Winter 12-15-2019

Physiologic and pathologic profiling of clonal variations

Wing Hing Wong
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Molecular Genetics and Genomics

Dissertation Examination Committee:
Todd Druley, Chair
Jamie Blundell
Grant Challen
Meagan Jacoby
John Welch

Physiologic and Pathologic Profiling of Clonal Variations

by
Wing Hing Wong

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

December 2019
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

Finishing this thesis would not have been possible, first and foremost, without the continuing support and guidance from my PI, Todd Druley. His supervision, constant encouragement, logistical support and general advice have been instrumental in shaping my perspective of science, both the fun parts as well as the dreadful political aspects. I have matured a lot working with Todd in the laboratory, both scientifically and personally. Thank you for believing in my work and me when I doubted myself.

I would also like to thank my committee members: Grant Challen, John Welch, Meagan Jacoby and Jamie Blundell for their insightful comments on the projects. Thank you Andrew Young, from whom I learned most of my skills. Thank you Mark Valentine, for your guidance both scientifically and otherwise. Thank you Maggie Ferris for your encouragement, advice, and for coffee breaks. Thank you Spencer Tong for your scientific companionship as my aisle mate, our journey to explore science together was fun. Thank you Nitin Mahajan for your encouragement, for telling me not to worry when I worried too much. Thank you Minori Tamai for lunch break at Chick-fil-a and talking about science and Japanese culture. Thank you Shailendra Maurya for your encouragement. Thank you Kate Alexander for your contagious enthusiasm, and thank you Amelia Bystry for all your help with my science, and for being my first ‘student’. Thank you James Skeath for your kind words, the speech you gave during my interview at WashU played a large part in shaping my decision to come to WashU. Thank you Melanie Relich for all you assistance with regards to graduate school chores. Thank you all my badminton and squash mates, for without you guys, I would probably be depressed. Just Kidding.

Thank you mom and bro for all your love and unwavering trust in me. These past 4.5 years have been really tough, for me and for the both of you. I hope I have not failed your expectations, and I hope I will not fail you in times to come. Thank you dad, even though you are no longer with us, your words were and will be remembered dearly.

Thank you Wuli YY, nuff said. I love you.

Wing Hing Wong (Ahlek)

Washington University in St. Louis

December 2019

ABSTRACT OF THE DISSERTATION

Physiologic and Pathologic Profiling of Clonal Variations

by

Wing Hing Wong

Doctor of Philosophy in Biology and Biomedical Sciences

Molecular Genetics and Genomics

Washington University in St. Louis, 2019

Associate Professor Todd Druley (Chair)

This thesis sought to provide a better understanding of clonality in various malignant and non-malignant settings using a variety of genomic analytical tools. Clonality is pre-defined as the presence of a mixed population of cells in which each sub-population has distinct somatic mutation profile. It is a common feature in cancers where subpopulations of cells arise as a result of independent, yet continual acquisition of somatic mutations. The clonal architecture of cancers can be used as a diagnostic and prognostic biomarker as well as to monitor disease progression or resolution. Besides cancer, clonal variability and expansion is also implicated in various non-malignant settings where somatically acquired mutations play a role in disease ontogeny and contribute to clinical morbidities. A prime example is clonal hematopoiesis of indeterminate potential (CHIP), which is associated with age-related, atherosclerotic cardiovascular disease due to aberrant inflammatory responses from *TET2* mutated hematopoietic clones. Specifically,

these clones increase the disease risk by modulating the interleukin-1β secretion pathway presumably resulting in direct endothelial damage that acts as a nidus for an atherosclerotic plaque.

Given the clinical importance of clonal variability, clonal profiling is a powerful method to study diseases resulting from acquired somatic mutation. However, our understanding of clonal variability in disease is generally limited by, 1) the relatively high error rate of high-throughput, next-generation sequencing (NGS) methods (approximately 2%), which obfuscates the detection of somatic mutations at low variant allele frequencies, 2) a lack of longitudinal data that would allow one to track the evolutionary dynamics of these somatic mutations, and 3) a lack of comprehensive multi-regional sampling, especially in the case of solid tumors that would enable one to define clonal heterogeneity spatially.

In this thesis, we first optimized an error-corrected sequencing (ECS) strategy that has approximately 100-fold higher limit of detection than standard NGS. We then applied ECS to longitudinally survey physiologic clonal hematopoiesis in healthy individuals aged 0 – 24. According to the thresholds defined by CHIP, which is limited to mutations at ≥2% variant allele frequency (VAF), this age group would not be expected to harbor any clonal hematopoietic mutations, but many researchers, including our group considered this information to be the result of technical inability to detect mutations with low variant allele frequency rather than a true absence of mutations. As a result, our group previously used ECS to determine that clonal hematopoiesis (CH) <0.02 VAF was ubiquitous in individuals at middle age, causing us in this thesis to examine and characterize CH in newborns, children, adolescent and young adults. With ECS, we examined the evolutionary dynamics of clonal mutations during normal hematopoiesis from birth to young adulthood and established that 30% of healthy infants were born with clonal hematopoietic somatic mutations in genes associated with leukemia.

Second, after we found that many healthy young individuals harbored potentially pathogenic somatic mutations in blood, we moved on to examine clonal transfer and clonal dynamics in the context of unrelated allogeneic hematopoietic stem cell transplantation (HSCT) where the majority of healthy HLA-matched, unrelated donors is between ages of 20 to 40. Our lab has previously shown that pre-existing hematopoietic progenitors with pathogenic mutations could be selected by chemotherapy and result in therapy-related AML, and as mentioned above, we have demonstrated that a significant percentage of healthy individuals of all ages harbored hematopoietic progenitors with pathogenic mutations. We therefore hypothesized that healthy donors would harbor mutated clones in blood that engraft the recipients, and the process of HSCT presents a potent selection pressure for donor clones. The most compelling finding was that 100% of the donor clones engrafted and the 84% of these clones harbored mutations that were pathogenic Our results also suggested a possible link between these engrafted pathogenic mutations and the development of chronic graft-versus-host disease in the recipients.

We next characterized clonal hematopoiesis in the background of Down Syndrome (DS) where individuals have approximately 150-fold increased risk of developing leukemia. This was done by comparing CH in DS children who were otherwise healthy with those that had developed myeloid leukemia of Down Syndreom (ML-DS). Besides trisomy 21, ML-DS is characterized by mutations in the X-linked transcription factor, *GATA1*, but GATA1 mutations have recently been demonstrated in about 30% of umbilical cord blood samples from DS children suggesting that additional lesions were required for leukemic transformation. We demonstrated that the clonal profiles in ML-DS differ from those in DS children without leukemia. Our results also suggest an alternative route in which *GATA1* mutated clones contribute to leukemogenesis via oligoclonal, cell extrinsic interactions.

Lastly, I investigated the tumor ontogeny of metastatic glioblastoma in a young adult (aged 27) with neurofibromatosis (NF1) using multi-region sequencing on widely disseminated tumor cells across different brain lesions. Glioblastoma in NF1 is rare, and has conventionally been thought to arise as a result of bi-allelic loss of the *NF1* gene. However, by examining the spatial genetic heterogeneity and the tumor phylogeny, our results suggested that the somatic loss of the second *NF1* allele occurred much later during disease progression, and pathogenic mutations in other genes such as *KMT2B* were involved in initial oncogenic transformation instead.

Collectively, these findings have augmented understanding of clonal hematopoiesis from birth through young adulthood, clonal variability in metastatic glioblastoma, and provide a foundational basis for further explorations in establishing causal links between clonal profiles and disease ontogenies.

Chapter 1: Introduction

1.1 Clonality in malignant diseases

In 1985, Secket-Walker[1] defined a clonal population as a group of cells arising from a single somatic cell via mitotic divisions. This phenomenon would consequently result in different lineages of cells harboring dissimilar genetic or cytogenetic abnormalities within the same individual. Based on this, clonality was thus further understood as the presence of mixed population of cells in which each sub-population has distinct profile of somatic alterations[2]. Over the past several decades, different distinct clonal populations of cells had been empirically observed in the laboratory, primarily in cancers[3], hence supporting the landmark perspective by Peter Nowell (1976)[4] on cancer as an evolutionary outcome driven by sequential acquisition of somatic alterations and clonal selection. Since then, clonality had been regarded as an important feature in cancers due to its direct influence in tumorigenesis. Some of the first demonstrations of clonality in cancers utilized differential patterns of X-chromosome inactivation in female cancer patients who displayed heterozygosity for certain X-linked polymorphisms such as the production of different glucose-6-phosphate dehydrogenase (*G6PD*) isoenzymes by cancer cells in uterine leiomyomas and teratomas[5-7]. This approach was later used by Fialkow and colleagues[8-10] in several classic works characterizing the monoclonal origin of chronic myelocytic leukemia (CML) and Burkitt's lymphoma, as well as the polyclonal origin of neurofibromas. These findings provided some of the first lines of evidence to suggest that cancer was a genetic disease likely driven by clonal selection of specific phenotype(s) and laid the foundation for the field of cancer genetics[11]. The subsequent developments of Maxam-Gilbert[12] and Sanger[13] sequencing methods further bolstered these concepts by providing researchers the initial means to begin asking questions about clonality and the evolutionary processes of tumorigenesis at the level of individual nucleotides[14,15]. These foundational results, coupled with the completion of the first human reference genome[16,17] in the early 21st century propelled us into an new era of cancer genomics. Some early large-scale Sanger sequencing studies leveraged the human reference genome to provide characterization of thousands of somatic alterations in colorectal and breast cancers[18,19], and demonstrated genetic heterogeneity in these cancers. It was also suggested then that each individual mutation in putative driver genes was associated with non-uniform fitness advantage, and cancer cells harboring specific mutations would experience clonal expansion to drive disease progression in a patient[20].

