, dominant-negative, or hypomorphic alleles. *EMBO J*, *26*(4), 1163-1175. https://doi.org/10.1038/sj.emboj.7601568

Matsumura, T., Nakamura-Ishizu, A., Takaoka, K., Maki, H., Muddineni, S. S. N. A., Wang, C. Q., . . . Suda, T. (2019). TUBB1 dysfunction in inherited thrombocytopenia causes genome instability. *Br J Haematol*, *185*(5), 888-902. https://doi.org/10.1111/bjh.15835

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., . . . DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, *20*(9), 1297-1303. https://doi.org/10.1101/gr.107524.110

McReynolds, L. J., Yang, Y., Yuen Wong, H., Tang, J., Zhang, Y., Mulé, M. P., . . . Hourigan, C. S. (2019). MDS-associated mutations in germline GATA2 mutated patients with hematologic manifestations. *Leuk Res*, *76*, 70-75. https://doi.org/10.1016/j.leukres.2018.11.013

Melazzini, F., Palombo, F., Balduini, A., De Rocco, D., Marconi, C., Noris, P., . . . Savoia, A. (2016). Clinical and pathogenetic features of ETV6 related thrombocytopenia with predisposition to acute lymphoblastic leukemia. *Haematologica*. https://doi.org/10.3324/haematol.2016.147496

Mills, J. A., Paluru, P., Weiss, M. J., Gadue, P., & French, D. L. (2014). Hematopoietic differentiation of pluripotent stem cells in culture. *Methods Mol Biol*, *1185*, 181-194. https://doi.org/10.1007/978-1-4939-1133-2_12

Minamiguchi, H., Kimura, T., Urata, Y., Miyazaki, H., Bamba, T., Abe, T., & Sonoda, Y. (2001). Simultaneous signalling through c-mpl, c-kit and CXCR4 enhances the proliferation and differentiation of human megakaryocyte progenitors: possible roles of the PI3-K, PKC and MAPK pathways. *Br J Haematol*, *115*(1), 175-185. https://doi.org/10.1046/j.1365-2141.2001.03068.x

Moriyama, T., Metzger, M. L., Wu, G., Nishii, R., Qian, M., Devidas, M., . . . Yang, J. J. (2015). Germline genetic variation in ETV6 and risk of childhood acute lymphoblastic leukaemia: a systematic genetic study. *Lancet Oncol*, *16*(16), 1659-1666. https://doi.org/10.1016/S1470-2045(15)00369-1

Nagamachi, A., Matsui, H., Asou, H., Ozaki, Y., Aki, D., Kanai, A., . . . Inaba, T. (2013). Haploinsufficiency of SAMD9L, an endosome fusion facilitator, causes myeloid malignancies in mice mimicking human diseases with monosomy 7. *Cancer Cell*, *24*(3), 305-317. https://doi.org/10.1016/j.ccr.2013.08.011

Narumi, S., Amano, N., Ishii, T., Katsumata, N., Muroya, K., Adachi, M., . . . Hasegawa, T. (2016). SAMD9 mutations cause a novel multisystem disorder, MIRAGE syndrome, and are associated with loss of chromosome 7. *Nat Genet*, *48*(7), 792-797. https://doi.org/10.1038/ng.3569

Niimi, H., Harada, H., Harada, Y., Ding, Y., Imagawa, J., Inaba, T., . . . Kimura, A. (2006). Hyperactivation of the RAS signaling pathway in myelodysplastic syndrome with AML1/RUNX1 point mutations. *Leukemia*, *20*(4), 635-644. https://doi.org/10.1038/sj.leu.2404136

Nishii, R., Baskin-Doerfler, R., Yang, W., Oak, N., Zhao, X., Hoshitsuki, K., . . . Yang, J. J. (2021). Molecular basis of ETV6-mediated predisposition to childhood acute lymphoblastic leukemia. *Blood*, *137*(3), 364-373. https://doi.org/10.1182/blood.2020006164

Noetzli, L., Lo, R. W., Lee-Sherick, A. B., Callaghan, M., Noris, P., Savoia, A., . . . Di Paola, J. (2015). Germline mutations in ETV6 are associated with thrombocytopenia, red cell macrocytosis and predisposition to lymphoblastic leukemia. *Nat Genet*, *47*(5), 535-538. https://doi.org/10.1038/ng.3253