The subsequent advent of next-generation sequencing[21] (NGS) in the following years reduced the cost of sequencing by approximately 100-fold, and democratized DNA sequencing by moving the technical capacity into small research laboratories. This technology enabled the first whole genome sequencing study of a cytogenetically normal adult *de novo* acute myeloid leukemia (AML)[22]. These developments led to a subsequent boom of studies that characterized genetic heterogeneity and clonal variability in cancers[23-26], and the field began to embrace and appreciate the clonal complexity of cancers. Focusing on liquid cancer, a recent study by Welch and colleagues (2012)[27] elegantly described the *in vitro* clonal architectures of M1 and M3 classified AML where they found that the AML samples were mostly oligoclonal at diagnosis with a majority of somatic mutations being benign and irrelevant for pathogenesis. In particular, they also showed that aged-matched healthy hematopoietic stem and progenitor cells (HPSCs) had similar mutation burden as the AML cells, suggesting that these *de novo* AML cases did not arise from an elevated intrinsic mutation rate but rather from positive selection on specific somatic alterations in genes such as *DNMT3A*, *TET2* and *IDH1*[28-30]. Similar findings were also reported in therapy-related AML where pre-existing *TP53* mutated clones were positively selected following chemotherapy to treat various primary diseases in the patients[31].

Unlike liquid cancers, solid cancers have an additional layer of clonal complexity in the form of spatial distributions, and this led to many studies examining the spatial heterogeneity and clonal diversity in these physically confined lesions via laser capture microdissection and sequencing[32]. One well-designed study sequenced multiple resected regions within single lesions of colon adenocarcinoma *in vivo* and evaluated different patterns of intra-tumoral heterogeneity[33]. Using multi-region sequencing, they demonstrated that the these colorectal cancers primarily exhibited neutral evolution where tumor growth resulted from a single expansion from the primordial tumor cells producing an intermixed sub-populations with distinct mutation profiles that were not subjected to stringent selection, a phenomenon termed the "Big Bang" model of tumor growth[34,35]. In other types of cancers such as renal cell carcinomas and multifocal lung cancers, distinct heterogeneous sub-populations of cancer cells often harbor different private mutations that map to a few specific key biological processes such as the MAPK and mTOR signaling pathways, suggesting a pattern of convergent evolution in malignant transformation in these cancers[36-38].

Besides tumorigenesis, perhaps one of the other most important implications of clonality in cancers is that it would result in resistance to treatment[3,39,40]. Regardless of the mode of evolution in treatment-naïve cancers, most therapies impose strong selective pressures that can lead to clonal selection and expansion of pre-existing resistant subclone[35]. For example, recent deep sequencing studies of diagnosis/relapse pairs identified relapsed AML as having arisen from minor subclones that survived treatment[23,41,42], and these relapsed cases typically harbor mutations in epigenetic regulators such as *DNMT3A*, *TET2* and *IDH1*, again suggesting a potential evolutionary convergence in treatment resistance. These relapsed cells were later shown to have originated from either rare leukemia stem cells or immunophenotypically committed subclones that retained stemness properties[43]. Other studies in *EGFR*-mutated solid cancers showed similar evolutionary processes[44,45]. Taken together, these results validated the fundamentals of cancers as an adaptive Darwinian system[46,47].

Due to its involvement in malignant transformation as well as disease progression, clonality has been proposed as a form of clinical assessment to stratify risk of disease and disease-related complications in cancer patients. The utility of clonality in the clinical setting is most markedly demonstrated in prognostic minimal/measurable residual disease (MRD) monitoring in leukemia patients after induction chemotherapy. As previously described, a growing body of evidence finds that certain subclonal mutations in leukemia survive treatment and spawn relapse[23,41]. The central idea of MRD monitoring is to identify surviving leukemia cells, thereby predicting relapse and post-treatment survival. In light of this, different modalities of high and low resolution MRD detection in leukemia such as flow cytometry, qRT-PCR for known fusions or cytogenetics[48-50] have been developed. In acute lymphoblastic leukemia (ALL), clonal B- or T-cell surface receptors enable flow cytometry to have a limit of detection of 1:10,000 leukemia cells, and for every order of magnitude more residual disease, the likelihood of overall survival decreases significantly[51]. The current gold standard modality for MRD detection in AML is multi-parameter flow cytometry (MPFC)[48] which identifies leukemic cells by either a leukemia-associated immunophenotype (LAIP) or a “different from normal” profile of surface markers not present in normal blood cells and leading to the inference of residual AML[52]. These methods offer a limit of detection of roughly 1:1000-2000. In pediatric AML, detecting residual disease above 1:1000 leukemic to normal cells after end-of-induction treatment via MPFC was associated with a significantly increased risk of relapse and a lower overall survival[53,54]. In general, flow cytometry – regardless of sensitivity – only offers a binary "yes/no" readout. For AML specifically, the limit of detection of MPFC is lower due to varying surface immunophenotypes in AML cells and a lack of genetic information that could facilitate detailed examination into clonal architecture[55,56]. While not standard practice and not currently approved by the FDA, newer sequencing based molecular MRD approaches have been developed[49,50]. Studies utilizing sequencing data had indeed demonstrated that relapse arose from a minor subclone in B-ALL[57]. In AML, it was also shown that clonal clearance of mutations in genes such as *ASXL1*, *CEBPA*, *FLT3*, and *DNMT3A* was significantly associated with lower risk of relapse and better prospect of overall survival[58,59]. Clonality assessment is similarly valuable to provide diagnostic information in patients, particularly those with leukemia. Several independent groups established that many cancer-free individuals, particularly those in their 60s, harbored a diverse array of leukemia-associated mutations in the hematopoietic compartments, and they were associated with a significant increase risk of developing hematologic malignancies later in life[60-62] – a condition called clonal hematopoiesis of indeterminate potential (CHIP)[63] or age-related clonal hematopoiesis (ARCH)[64]. Follow-up studies showed that rare hematopoietic clones with specific mutations in genes such as *TP53, IDH1* and *DNMT3A* (in particular R882H, which acts as a dominant negative[65]) significantly increased the likelihood of developing AML in otherwise healthy individuals with a median latency of approximately 10 years[28-30].

In contrast to leukemia, recent advances in sequencing of "liquid biopsies" via cell-free, circulating tumor DNA (ctDNA) for MRD have also been employed to monitor for relapse in solid cancers[66]. One study in breast cancer used mutations identified in ctDNA to show that the cancer cells underwent constant clonal selection throughout treatment, resulting in subclonal mutation in *RB1*, a potent cell cycle regulatory, which increased to relative high frequency at relapse[67]. In general, MRD detection by ctDNA correlates with worse outcomes in patients with various solid cancers[66-69]. In addition, the clonal architecture of colorectal cancer has been examined via phylogenetic approaches to identify early from late metastases in lymph nodes and distant organs[70,71]. The data showed that majority of distant metastatic lesions were not sequentially seeded by lymph node metastasis[70], but instead shared a common origin with lymph node metastasis[70,72]. Metastasis-associated mutations were also identified in subclones at diagnosis and were found to collectively converge on a few key signaling pathways such as *WNT*, *TP53* and *EGFR*[71]. This has profound implications for improving clinical outcomes. Patients who harbor mutations that confer a high risk of metastasis may benefit from adjuvant therapies that target micro-metastatic disease[73]. Likewise, malignancy-specific mutations found in ctDNA also offers a practical avenue as early detection markers in diagnosis[74,75].

Taken together, the clonality of malignant diseases provides rich information that augments our understanding of disease etiology and serves as biomarkers to improve clinical detection, surveillance, therapeutic selection and ultimately, outcomes.

1.2 Clonality in non-malignant disease settings

Whereas clonality generally implies cancer[76], the concept of clonality has been historically associated with various non-malignant physiological processes and disorders in hematology[77,78]. This is mainly due to the study of adaptive immunity, where active cycling of hematopoietic stem and progenitor cells gives rise to clonal populations of renewed stem cells or terminal cells with specific genotypes or phenotypes[79,80]. For example, in cell-mediated adaptive immunity, clonal T-cells with unique surface receptors are produced in response to various infectious or immunologic threats. The process of T-cell maturation involves directed somatic alterations of the T-cell receptor (TCR) in order to produce a TCR specific to a given antigen[81-83]. Similarly, clonal expansion of B-cells with specific functional mutations in the B-cell receptor gene could result in viral-mediated lymphoma (e.g. Burkitt's lymphoma secondary to Epstein-Barr infection) or autoimmunity[84-86]. Overall, abundant evidence demonstrates that physiological immune responses are clonal[76], these processes are usually reactive in nature, and are rarely associated with malignancies.

As recent as 2014, CHIP as a benign physiological process was proposed and demonstrated in individuals without any form of detectable hematologic disorders via the use of NGS[28-30]. Along with *DNMT3A*, *TET2* followed by *ASXL1*, which are frequently found to be mutated in adult *de novo* AML, are the most recurrently mutated genes in CHIP and demonstrating that CHIP represents an early precursor of malignant leukemic transformation, forming a bridge between malignant disease and non-malignant precursors[87] by increasing the risk of AML by 0.5-1.0%/year[60-62]. However, there is a growing list of pathophysiologic, non-malignant implications of CHIP. Perhaps the most representative example entails the clonal expansion of *TET2* mutated hematopoietic clones and their contribution to cardiovascular diseases[87]. Individuals with *TET2* mutations in hematopoietic cells are at a significantly higher risk of atherosclerosis, resulting in myocardial infarction and ischemic stroke[88,89], due to aberrant pro-inflammatory responses involving interactions between clonal leukocytes and vascular endothelium in the IL-1β/NLRP3 pathway[90]. Additionally, CHIP has been implicated to contribute to unexplained cytopenias[91] and related-donor, allogeneic transplant-related morbidities[92] during hematopoietic stem cell transplantation (HSCT). At the time of stem cell donation, otherwise healthy donors, without strong selective pressures other than aging, are likely to possess hematopoietic clones[93]. The recipient, however, has been subjected to months of systemic radiotherapy, chemotherapy, immunosuppression and antibiotics, which can result in strong selective pressures for a recovering bone marrow. Donor stem cells harboring mutations that confer self-renewal or growth advantages are unknowingly transferred from donor to the recipient, and the mutated donor-derived hematopoietic clones may be positively selected and preferentially engraft the recipient, contributing to various acute or chronic morbidities. Several retrospective studies looking at isolated cases of post-HSCT complications have indeed shown that donor HSC clones harboring low VAF mutations in *IDH2* and *DNMT3A* have undergone clonal expansion in the HSCT recipients, causing transplant-related leukemia after months to years of latency[94]. And strikingly, a recent study by Frick and colleagues (2019)[92] demonstrated that the presence of CHIP (in particular *DNMT3A* mutated clones) in related donors, who are typically older siblings aged 50-70 years, was correlated with chronic graft-versus-host-disease (GvHD) in recipients, possibly via inflammatory responses induced by the mutated clones engrafted from donors.

1.3 Clonality with respect to various disease states in pediatric and adolescent/young adult

Notwithstanding the usefulness of clonality assessment, there are several extrinsic limitations that prevent us from harnessing the full potential of clonality to derive insights into disease ontogenies, and to inform clinical practices especially in the pediatric and adolescent/young adult (AYA) age groups. Even with the advances in NGS, most of our current understanding of clonal architecture in genetic diseases, particularly in cancers, is limited to subclones with >0.02 VAF (or 1 in 100 cells for heterozygous mutations), which is derived due to the inherent error rates of 0.005-0.02 of most sequencing platforms[95]. In other words, whole-genome or whole-exome sequencings would not be able to reliably reveal clones with VAF <0.02. Recognizing this issue, several modified sequencing strategies had been developed to circumvent the sequencing error rate[96-98]. These methods, collectively termed error-corrected sequencing (ECS), utilize molecular indexing: each molecule or genomic fragment in the sequencing library would be tagged with a unique molecular identifier (UMI) that is specific to that molecule. Raw sequencing reads sharing the same UMIs would be grouped together in order to create a consensus sequence, and sequencing artifacts would be removed in the process. ECS enables us to detect somatic mutations at 2 orders of magnitude below the sequencing error rate. Using ECS, Young and colleagues (2016)[93] found that clonal hematopoiesis with low VAF mutations was ubiquitously present in almost every individual in their 50s-60s. Given the establishment of 0.02 as the benchmark of CHIP, several investigators openly doubted whether finding such rare events was clinically significant. However, several follow-up studies using ECS showed that hematopoietic mutations as rare as 0.005 VAF are clinically informative as risk factors for subsequent leukemia transformation and relapse[30,99], reinforcing the conclusion that the threshold of 0.02 for CHIP was not a biological tipping point, but rather a technical limitation of the methodology utilized. As a result, adoption of ECS in cancer research and early detection of diseases is gaining traction[100,101] among the scientific community, but there is still a large gap in knowledge with regards to our current understanding in the true complexity of clonality. And since more research emphasis in the field in general have been placed in the older adult cohorts where incidence of cancers is the highest, our understanding of clonality in various disease states in younger individuals is wholly inadequate. As the clinical significance of rare clonal mutations continues to be demonstrated and technologies continually improve detection of more rare clonal and subclonal events, a distinction between normal, physiologic clonal variability and disease-associated clonal mutation and selection is imperative for clinical progress.

Against that context, much of the work presented in this thesis is aimed at characterizing the physiologic clonal hematopoietic profile in various subsets of individuals and compare against cancer in those populations. Hence, in this thesis, we first examined and characterized clonal hematopoiesis in individuals who were aged 0 through 24 using ECS. With this information, we further aimed to gain an understanding of the clonal architecture in AYA individuals as they are involved in several clinically important practices such as hematopoietic stem cell donation for transplantation[92] where there is strong clonal selection, and risk stratification for pediatric diseases by serving as baseline background profiles. In addition, we examined the hematopoietic clonal architecture in Down syndrome children with or without myeloid leukemia because these children were at higher risk of developing myeloid malignancies. Lastly, using multi-region intra-tumoral sampling as advocated previously, we examined the disease ontogeny of neurofibromatosis type 1 (NF1) associated glioblastoma. In summary, this thesis sought to gain a better understanding of physiologic and pathologic clonality of various disease states in the pediatric and AYA cohorts using appropriate techniques, an undertaking which is imperative, but often overlooked.

Chapter 2: Error-corrected sequencing for rare event detection

Note: This chapter has been published in *Journal of Visualized Experiments* as of thesis submission (Wong *et al.* 2018)[102]. I conceptualized the manuscript with the help of Mr. R. Spencer Tong and Dr. Todd Druley. I wrote approximately 85% of the published manuscript, which included two-third of Methods, whole of Introduction, Results and Discussions, and I generated the figures. Materials from the manuscript were re-formatted and re-used in writing this chapter. Parts that were not written by me have been removed from my thesis.

2.1 Introduction

As we age, exposure to mutagens and stochastic errors during cell division result in the accumulation of somatic aberrations in the genome, and this underlies the fundamental pathogenesis of malignant transformation, neuro-developmental diseases, pediatric disorders and normal aging[103,104]. Somatic mutations with disease driving potential are important diagnostic and prognostic biomarkers for early detection and risk management[105-107]. In order to better understand physiologic clonogenesis, which will inform clinical and research decisions, the accurate quantification and characterization of these mutations is of primary importance. In this regard, the advent of next-generation sequencing (NGS) has revolutionized life science research over the past decade by providing various means for high-throughput genomic and epigenomic characterization. Specifically in biomedical research, NGS has been extensively used to examine germline (mosaicism) and clonal somatic variants in heterogeneous DNA samples[108,109];

however, NGS is limited to identifying mutations at >0.02 variant allele fraction (VAF) — due to the inherent error-rate of ≤2.0% (or 0.02) of most sequencing platforms[95]. As a result, tracking diagnostically and prognostically significant somatic variants at lower VAF cannot be achieved using standard NGS. The clinical significance of rare mutations with low VAF (i.e. 0.005) has recently been demonstrated in AML prediction[30], monitoring of minimal residual disease (MRD) and detection of chimerism in hematopoietic stem cell transplantation (HSCT)[99], thus highlighting a need to increase the sensitivity of high-throughput sequencing detection techniques for somatic mutations.

In light of this, various methods have been developed over the last few years in order to circumvent the error rate of NGS[96-98]. These methods utilize some form of molecular tagging, which enables computational error correction of the resultant sequencing reads. Collectively, these strategies were termed error-corrected sequencing (ECS). In essence, each molecule or genomic fragment in the sequencing library is tagged with a random Unique Molecular Identifier (UMI) that is specific to that molecule. The UMIs are constructed by permutations of a string of randomized nucleotides (usually 8 – 16 N) that are built into standard NGS adapter sequences or added through sequential PCR reactions. A second sample-specific barcode is also integrated into the workflow that enables multiplexing of multiple samples into the same NGS sequencing run. PCR amplification is performed on the molecularly tagged library, and subsequently the library is sequenced. During library preparation, it is expected that errors will be randomly introduced to amplicons during PCR amplification and additional errors (predominantly due to de-phasing of bridge amplification) will be introduced during sequencing. To remove random sequencing errors, raw sequencing reads are first computationally grouped according to individual UMIs. Artifacts from sequencing are not expected to be present in all reads or the same type of mutation with the same UMI at the same genomic position due to the stochastic nature of introduction. True variants will be faithfully recapitulated at the same relative position in all reads that share the same UMI. The artifacts are bioinformatically removed (Figure. 2.1).

Previously, the laboratory had used several different versions of ECS to answer various scientific questions pertaining to CH[30,31,93,98]. In this chapter, I built upon previous work and further optimized the protocol to combine ECS with a custom gene panel using Illumina TruSeq Custom Amplicon chemistry. I put forth a standardized pipeline. I will also discuss ways various parameters could be changed by other users to achieve their desired specificity and limit of detection. The optimized ECS from this chapter was subsequently used to characterize clonality in various physiologic and disease settings in the pediatric and adolescent/young adult (AYA) population in the following Chapters 3 – 5.

2.2 Materials and Methods

2.2.1 Customizing an ECS gene panel for pediatric and AYA

Due to the age-related linear model of somatic mutation acquisition and prior work demonstrating the difference in genes commonly mutated in pediatric versus adult AML[110], we hypothesized that pediatric and AYA individuals would have different clonal hematopoietic profiles compared to older adults. And since genes (i.e. *DNMT3A* and *TET2*) commonly associated with adult *de novo* AML were also recurrently mutated in asymptomatic individuals with clonal hematopoiesis[93], we therefore theorized that genes implicated in pediatric leukemia would be relevant in studying clonal hematopoiesis in the healthy or asymptomatic individuals from the pediatric and AYA groups. To our knowledge, we are presenting the first candidate gene panel designed to quantify clonal and subclonal events in genes associated with pediatric leukemia. Therefore, to design a gene panel that would allow us to query somatic mutations across the relevant age spectrum, we worked closely with the Children's Oncology Group (COG) and the TARGET project to identify genes that were recurrently mutated in pediatric AML[110]. We then created a custom panel combining these pediatric genes with adult AML genes (taken from Illumina TruSight Myeloid Sequencing kit) via the Illumina Concierge service. The resulting panel consisted of 1063 amplicons enriching all or some exons of 80 genes that were applicable to a wide age range (Table 2.1). We opted to include the adult genes because it was then unknown if these genes were as commonly mutated in pediatric AML samples at <0.02 VAF as in adult AML samples[93]. Quality-check experiments with a mean coverage per amplicon of approximately 3000x showed a low dropout rate of 0.19%, and a 96% uniformity (with 0.2x of mean coverage).

2.2.2 Standardized step-by-step protocol

The Illumina panels do not offer UMI that would facilitate error correction, so we have incorporated our own ECS strategy to these panels. Therefore, we made substantial modifications to Illumina's commercial protocol, and the modified protocol was based closely upon a previously published manuscript from the laboratory[93]. Here, I standardized the previous protocol in a step-by-step manner, and provided recommendations on different changes other users could make at various steps to achieve the desired specificity and limit of detection.

1. Hybridization of oligos from gene panels
    a. Hybridize oligos onto genomic fragments following manufacturer's protocol. Use 100 ng – 250 ng of DNA template.

b. Remove unbound oligos following manufacturer’s protocol.

c. Perform extention-ligation following manufacturer’s protocol.

Note: Modifications to the manufacturer’s protocol begin below.

2. Incorporation of UMIs via PCR.

   a. Replace the standard i5 8-nucleotide index sequence with a random string of 16 nucleotides (16N) which serve as the UMI. The Illumina i5 adapter sequences were retained. The 16N UMI i5 adapter sequences can be manufactured through Integrated DNA Technologies using the following input: AATGATACGGCGACCACCGAGATCTACAC(N1:25252525)(N1)(N1)(N1)(N1)(N1)(N1)(N1)(N1)(N1)(N1)(N1)(N1)(N1)(N1)(N1)ACACTCTTTCCCTACACGACGCTCTTCCGATCT.

   b. Prepare PCR mastermix by pipetting the following reagents into a tube of appropriate volume size: 37.5 µL of Q5 Hot Start High-Fideliy 2X Mastermix, 6 µL of 10 µM 16N i5 adapters, 6 µL of i7 adapters (Use different i7 adapters for separate samples for multiplexing), and 22 µL of extension-ligation solution

      Note: The Q5 mastermix replaces the polymerase mastermix provided in the Illumina kit. The Q5 polymerase amplifies the genomic fragment with higher fidelity and fewer introduced errors.

   c. Run PCR program on a thermal cycler using the following parameters: 30 s at 98 °C, 4-6 cycles of 10 s at 98°C, 30 s at 66°C, 30 s at 72°C; 2 minutes at 72°C, and then hold at 4°C.