Noris, P., Favier, R., Alessi, M. C., Geddis, A. E., Kunishima, S., Heller, P. G., . . . Balduini, C. L. (2013). ANKRD26-related thrombocytopenia and myeloid malignancies. *Blood*, *122*(11), 1987-1989. https://doi.org/10.1182/blood-2013-04-499319

Noris, P., Perrotta, S., Seri, M., Pecci, A., Gnan, C., Loffredo, G., . . . Savoia, A. (2011). Mutations in ANKRD26 are responsible for a frequent form of inherited thrombocytopenia: analysis of 78 patients from 21 families. *Blood*, *117*(24), 6673-6680. https://doi.org/10.1182/blood-2011-02-336537

Odero, M. D., Carlson, K., Calasanz, M. J., Lahortiga, I., Chinwalla, V., & Rowley, J. D. (2001). Identification of new translocations involving ETV6 in hematologic malignancies by fluorescence in situ hybridization and spectral karyotyping. *Genes Chromosomes Cancer*, *31*(2), 134-142. https://doi.org/10.1002/gcc.1127

Papaemmanuil, E., Hosking, F. J., Vijayakrishnan, J., Price, A., Olver, B., Sheridan, E., . . . Houlston, R. S. (2009). Loci on 7p12.2, 10q21.2 and 14q11.2 are associated

with risk of childhood acute lymphoblastic leukemia. *Nat Genet*, *41*(9), 1006-1010. https://doi.org/10.1038/ng.430

Parta, M., Shah, N. N., Baird, K., Rafei, H., Calvo, K. R., Hughes, T., . . . Hickstein, D. D. (2018). Allogeneic Hematopoietic Stem Cell Transplantation for GATA2 Deficiency Using a Busulfan-Based Regimen. *Biol Blood Marrow Transplant*, *24*(6), 1250-1259. https://doi.org/10.1016/j.bbmt.2018.01.030

Pavlovic, B. J., Blake, L. E., Roux, J., Chavarria, C., & Gilad, Y. (2018). A Comparative Assessment of Human and Chimpanzee iPSC-derived Cardiomyocytes with Primary Heart Tissues. *Sci Rep*, *8*(1), 15312. https://doi.org/10.1038/s41598-018-33478-9

Perez Botero, J., Chen, D., He, R., Viswanatha, D. S., Majerus, J. A., Coon, L. M., . . . Patnaik, M. M. (2016). Clinical and laboratory characteristics in congenital ANKRD26 mutation-associated thrombocytopenia: A detailed phenotypic study of a family. *Platelets*, *27*(7), 712-715. https://doi.org/10.3109/09537104.2016.1171305

Pippucci, T., Savoia, A., Perrotta, S., Pujol-Moix, N., Noris, P., Castegnaro, G., . . . Balduini, C. L. (2011). Mutations in the 5' UTR of ANKRD26, the ankirin repeat domain 26 gene, cause an autosomal-dominant form of inherited thrombocytopenia, THC2. *Am J Hum Genet*, *88*(1), 115-120. https://doi.org/10.1016/j.ajhg.2010.12.006

Polprasert, C., Schulze, I., Sekeres, M. A., Makishima, H., Przychodzen, B., Hosono, N., . . . Maciejewski, J. P. (2015). Inherited and Somatic Defects in DDX41 in Myeloid Neoplasms. *Cancer Cell*, *27*(5), 658-670. https://doi.org/10.1016/j.ccell.2015.03.017

Pouliot, G. P., Degar, J., Hinze, L., Kochupurakkal, B., Vo, C. D., Burns, M. A., . . . Gutierrez, A. (2019). Fanconi-BRCA pathway mutations in childhood T-cell acute lymphoblastic leukemia. *PLoS One*, *14*(11), e0221288. https://doi.org/10.1371/journal.pone.0221288

Quesada, A. E., Routbort, M. J., DiNardo, C. D., Bueso-Ramos, C. E., Kanagal-Shamanna, R., Khoury, J. D., . . . Patel, K. P. (2019). DDX41 mutations in myeloid neoplasms are associated with male gender, TP53 mutations and high-risk disease. *Am J Hematol*, *94*(7), 757-766. https://doi.org/10.1002/ajh.25486

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., . . . Committee, A. L. Q. A. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*, *17*(5), 405-424. https://doi.org/10.1038/gim.2015.30

Risch, H. A., McLaughlin, J. R., Cole, D. E., Rosen, B., Bradley, L., Kwan, E., . . . Narod, S. A. (2001). Prevalence and penetrance of germline BRCA1 and BRCA2 mutations in a population series of 649 women with ovarian cancer. *Am J Hum Genet*, *68*(3), 700-710. https://doi.org/10.1086/318787

Rogowski, K., van Dijk, J., Magiera, M. M., Bosc, C., Deloulme, J. C., Bosson, A., . . . Janke, C. (2010). A family of protein-deglutamylating enzymes associated with neurodegeneration. *Cell*, *143*(4), 564-578. https://doi.org/10.1016/j.cell.2010.10.014

Rojek, K., Nickels, E., Neistadt, B., Marquez, R., Wickrema, A., Artz, A., . . . Godley, L. A. (2016). Identifying Inherited and Acquired Genetic Factors Involved in Poor Stem Cell Mobilization and Donor-Derived Malignancy. *Biol Blood Marrow Transplant*, *22*(11), 2100-2103. https://doi.org/10.1016/j.bbmt.2016.08.002

Roloff GW, D. M., Godley LA. (2021.). Inherited susceptibility to hematopoietic malignancies in the era of precision oncology. In (Vol. 5, pp. 107-122). JCO Precision Oncology.

Roloff, G. W., Godley, L. A., & Drazer, M. W. (2020). Assessment of technical heterogeneity among diagnostic tests to detect germline risk variants for hematopoietic malignancies. *Genet Med*. https://doi.org/10.1038/s41436-020-0934-y

Romana, S. P., Poirel, H., Leconiat, M., Flexor, M. A., Mauchauffé, M., Jonveaux, P., . . . Bernard, O. A. (1995). High frequency of t(12;21) in childhood B-lineage acute lymphoblastic leukemia. *Blood*, *86*(11), 4263-4269.

Sanders, M. A. (2019). Lifting the veil on germline DDX41 mutations. *Blood*, *134*(17), 1368-1370. https://doi.org/10.1182/blood.2019002982

Scott, R. H., Homfray, T., Huxter, N. L., Mitton, S. G., Nash, R., Potter, M. N., . . . Rahman, N. (2007). Familial T-cell non-Hodgkin lymphoma caused by biallelic MSH2 mutations. *J Med Genet*, *44*(7), e83. https://doi.org/10.1136/jmg.2007.048942

Selvaraj, N., Kedage, V., & Hollenhorst, P. C. (2015). Comparison of MAPK specificity across the ETS transcription factor family identifies a high-affinity ERK interaction required for ERG function in prostate cells. *Cell Commun Signal*, *13*, 12. https://doi.org/10.1186/s12964-015-0089-7

Song, W. J., Sullivan, M. G., Legare, R. D., Hutchings, S., Tan, X., Kufrin, D., . . . Gilliland, D. G. (1999). Haploinsufficiency of CBFA2 causes familial thrombocytopenia with propensity to develop acute myelogenous leukaemia. *Nat Genet*, *23*(2), 166-175. https://doi.org/10.1038/13793

Sorrell, A., Espenschied, C., Wang, W., Weitzel, J., Chu, S., Parker, P., . . . Bhatia, R. (2012). Hereditary leukemia due to rare RUNX1c splice variant (L472X) presents with eczematous phenotype. *Int J Clin Med*, *3*(7). https://doi.org/10.4236/ijcm.2012.37110

Spinner, M. A., Sanchez, L. A., Hsu, A. P., Shaw, P. A., Zerbe, C. S., Calvo, K. R., . . . Holland, S. M. (2014). GATA2 deficiency: a protean disorder of hematopoiesis, lymphatics, and immunity. *Blood*, *123*(6), 809-821. https://doi.org/10.1182/blood-2013-07-515528

Steensma, D. P. (2018). Clinical Implications of Clonal Hematopoiesis. *Mayo Clin Proc*, *93*(8), 1122-1130. https://doi.org/10.1016/j.mayocp.2018.04.002

Sun, W., & Downing, J. R. (2004). Haploinsufficiency of AML1 results in a decrease in the number of LTR-HSCs while simultaneously inducing an increase in more mature progenitors. *Blood*, *104*(12), 3565-3572. https://doi.org/10.1182/blood-2003-12-4349