      Note: The number of cycles depends on the panel size. From our experience, a 4-cycle PCR is sufficient if the gene panel has about 1500 different pairs of gene specific oligos, whereas a panel with 500 - 600 pairs of oligos requires 6 cycles of PCR.

   d. Clean up PCR reactions with magnetic beads (i.e. AMPure XP Beads): Add the PCR reaction to magnetic beads in a modified 1 PCR reaction: 0.75 magnetic bead ratio. Proceed with clean-up protocol following manufacturer's recommendations.

3. Quantify molecules in the amplified libraries using a QX200 ddPCR platform with EvaGreen, and follow manufacturer's manual.

   Note: A primer pair of the standard P5 and P7 is required. Molecules that can be detected using P5 and P7 primers are UMI-tagged. Precise mutation quantification requires strict observance of the number of molecules of each library that are loaded onto the sequencer. To achieve this, quantifying the number of molecules for individual libraries per unit volume is performed using the QX200 droplet digital PCR (ddPCR) platform – quantitative PCR is an alternative option. Following ddPCR analysis, the readout will specify the number of molecules per µL per library.

4. Perform second amplification and then normalize the libraries for sequencing.

   a. After quantification of the number of molecules that are UMI-tagged, calculate the desired number of molecules that should be sent for sequencing. This number will directly influence the limit of detection. The limit of detection can be calculated using the following equation:

      $$L = S/CNA$$

      where L represents limit of detection, S represents sequencing output (e.g. 400 millions reads of a NextSeq High Output run), C represents constant (10), N represents number of samples to be pooled per sequencing run, and A represents number of amplicons in the gene panel. To calculate the number of molecules, multiply L by A.

b. Amplify the desired number of molecules using the following mastermix for the second PCR totaling 50 µL: 25 µL of Q5 Mastermix, 2 µL of P5 Primer (1 µM), 2 µL of P7 Primer (1 µM), and 21 µL of DNA molecules.

c. Run PCR program on a thermal cycler using the following parameter: 30 s at 98 °C; 16 cycles of 10 s at 98 °C, 30 s at 66 °C, 30 s at 72 °C; 2 min at 72 °C; and then hold at 4 °C.

d. Clean up sequencing libraries using magnetic beads (i.e. AMPure XP Beads): Add the PCR reaction to magnetic beads in a modified 1 PCR reaction: 0.75 magnetic bead ratio. Proceed with clean-up protocol following manufacturer's recommendations.

e. Quantify concentration of DNA using a bioanalyzer to determine concentration of the ECS libraries.

f. Pool the libraries in equimolar amounts for sequencing. Refer to Step 4.a.

g. Provide the following sequencing settings to Illumina sequencing platforms (MiSeq, HiSeq, NextSeq or NovaSeq): 2x144 paired-end reads, 8 cycles Index 1 and 16 cycles Index 2.

5. ECS bioinformatics processing and analysis.

   a. Obtain the sample-demultiplexed reads from the sequencer or perform demultiplexing of raw sequence reads into different samples using i7 adapter sequences bioinformatically with a custom script.

b. Trim off the first 30 nucleotides of each demultiplexed read to remove oligo sequences from the gene panel.

   Note: this step is optional. The first 30 nucleotides are usually of low Phred quality, thus removal is recommended.

c. Align reads that share the same UMIs to one another to form read families.

   Note: Researchers can use UMI-aware software such as MAGERI[111] to extract read families. No hamming distance was allowed within the UMI sequence in this experiment to increase the specificity of the method.

d. Perform de-duplication and error-correction using ≥5 read pairs in the same read family. A minimum of 3 read pairs is recommended.

e. During error-correction, compare nucleotide at every position across all reads in the same read family, and generate a consensus nucleotide if there is at least 90% concordance among the reads for the particular nucleotide. Call an N if there is less than 90% agreement for the nucleotide position.

f. Discard consensus reads that have >10% of the total number of consensus nucleotides being called as N.

g. Align all retained consensus reads locally to either hg19 or hg38 human reference genome using researcher's preferred aligner(s) such as Bowtie2 and BWA.

h. Process aligned reads with Mpileup using parameters –BQ0 –d 10,000,000,000,000 to remove coverage thresholds to ensure a proper pileup output regardless of VAF.

i. Filter out positions with less than 1000x consensus read coverage.

Note: Researcher determines the minimum coverage for each nucleotide position arbitrarily, it is recommended to have at least 500x consensus read coverage for downstream analysis

j. Use binomial distribution to call single nucleotide variants (SNVs) in retained data from previous. The binomial statistic will be based on a genomic position-specific error model. Each genomic position is modeled independently after summing out the error rates of all samples for that particular position. Following the example (refer to Sample K in Table 2.2):

   i. Probability of nucleotide profile, P = $\sum Variant\ RF$ / $\sum Total\ RF$

   ii. P-value of observation with binomial distribution

   = 1 – binomial(24, 25911, P)

Note: For each genomic position queried, there would be three possible mutational changes (*i.e.*,A>T, A>C, A>G), and each of which would be represented as background artifact. Somatic events that are significantly different from the background after Bonferroni correction are retained. In the example shown in Table 2.2, the number of tests performed was 11, hence a Bonferroni corrected *p*-value ≤ 0.00454545 (0.05/11) was required to call an event as statistically significant.

2.3 Discussion

To characterize physiologic and disease-associated clonality in the pediatric and AYA populations, we designed a custom candidate gene panel that included genes implicated in both adult and pediatric AML. Here, we demonstrate a method for combining our UMI-tagging strategy with the Illumina TruSeq gene panel to survey for clonal mutations and small insertions/deletions (Indels) as rare as 0.0001 VAF. We should note that our ECS strategy is is robust with various chemistries and not limited to integration with the Illumina kit. We merely started with this kit due to its existing adult AML content.

By mitigating the NGS error rate, ECS enables quantification of somatic mutations at a level of detection 100-fold lower than standard NGS. For instance, detection of the new presence of pathogenic mutations (therefore having low VAF) is imperative to inform early intervention of disease[31,112]. In the clinical management of leukemia, which has been done via flow cytometry to date, the detection of MRD (residual leukemic cells post-treatment) informs risk stratification and is used to inform non-targeted treatment options. Given the binary "yes/no" flow cytometric assessments of MRD, an NGS approach could not only match flow in terms of limit of detection but also provide gene-specific data that further informs the use of targeted therapies. In addition, ECS is applicable to detecting circulating tumor DNA (ctDNA), so-called "liquid biopsy", and evaluate therapeutic response and potential relapse in solid tumor patients by assessing for the presence/absence of ctDNA as well as the burden of specific clonal or subclonal mutations that can inform salvage therapy[113].

As demonstrated in Table 2.2, the power of using binomial distribution-based position-specific error model to call variants depends largely on the number of sequenced libraries as well as the depth of sequencing used to build the error model. The robustness of the error model increases with higher number of samples and more sequencing depth. It is recommended to use at least 10 sequenced samples with an average error-corrected read coverage of 3000x per genome position per sample to build an error profile for each sample. The position-specific approach is similar to MAGERI[111], but instead of using an aggregate error rate for all six different substitution types (A>C/T>G, A>G/T>C, A>T/T>A, C>A/G>T, C>G/G>C, C>T/G>A), we model each substitution independently at every position. For instance, an error rate of C>T at a given genomic position is different from another position. Our approach also takes into account a sequencing batch effect, as the base substitution rate observed in one sequencing run is different from another run. Hence, it is important to model each position for all substitution types especially when samples from different sequencing runs are pooled to build the model.

An important consideration when designing an ECS experiment is the desired detection threshold. The beauty of NGS studies is that they can be readily scaled in terms of genes/targets of interest, detection threshold (dictated by depth of sequencing), and number of individuals queried. For example, if the researchers are interested to find rare mutations in two amplicons with a detection threshold of 0.0001, they can pool maximally 75 samples in a single sequencing run using MiSeq V2 chemistry which outputs up to 15 million reads (2 amplicons * 10,000 molecules * 10 reads for error-correction * 75 samples = 15 million sequencing reads). Researchers can vary the number of molecules going into sequencing or the number of pooled samples in a single sequencing run to adjust the detection threshold. In our studies, we aimed to find mutations with a detection threshold of 0.0001 VAF (1:10,000) using the Illumina gene panel. We routinely use 250 ng of starting DNA to ensure that sufficient molecules are captured in order to achieve the aforementioned detection threshold. Researchers can opt to start with lower amount of DNA (>50 ng is recommended) if the desired detection limit is >0.001 VAF.

As the UMIs are appended onto the i5 indexes, sequencing settings have to be amended accordingly. For example, we used 16 N UMIs, and the sequencing settings were 2x144 paired end reads, 8 cycles of Index 1 and 16 cycles of Index 2 as opposed to the usual 8 cycles of Index

2. The increase in the Index 2 cycle is compensated by a decrease in the total number of cycles allocated to the reads. If researchers opt to use 12N UMIs[114], the settings should be changed to 12 cycles of Index 2.