Tan, C., Dai, L., Yang, W., Li, F., Wang, L., Xiao, Y., . . . Chen, L. (2020). Generation of the human induced pluripotent stem cell line (SHAMUi001-A) carrying the heterozygous c.-128G>T mutation in the 5'-UTR of the ANKRD26 gene. *Stem Cell Res*, *48*, 102002. https://doi.org/10.1016/j.scr.2020.102002

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., . . . Forbes, S. A. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res*, *47*(D1), D941-D947. https://doi.org/10.1093/nar/gky1015

Tawana, K., & Fitzgibbon, J. (2016). Inherited DDX41 mutations: 11 genes and counting. *Blood*, *127*(8), 960-961. https://doi.org/10.1182/blood-2016-01-690909

Tegg, E. M., Thomson, R. J., Stankovich, J. M., Banks, A., Marsden, K. A., Lowenthal, R. M., . . . Dickinson, J. L. (2011). Anticipation in familial hematologic malignancies. *Blood*, *117*(4), 1308-1310. https://doi.org/10.1182/blood-2010-07-296475

Tesi, B., Davidsson, J., Voss, M., Rahikkala, E., Holmes, T. D., Chiang, S. C. C., . . . Bryceson, Y. T. (2017). Gain-of-function SAMD9L mutations cause a syndrome of cytopenia, immunodeficiency, MDS, and neurological symptoms. *Blood*, *129*(16), 2266-2279. https://doi.org/10.1182/blood-2016-10-743302

Thiede, C., Steudel, C., Mohr, B., Schaich, M., Schäkel, U., Platzbecker, U., . . . Illmer, T. (2002). Analysis of FLT3-activating mutations in 979 patients with acute myelogenous leukemia: association with FAB subtypes and identification of subgroups with poor prognosis. *Blood*, *99*(12), 4326-4335. https://doi.org/10.1182/blood.v99.12.4326

Thomas, S. M., Kagan, C., Pavlovic, B. J., Burnett, J., Patterson, K., Pritchard, J. K., & Gilad, Y. (2015). Reprogramming LCLs to iPSCs Results in Recovery of Donor-Specific Gene Expression Signature. *PLoS Genet*, *11*(5), e1005216. https://doi.org/10.1371/journal.pgen.1005216

Tijssen, M. R., Cvejic, A., Joshi, A., Hannah, R. L., Ferreira, R., Forrai, A., . . . Göttgens, B. (2011). Genome-wide analysis of simultaneous GATA1/2, RUNX1, FLI1, and SCL binding in megakaryocytes identifies hematopoietic regulators. *Dev Cell*, *20*(5), 597-609. https://doi.org/10.1016/j.devcel.2011.04.008

Topka, S., Vijai, J., Walsh, M. F., Jacobs, L., Maria, A., Villano, D., . . . Offit, K. (2015). Germline ETV6 Mutations Confer Susceptibility to Acute Lymphoblastic Leukemia and Thrombocytopenia. *PLoS Genet*, *11*(6), e1005262. https://doi.org/10.1371/journal.pgen.1005262

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., . . . Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*, *7*(3), 562-578. https://doi.org/10.1038/nprot.2012.016

Tsang, H. C., Bussel, J. B., Mathew, S., Liu, Y. C., Imahiyerobo, A. A., Orazi, A., & Geyer, J. T. (2017). Bone marrow morphology and disease progression in congenital thrombocytopenia: a detailed clinicopathologic and genetic study of eight cases. *Mod Pathol*, *30*(4), 486-498. https://doi.org/10.1038/modpathol.2016.218

Valenzuela, D. M., Murphy, A. J., Frendewey, D., Gale, N. W., Economides, A. N., Auerbach, W., . . . Yancopoulos, G. D. (2003). High-throughput engineering of the mouse genome coupled with high-resolution expression analysis. *Nat Biotechnol*, *21*(6), 652-659. https://doi.org/10.1038/nbt822

Vasanthakumar, A., Arnovitz, S., Marquez, R., Lepore, J., Rafidi, G., Asom, A., . . . Churpek, J. E. (2016). Brca1 deficiency causes bone marrow failure and spontaneous hematologic malignancies in mice. *Blood*, *127*(3), 310-313. https://doi.org/10.1182/blood-2015-03-635599

Videbaek, A. (1947). Heredity in Human Leukemia and Its Relation to Cancer. In. Einar Munksgaard, Copenhagen.