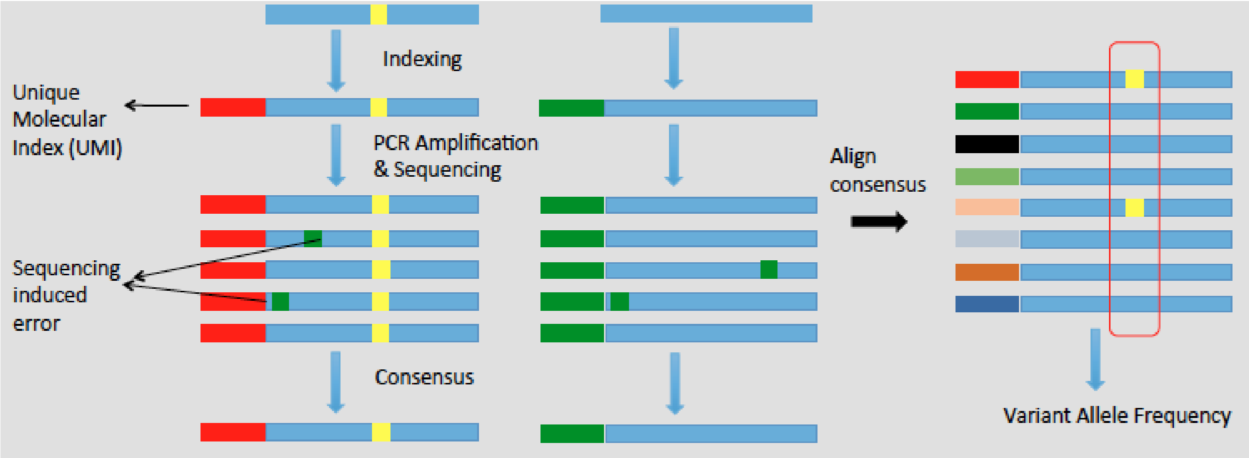


Figure 2.1: An illustration of ECS workflow. Starting on the left, each molecule in the library is tagged with a unique molecular index (UMI), color-coded as red and green bars. The true mutation is colored in yellow. During sequencing, random errors will be introduced across the genomic fragment of interest, whereas the true mutations will be faithfully amplified in all daughter strands (yellow square). After sequencing, reads that share the same UMI will be group together, and a consensus sequence is obtain by removing nucleotide change(s) not observed in all reads. After that, all consensus sequences are aligned to one another, enabling the calculation the VAF of a given nucleotide change (right).

Table 2.1: Recurrently mutated genes in adult and pediatric AML included in the custom gene panel. Genes listed as Adult were taken from the Illumina TruSight Myeloid Sequencing Panel, while the genes listed as Pediatrics were obtained from the studies of pediatric leukemia at Children's Oncology Group.

Adult	ABL1	ASXL1	ATRX	BCOR	BRAF	CALR	CBL
	CBLB	CBLC	CDKN2A	CEBPA	CSF3R	DNMT3A	ETV6
	EZH2	FBXW7	FLT3	GATA1	GATA2	GNAS	HRAS
	IDH1	IDH2	IKZF1	JAK2	JAK3	KDM6A	KIT
	KRAS	MPL	MYD88	NOTCH1	NPM1	NRAS	PDGFRA
	PHF6	PTEN	PTPN11	RAD21	RUNX1	SETBP1	SF3B1
	SMC1A	STAG2	TET2	TP53	U2AF1	WT1	ZRSR2
Pediatric	AFF3	ARFGEF3	ASXL2	CACNA1A	CCND3	COL12A1	CREBBP
	DHX15	DNAH2	DNAH9	FAT1	FLT4	KMT2A	LAMA1
	MED23	MFSD11	MYC	MYH11	MYH2	MYH4	NCOR1
	NF1	NUP205	NUP214	RELN	SETD2	SPI1	SRCAP
	TRIM24	TRRAP	USP34				

Table 2.2: Example demonstrating the way to construct a position-specific binomial error model. 1) take the sum of Variant RFs (read families); 2) take the sum of Total RFs; 3) Calculate probability of observing a nucleotide change (in this case G to A) in the study cohort by dividing the sum of Total RFs by the sum of Variant RFs; 4) Using the probability obtained in step 3, for each sample, calculate the p-value of observing the number of Variants RFs using binomial distribution. In the above example, the p-value of observing 24 Variants RFs out of 35911 Total RFs in Sample K was 2.265E^{-13}.

Sample ID	Chromosome	Genomic Position	Nucleotide Change	Variant RFs	Total RFs	Binomial p-value
A	chr4	106158046	G>A	0	11783	0.698534317
B	chr4	106158046	G>A	0	14855	0.779470039
C	chr4	106158046	G>A	0	21557	0.88850237
D	chr4	106158046	G>A	0	21777	0.89097088
E	chr4	106158046	G>A	0	22502	0.89872544
F	chr4	106158046	G>A	0	24493	0.917299903
G	chr4	106158046	G>A	0	25145	0.922609048
H	chr4	106158046	G>A	0	25731	0.927089294
I	chr4	106158046	G>A	0	27281	0.937728774
J	chr4	106158046	G>A	2	24470	0.453642856
K	chr4	106158046	G>A	24	35911	2.26485E-13
Total				26	255505	

Chapter 3: Longitudinal clonal hematopoiesis in healthy individuals from birth to young adulthood

3.1 Introduction

Clonal hematopoiesis (CH) describes an asymptomatic expansion of blood cells from a HSPC carrying one or more unique mutations, while clonal hematopoiesis of indeterminate potential (CHIP), also known as age-related clonal hematopoiesis (ARCH), is further defined by age-related the presence and expansion of AML-associated hematopoietic somatic mutations at >0.02 VAF in the blood of asymptomatic individuals who are without hematologic diseases[87]. CHIP has been shown to increase the risk of hematologic malignancies by approximately 0.5-1.0% per year[60-62]. The 0.02 VAF threshold for CHIP was not a biological marker, and was largely defined based on the limit of detection for standard NGS without unique molecular indexing (UMIs), which generates systematic sequencing errors at a rate of around 2% (or 0.02 VAF)[95]. Our laboratory has a long-standing interest in investigating rare clonal mutations below this 0.02 VAF threshold for various reasons such as detection of pre-leukemic clones, and surveillance monitoring of leukemia after therapy, so called minimal residual disease (MRD). In light of this, the Druley laboratory had optimized and implemented an ECS strategy that can be coupled seamlessly with Illumina TruSight[93] or TruSeq gene panels to mitigate the error rate of NGS through the introduction of UMIs that would offer a limit of detection of 0.0001 as validated by droplet digital PCR (ddPCR). Three observations using ECS led to the conceptualization of this chapter. First, in adults with *TP53*-mutated therapy-related AML (t-

AML) who developed their leukemia as much as 12 years after their initial diagnosis and treatment for a primary malignancy, the laboratory demonstrated from archived blood and marrow samples that the *TP53*-mutated hematopoietic progenitors pre-existed treatment and were selected by chemotherapy[31]. This is in contrast to the long-held perception that chemotherapy introduces these mutations with leukemogenesis potential. In the same study, it was demonstrated that 40% of healthy senior citizens in their 80s harbored cancer-associated clonal hematopoietic *TP53* mutations without any evidence of hematologic disease. Second, the laboratory expanded this observation by surveying longitudinal blood samples from healthy elderly participants who were in their 50s – 60s in the Nurses' Health Study and found that 95% of these healthy adults possessed mutations in AML-associated genes by their 50s[93]. Third, we also found that rare pre-leukemic mutations with as low a VAF as 0.005 were informative to future risk of leukemia development[30].

Previous studies have focused on elderly individuals in their 50s – 60s or older since CHIP was virtually undetectable in younger populations by standard NGS[60-62]. Therefore, it was presumed that younger populations rarely harbor clonal mutations in their blood. To date, there has not been a longitudinal, systematic characterization of CH below the 0.02 VAF threshold using an ECS approach from birth through young adulthood. Thus our current knowledge of physiological CH below 0.02 VAF in the pediatric and AYA groups is critically lacking. Here, we used our optimized ECS approach as described in Chapter 2 to characterize longitudinal clonal dynamics in 80 genes that are recurrently mutated in adult and pediatric leukemia[110] in 30 random healthy individuals from birth through young adulthood. We expected to see a lower prevalence of CH and different mutation profile of genes compared to older adults[110]. We found CH in 30% of healthy individuals at birth, and we did not identify any recurrently mutated genes across the 30 individuals. Notably, the genes associated with CHIP/ARCH in older adults were not preferentially mutated at young age. In addition, we did not observe any enrichment in C to T transitions which is a characteristic of aged-related mutagenesis[115,116] over all four time points. Many mutations were quiescent either by ages 17 or 24, suggesting that the clones or stem cells harboring these mutations were inactive during hematopoiesis at that specific age[117].

3.2 Materials and Methods

3.2.1 Study population

Longitudinal samples of 30 healthy individuals spanning 4 time points (birth, age 7, 17 and 24) were obtained from the Avon Longitudinal Study of Parents and Children (ALSPAC) at University of Bristol, United Kingdom (total 120 samples). The study was approved under the Institutional Review Board (IRB) #201710007. The ALSPAC enrolled more than 14,000 pregnant women from 1991 to 1992, and longitudinally follows the general health of these women and their children over two decades, and it is still ongoing. The study curates detailed information pertaining to environmental and genetic factors that would affect the health and development of both mothers and children, and it has collected saliva and blood DNA (from white blood cells) samples from the participants at multiple time points. We obtained a total of 250 ng of blood DNA per sample from ALSPAC. The DNA samples were randomized in a 96-well plate. Upon receipt, the DNA samples were stored in -80C freezer until ECS library preparation.

3.2.2 ECS library preparation

A total of 100 ng of DNA per sample was used to prepare an ECS library with a custom-made Illumina TruSeq Custom Amplicon gene panel designed to enrich for 1063 amplicons in some or all exons from 80 genes that are recurrently mutated in adult and pediatric AML, and are also commonly associated with CHIP in adults. The libraries were prepared by following the protocol detailed in Chapter 2[102]. We made several modifications to the protocol for this study. First, after ddPCR quantification (Step 3 in Chapter 2), each library was normalized to 7.1 million UMI-tagged molecules, giving a theoretical mean error-corrected consensus coverage of approximately 6680x per genomic position. Second, at the end of library preparation, six purified libraries were normalized to equimolarity, pooled and sequenced per lane in an Illumina NovaSeq S4 flowcell with the following settings: 2x144 paired-end, 8 cycles Index 1, 16 cycles Index 2 (account for 16N random bases used as UMI). In total, 120 samples were processed.