Wahlster, L., Verboon, J. M., Ludwig, L. S., Black, S. C., Luo, W., Garg, K., . . . Sankaran, V. G. (2021). Familial thrombocytopenia due to a complex structural variant resulting in a WAC-ANKRD26 fusion transcript. *J Exp Med*, *218*(6). https://doi.org/10.1084/jem.20210444

Wang, D., Boerner, S. A., Winkler, J. D., & LoRusso, P. M. (2007). Clinical experience of MEK inhibitors in cancer therapy. *Biochim Biophys Acta*, *1773*(8), 1248-1255. https://doi.org/10.1016/j.bbamcr.2006.11.009

Wang, X., Muramatsu, H., Okuno, Y., Sakaguchi, H., Yoshida, K., Kawashima, N., . . . Kojima, S. (2015). GATA2 and secondary mutations in familial myelodysplastic syndromes and pediatric myeloid malignancies. *Haematologica*, *100*(10), e398-401. https://doi.org/10.3324/haematol.2015.127092

West, A. H., Godley, L. A., & Churpek, J. E. (2014). Familial myelodysplastic syndrome/acute leukemia syndromes: a review and utility for translational investigations. *Ann N Y Acad Sci*. https://doi.org/10.1111/nyas.12346

Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., . . . Rothberg, J. M. (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature*, *452*(7189), 872-876. https://doi.org/10.1038/nature06884

Wlodarska, I., Mecucci, C., Baens, M., Marynen, P., & van den Berghe, H. (1996). ETV6 gene rearrangements in hematopoietic malignant disorders. *Leuk Lymphoma*, *23*(3-4), 287-295. https://doi.org/10.3109/10428199609054831

Wlodarski, M. W., Hirabayashi, S., Pastor, V., Starý, J., Hasle, H., Masetti, R., . . . EWOG-MDS. (2016). Prevalence, clinical characteristics, and prognosis of GATA2-related myelodysplastic syndromes in children and adolescents. *Blood*, *127*(11), 1387-1397. https://doi.org/10.1182/blood-2015-09-669937

Wooster, R., Neuhausen, S. L., Mangion, J., Quirk, Y., Ford, D., Collins, N., . . . Averill, D. (1994). Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science*, *265*(5181), 2088-2090. https://doi.org/10.1126/science.8091231

Wu, Y., & Li, S. (2020). Role of Post-Translational Modifications of cGAS in Innate Immunity. *Int J Mol Sci*, *21*(21). https://doi.org/10.3390/ijms21217842

Xia, P., Ye, B., Wang, S., Zhu, X., Du, Y., Xiong, Z., . . . Fan, Z. (2016). Glutamylation of the DNA sensor cGAS regulates its binding and synthase activity in antiviral immunity. *Nat Immunol*, *17*(4), 369-378. https://doi.org/10.1038/ni.3356

Ye, B., Li, C., Yang, Z., Wang, Y., Hao, J., Wang, L., . . . Fan, Z. (2014). Cytosolic carboxypeptidase CCP6 is required for megakaryopoiesis by modulating Mad2 polyglutamylation. *J Exp Med*, *211*(12), 2439-2454. https://doi.org/10.1084/jem.20141123

Zar, J. H. (1999). Biostatistical Analysis. In (4th edition ed., pp. 523). NJ Prentice Hall.

Zhang, M. Y., Churpek, J. E., Keel, S. B., Walsh, T., Lee, M. K., Loeb, K. R., . . . Shimamura, A. (2015). Germline ETV6 mutations in familial thrombocytopenia and hematologic malignancy. *Nat Genet*, *47*(2), 180-185. https://doi.org/10.1038/ng.3177

Zhang, Z., Yuan, B., Bao, M., Lu, N., Kim, T., & Liu, Y. J. (2011). The helicase DDX41 senses intracellular DNA mediated by the adaptor STING in dendritic cells. *Nat Immunol*, *12*(10), 959-965. https://doi.org/10.1038/ni.2091

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., . . . Chanda, S. K. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun*, *10*(1), 1523. https://doi.org/10.1038/s41467-019-09234-6