3.2.3 ECS bioinformatics processing

The data processing follows the pipeline outline in Chapter 2 closely with several modifications. First, after read alignment, the output was filtered to include bases with ≥700x consensus read coverage and within the target regions of the Illumina TruSeq panel and are not common variants (≥0.01 minor allele fraction) identified by the 1,000 Genomes Project[118]. For single nucleotide variants (SNVs), a position-specific binomial background error model was implemented in variant calling. Each genomic position is modeled independently by compiling the background error rate of all samples for that specific genomic position (sum of all variant bases relative to the sum of reference bases). A sample with a number of variant bases at genomic position that is significantly different from the background error-rate based on binomial distribution after Bonferroni correction is considered a positive for that position. Typically, the p-value (after Bonferroni correction) for calling a variant as positive was <0.00000001. After variant-calling, manual curation and several other filters were applied to remove potential artifacts in order to obtain high-confidence variants: 1) germline variants were removed, 2) variants called due to sequencing batch effect were removed, 3) variants identified in more than one individuals at any time point were removed, 4) variants that were present in less than 3 time points of an individual were removed, 5) variants that were deemed alignment artifacts by manual curation via IGV browser[119] were removed. In the end, we retained a set of high-confidence variants by removing potential false-positive calls and common variants that are observed in the general population. VarScan2[120] was used to call insertions/deletions (Indels) with the mpileup2indel setting.

3.3 Results

3.3.1 Variant characterization in rare hematopoietic clones

Our study population consisted of 30 healthy individuals with 4 longitudinal samples collected each from birth, and ages 7, 17 and 24 (120 samples in total). ECS facilitated the investigation of CH at <0.02 VAF in age groups that were previously demonstrated to have essentially no CHIP above 0.02 VAF, and were presumed to have no clinically relevant CH below 0.02 VAF[60-62]. We identified a total of 27 clonal single nucleotide variants (SNVs) that were present in at least 3 time points in 30% (9 of 30) of individuals. We further observed that all of these SNVs were present at birth, and none of these SNVs were associated with malignancies as annotated with the COSMIC database (Table 3.1). We did not detect any Indels that passed our bioinformatics pipeline at any time point in all individuals. The VAFs of the detected SNVs ranged from 0.00054 – 0.02381 with a median of 0.00096 at birth, from 0.00063 – 0.33779 with a median of 0.00119 at age 7, from 0.00052 – 0.35925 with a median of 0.00079 at age 17, and from 0.00061 – 0.32585 with a median of 0.00074 at age 24. When we quantified the different types of mutations called, we detected 5 intronic and 22 exonic clonal mutations with 18 missense, 1 nonsense, and 3 silent SNVs in 21 genes (Figure 3.1). With the exception of the *DNAH2* mutation with >0.3 VAF from age 7 onwards in individual als14, all other SNVs remained at <0.02 VAFs at later time points. One particular individual (als27) displayed a hyper-mutated phenotype, and several genes including *ASXL1*, *BCOR*, *FAT1*, *IKZF1* and *KDM6A* were mutated twice in this individual, but we did not observe any recurrently mutated genes amongst the studied cohort as a whole. In addition, we found that the most commonly observed substitution type in previous studies of CH in older individuals – cytosine to thymine (C to T) transition[60-62,93,] was not enriched in our studied cohort of young individuals. Instead, the SNVs were evenly distributed across substitution types (Figure 3.2).

3.3.2 Temporal dynamics of rare clonal mutations

With longitudinal samples collected at 4 different time points, we were able to examine the temporal dynamics of these detected rare clonal SNVs. All detected SNVs were present at birth among the 9 individuals, and some of these SNVs were undetectable via ECS (with a limit of detection of 0.0005 in this study) at a later time point. Of the 27 SNVs, only 4 SNVs were present in all 4 time points while the rest were present in 3 time points. Of those present in 3 time points, 11 SNVs were stable from birth through age 7, became undetectable via ECS at age 17 but resurfaced at age 24; 11 SNVs were stable from birth through age 17, and became undetectable via ECS at age 24; and 1 was undetectable at age 7 but present in other time points (Figure 3.3). All except 1 of these mutations were relatively stable in terms of VAF across different time points. The exception was a missense SNV in *DNAH2* which saw an increase in VAF from 0.0026 at birth to 0.3378 at age 7, and remained at approximately 0.3 thereafter.

3.4 Discussion

Our findings suggest that low-VAF CH harboring somatic mutations in AML-associated genes is much more prevalent in young populations than previously appreciated. The individuals in our studied age group were previously presumed not to (or rarely) harbor somatic clonal mutations in blood cells at >0.02 VAF[60-62], but using ECS, which offers a limit of detection 2 orders of magnitude lower, we detected somatic mutations present since birth in 30% of individuals, indicating that a substantial proportion of the population were born carrying somatically mutated hematopoietic clones. Although these SNVs were present in genes known to be recurrently mutated in adult and pediatric leukemia, none of the SNVs were mutated at specific genomic positions associated with malignancies. In addition, contrary to studies in adults, we also did not observe an enrichment of mutations in *DNMT3A* or *TET2* that were canonically associated with CHIP/ARCH in adults and future risk of AML development. These results are suggestive that leukemia specific mutations are generally acquired at a later age[64,121]. The corollary to this conclusion, which remains to be further studied, is whether children who develop pediatric leukemia, the most common form of malignancy in childhood, derive their disease from stochastic clonal mutations at base positions that confer a functional impact on the resulting proteins. We observed that the most common mutation signature of CHIP in older adults – spontaneous deamination of methylated cytosine into thymine (C to T) was not enriched in the young individuals, and the mutation types were evenly distributed. The C to T signature reflects a cell-cycle dependent mutation clock and is associated with the rate of replicative turnover[116,122]. Due to the young age of the individuals in this study, the number of cell cycle that a given number of stem cells have undergone is presumably low.

Notably, we showed that these SNVs were stable longitudinally, and were present at multiple time points at relatively consistent VAFs across a period of 20 years. This suggests that the SNVs were likely present in the long-term HSCs[93]. The clonal dynamics in the form of 'disappearance' and subsequent resurfacing of clonal mutations likely indicates that the HSCs underwent entry into and exit out of cell cycle over time[123]. With the exception of the *DNAH2* mutation in individual als14 that underwent a rapid expansion, increasing its VAF from 0.0026 at birth to 0.3378 at age 7 and thereafter, the other SNVs did not show strong evidence that they were selected for since their VAFs stayed relatively flat across 20 years. The gene *DNAH2* encodes for a heavy chain of axonemal dynein which is part of the microtubule-associated motor protein complex. Since this individual was recorded as healthy at the time of sample collections at all 4 time points, the sudden and rapid expansion of *DNAH2* has unknown implications in the health and development of this individual at the conclusion of this study. Previously in another study, it was shown that *PPM1D*-mutated clones expanded in response cytotoxic treatment, and these clones had little leukemogenic potential[124]. It is possible that the expansion of this *DNAH2*-mutated clone was in response to an extrinsic stressor perhaps in the form of infection or inflammatory response not unlike *PPM1D*-mutated clones, but the expansion itself had little implication in leukemia development.

Due to the technical limitations of our ECS pipeline, we likely under-called the number of SNVs in our studied cohort[93]. Specifically, we required that a SNV be observed in at least 3 time points of an individual to be considered positive. This increased the specificity of our pipeline, but we had potentially missed true biological mutation(s) that were present in two time points or less. Lowering the stringency of the pipeline was not desirable as we observed a high number of potential false positive calls. In order to mitigate this issue, an independent library should be prepared and sequenced, and we should retain SNVs called in both replicates.

In summary, we demonstrated that the prevalence of CH in young populations was much higher than previously appreciated, and a sizable proportion of babies were born carrying mutated hematopoietic clones. The clinical significance of these mutations were not known, though at this stage they were presumably benign since the studied individuals were healthy at the time of sample collection. However, these mutated clones might contribute to future risk of diseases including leukemia and inflammatory disorders. Previous studies had correlated the risk of leukemia development with CH of any mutations at >0.005 VAF in older individuals who were in their 50s[29,30]. Although the mutations detected in individuals below 20s in this study were relatively lower in VAF compared to older adults, one would expect clonal outgrowth either by drift or selection given sufficient time[125]. Our results warrant a further study to detail the eventual clinical outcomes of these young individuals, and to correlate the outcomes with the mutation profiles detected at young age or even at birth. If these two variables were to be positively and significantly correlated, we could then look into establishing a CH clinic to identify the individuals most at risk for certain hematologic diseases decades before clinical diagnosis, and to take preventive steps[126]. The ALSPAC study which will soon have another longitudinal sample collection at age 30 of enrolled participants as of this thesis submission, is well positioned to provide the necessary resources and sample cohorts to answer this question. In addition, the absence of enrichment of *DNMT3A* and *TET2* mutated clones with known leukemogenesis potential raises questions regarding the age window in which these genes are mutated and the developmental significance of that age window. Future research should focus on identifying this age window (likely between ages 20 to 40), and the factors contributing to mutations in *DNMT3A* and *TET2* (e.g. Are the individuals born with CH more prone to further mutations?). For individuals who had CH at birth, it is presumed that mutagenesis happened *in utero*[127]. One utility stemming from this information is genetic counseling for couples hoping to have children by surveying fetal blood for possible disease-associated somatic mutations. Furthermore, the presence of CH in a substantial proportion of healthy young individuals may have an impact in clinical settings such as hematopoietic stem cell transplantation where the majority of unrelated blood donors are relatively young at ages 18-44[128]. Physicians might unwittingly transfer mutated hematopoietic clones of unknown clinical significance from otherwise healthy donors to the recipients. Since the process of transplantation itself could act as a potent selection pressure, certain mutated hematopoietic clones might be conferred an advantage at survival proliferation, and these clones might consequently cause donor-derived complications in recipients. The clinical impacts of the presence of mutated donor clone(s) in HSCT, especially in unrelated HSCT are poorly known, and this was probably due to the technical limitations of standard NGS that precluded comprehensive examination of CH in young healthy donors. With the availability of an ECS approach, it now seems ripe to explore this question of clinical importance.

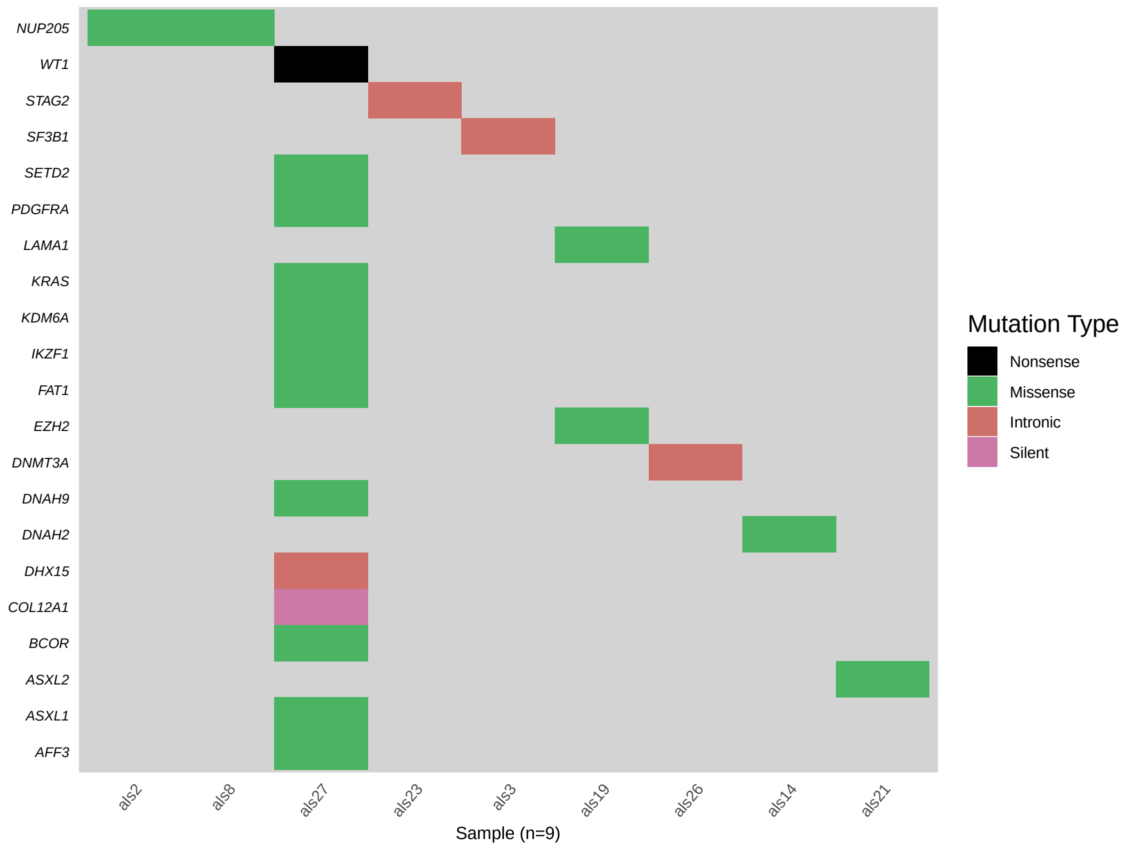


Figure 3.1: Mutation spectrum of CH in ALSPAC samples at birth. A total of 30% individuals (9 of 30) were found to have at least one somatic mutation in the blood cells in the 80 genes queried. Most mutations were found to be missense. An individual (als27) displayed a hyper-mutated phenotype in the hematopoietic compartment.

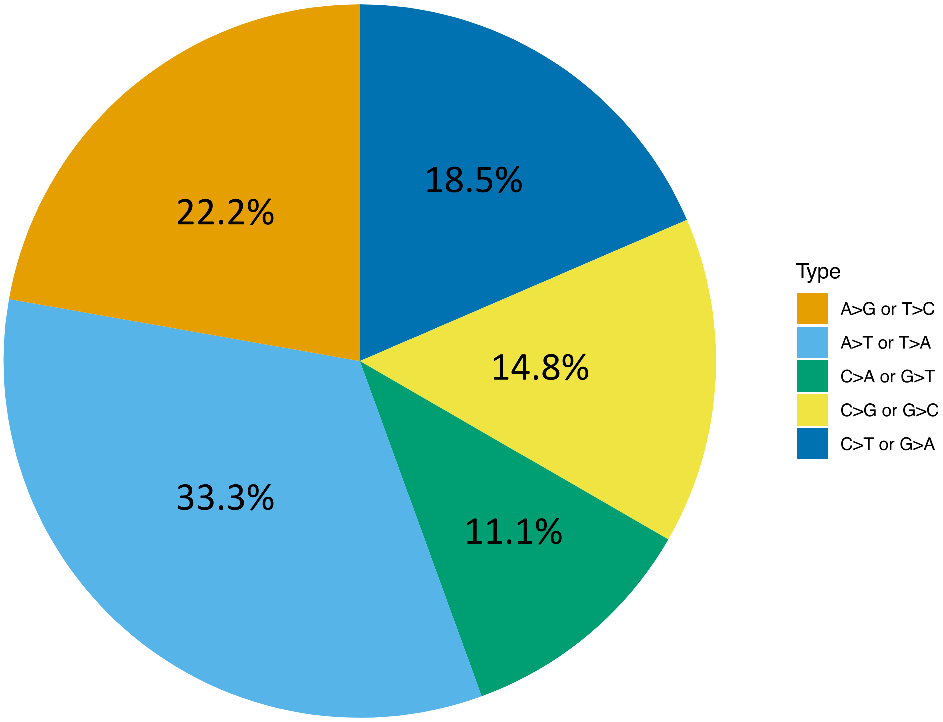


Figure 3.2: Distribution of types of nucleotide changes. The canonical enrichment of C to T transitions in CHIP of adults was not observed in the ALSPAC individuals at birth. Different types of nucleotide changes were evenly distributed.

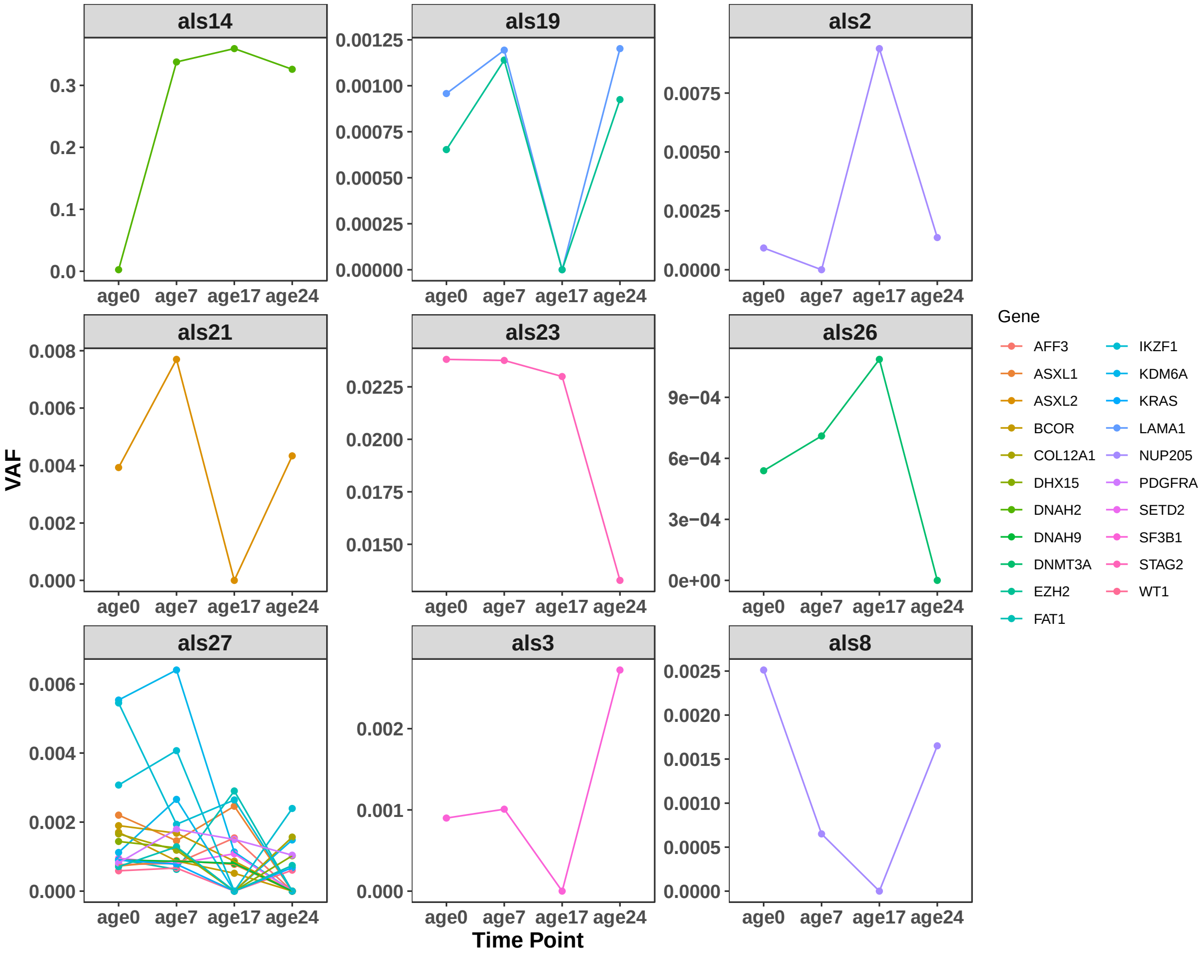


Figure 3.3: Longitudinal detection of somatic mutations in blood cells in the ALSPAC individuals. Somatic mutations were detected in at least 3 time points in 9 individuals. The VAFs of the detected mutations were plotted on the y-axis, and the corresponding time points were plotted on the x-axis. Individual als27 displayed a hypermutated phenotype. With the exception of DNAH2 mutation in als14 which increased from 0.002 VAF (birth) to 0.3 VAF (age 7 and thereafter), all other mutations remained fairly consistent in terms of VAFs.

Table 3.1: Somatic mutations identified in the ALSPAC individuals by ECS using the 80-gene targeted sequencing panel.

patientID	Time	VAF	Chr	Start	Stop	Ref	Mut	Gene	Type
als27	age0	0.0022	chr20	31024434	31024434	T	A	ASXL1	missense
als27	age7	0.001457	chr20	31024434	31024434	T	A	ASXL1	missense
als27	age17	0.00246	chr20	31024434	31024434	T	A	ASXL1	missense
als3	age0	0.000899	chr2	198267645	198267645	T	C	SF3B1	intronic
als3	age7	0.001008	chr2	198267645	198267645	T	C	SF3B1	intronic
als3	age24	0.002721	chr2	198267645	198267645	T	C	SF3B1	intronic
als27	age0	0.000917	chr4	187629162	187629162	T	C	FAT1	missense
als27	age7	0.000629	chr4	187629162	187629162	T	C	FAT1	missense
als27	age17	0.0029	chr4	187629162	187629162	T	C	FAT1	missense
als27	age0	0.000588	chr11	32413596	32413596	G	A	WT1	nonsense
als27	age7	0.000669	chr11	32413596	32413596	G	A	WT1	nonsense
als27	age24	0.000607	chr11	32413596	32413596	G	A	WT1	nonsense
als27	age0	0.005533	chrX	44941911	44941911	G	A	KDM6A	intronic
als27	age7	0.006403	chrX	44941911	44941911	G	A	KDM6A	intronic
als27	age17	0.001133	chrX	44941911	44941911	G	A	KDM6A	intronic
als27	age0	0.000768	chr2	100217924	100217924	G	C	AFF3	missense
als27	age7	0.000806	chr2	100217924	100217924	G	C	AFF3	missense
als27	age17	0.001542	chr2	100217924	100217924	G	C	AFF3	missense
als2	age0	0.000921	chr7	135285705	135285705	G	C	NUP205	missense
als2	age17	0.00939	chr7	135285705	135285705	G	C	NUP205	missense
als2	age24	0.001364	chr7	135285705	135285705	G	C	NUP205	missense
als19	age0	0.000958	chr18	6958484	6958484	G	C	LAMA1	missense
als19	age7	0.001194	chr18	6958484	6958484	G	C	LAMA1	missense
als19	age24	0.001202	chr18	6958484	6958484	G	C	LAMA1	missense
als27	age0	0.001896	chrX	39933273	39933273	G	T	BCOR	missense
als27	age7	0.001678	chrX	39933273	39933273	G	T	BCOR	missense
als27	age17	0.000867	chrX	39933273	39933273	G	T	BCOR	missense
als27	age0	0.000951	chr3	47163685	47163685	G	T	SETD2	missense
als27	age7	0.000798	chr3	47163685	47163685	G	T	SETD2	missense
als27	age17	0.001082	chr3	47163685	47163685	G	T	SETD2	missense
als27	age0	0.001706	chrX	39923753	39923753	C	A	BCOR	missense
als27	age7	0.000872	chrX	39923753	39923753	C	A	BCOR	missense
als27	age17	0.000518	chrX	39923753	39923753	C	A	BCOR	missense
als27	age0	0.000725	chr20	31022999	31022999	C	G	ASXL1	silent
als27	age7	0.000888	chr20	31022999	31022999	C	G	ASXL1	silent
als27	age17	0.000775	chr20	31022999	31022999	C	G	ASXL1	silent
als19	age0	0.000653	chr7	148529782	148529782	C	T	EZH2	missense

als19	age7	0.00114	chr7	148529782	148529782	C	T	EZH2	missense
als19	age24	0.000925	chr7	148529782	148529782	C	T	EZH2	missense
als21	age0	0.003931	chr2	25965982	25965982	C	T	ASXL2	missense
als21	age7	0.007699	chr2	25965982	25965982	C	T	ASXL2	missense
als21	age24	0.004338	chr2	25965982	25965982	C	T	ASXL2	missense
als27	age0	0.001119	chrX	44929034	44929034	A	G	KDM6A	missense
als27	age7	0.002658	chrX	44929034	44929034	A	G	KDM6A	missense
als27	age24	0.001482	chrX	44929034	44929034	A	G	KDM6A	missense
als27	age0	0.001656	chr6	75834077	75834077	A	G	COL12A1	silent
als27	age7	0.001186	chr6	75834077	75834077	A	G	COL12A1	silent
als27	age24	0.001565	chr6	75834077	75834077	A	G	COL12A1	silent
als27	age0	0.000902	chr17	11872785	11872785	A	G	DNAH9	missense
als27	age7	0.000865	chr17	11872785	11872785	A	G	DNAH9	missense
als27	age17	0.000804	chr17	11872785	11872785	A	G	DNAH9	missense
als27	age0	0.000899	chr12	25380190	25380190	A	G	KRAS	missense
als27	age7	0.000779	chr12	25380190	25380190	A	G	KRAS	missense
als27	age24	0.00068	chr12	25380190	25380190	A	G	KRAS	missense
als27	age0	0.001439	chr4	24578381	24578381	A	T	DHX15	intronic
als27	age7	0.001254	chr4	24578381	24578381	A	T	DHX15	intronic
als27	age24	0.001023	chr4	24578381	24578381	A	T	DHX15	intronic
als26	age0	0.000539	chr2	25469754	25469754	A	T	DNMT3A	intronic
als26	age7	0.00071	chr2	25469754	25469754	A	T	DNMT3A	intronic
als26	age17	0.001087	chr2	25469754	25469754	A	T	DNMT3A	intronic
als27	age0	0.000708	chr4	187541008	187541008	A	T	FAT1	missense
als27	age7	0.001282	chr4	187541008	187541008	A	T	FAT1	missense
als27	age24	0.000743	chr4	187541008	187541008	A	T	FAT1	missense
als27	age0	0.003071	chr7	50367319	50367319	A	T	IKZF1	silent
als27	age7	0.004069	chr7	50367319	50367319	A	T	IKZF1	silent
als27	age24	0.002395	chr7	50367319	50367319	A	T	IKZF1	silent
als27	age0	0.00545	chr7	50467723	50467723	A	T	IKZF1	missense
als27	age7	0.001933	chr7	50467723	50467723	A	T	IKZF1	missense
als27	age17	0.002639	chr7	50467723	50467723	A	T	IKZF1	missense
als27	age0	0.00081411	chr4	55152051	55152051	A	T	PDGFRA	missense
als27	age7	0.001791	chr4	55152051	55152051	A	T	PDGFRA	missense
als27	age17	0.001496	chr4	55152051	55152051	A	T	PDGFRA	missense
als27	age24	0.00104	chr4	55152051	55152051	A	T	PDGFRA	missense
als23	age0	0.02381	chrX	123200408	123200408	A	T	STAG2	intronic
als23	age7	0.023757	chrX	123200408	123200408	A	T	STAG2	intronic
als23	age17	0.022992	chrX	123200408	123200408	A	T	STAG2	intronic
als23	age24	0.013279	chrX	123200408	123200408	A	T	STAG2	intronic
als8	age0	0.002512	chr7	135329698	135329698	A	T	NUP205	missense
als8	age7	0.00065	chr7	135329698	135329698	A	T	NUP205	missense

als8	age24	0.001651	chr7	135329698	135329698	A	T	NUP205	missense
als14	age0	0.002594	chr17	7667554	7667554	C	T	DNAH2	missense
als14	age7	0.337785	chr17	7667554	7667554	C	T	DNAH2	missense
als14	age17	0.359248	chr17	7667554	7667554	C	T	DNAH2	missense
als14	age24	0.325853	chr17	7667554	7667554	C	T	DNAH2	missense

Chapter 4: Engraftment of rare, pathogenic donor hematopoietic clones in unrelated allogeneic hematopoietic stem cell transplantation

Note: At the time of my thesis defense, the work presented in this chapter is in press at *Science Translational Medicine* (Wong WH *et al.* 2019)[129]. I conceptualized the manuscript with the help of Dr. Todd Druley. I developed the original idea, proposed the study to our collaborators (John DiPersio, MD, PhD and the Center for International Bone Marrow Transplant Research, CIBMTR), generated, processed, and analyzed the data. I wrote the accepted manuscript in its entirety with guidance from Todd Druley and editorial comments from co-authors, and generated the figures. Materials from the manuscript were re-formatted and re-used in writing this chapter.

4.1 Introduction

Allogeneic hematopoietic stem-cell transplantation (HSCT) is a curative therapy for a variety of hematologic disorders such as non-malignant beta-globinopathies[130], constitutional enzyme deficiencies and hematologic malignancies[131]. However, beyond pre-HSCT primary disease progression, HSCT recipients often suffer multiple early and late post-HSCT morbidities[132] such as cardiac dysfunction, coronary artery disease[133], graft-versus-host-disease (GvHD)[92], immune dysfunction/infection, cytopenias, myelodysplasia and donor-cell leukemia[134]. Many of these common morbidities have been anecdotally attributed to donor clone(s) with pathogenic mutations in a discrete panel of candidate genes[91,92,135]. These anecdotal clones would qualify as clonal hematopoiesis of indeterminate potential (CHIP; with ≥2% variant allele frequency or VAF) in an otherwise healthy person[63], and about 5% of healthy individuals above 50 years harbor CHIP clones[60-62]. However, this definition of CHIP is not based on biology, but rather the limit of detection of standard next-generation sequencing (NGS), hence the age-related prevalence takes decades of selection for some clones to expand to the level of this detection. In contrast, error-corrected sequencing (ECS) has a limit of detection of 0.0001 and has revealed that nearly everyone older than 50 years harbors hematopoietic clones with acute myeloid leukemia (AML) and atherosclerotic-associated mutations[31,93], and there are very few differences in clonal variability and frequency between those that stay healthy and those that actually develop AML[30]. The clinical relevance of hematopoietic clones <2% VAF was recently demonstrated in AML prediction[29] and mutation clearance following allogeneic HSCT for myelodysplastic syndrome[99], where clones as rare as 0.005 VAF were clinically relevant for disease progression.

Recently, Frick and colleagues[92] studied common clonal mutations in the context of CHIP from older, matched, related HSCT donors >55 years old where approximately ≥5% of this population would be expected to harbor CHIP clones based on prior studies[60-62]. This study found that the presence of CHIP correlated with the development of chronic GvHD. However, that study is limited by only examining older, related donors and mutations above 0.02 VAF.

Unlike older related HSCT donors who are expected to have CHIP, 86% of eligible unrelated donors are adolescents and young adults (AYA) aged 18-44[128], an age group where CHIP is virtually non-detectable[60-62] via standard NGS but recipient morbidity generally exceeds that seen in related HSCT. Despite not having CHIP, it has been hypothesized that the AYA population harbors hematopoietic somatic mutations of low VAF below 0.02 VAF (as CH), otherwise undetectable via standard NGS[125], and these mutations could serve as a reservoir for future disease development when relevant selective pressure is present[136]. Hence, the appropriate way to study CH with low VAFs in the AYA group and the effects thereof in HSCT recipients is via ultra-sensitive sequencing techniques that could circumvent the error rate of standard NGS, such as ECS[93,102]. Indeed, as previously demonstrated in Chapter 3, 30% of healthy individuals were born with clonal mutations at low VAFs in their blood using ECS.

In addition, the genes frequently mutated in AYA leukemia[110] differ substantially from leukemia in older adults[137], suggesting that the AYA population may harbor a different clonal hematopoietic mutation spectrum than that seen in the CHIP literature. However, the physiologic prevalence and mutation spectrum of hematopoietic clones with mutations <0.02 VAF in the AYA population has not been quantitatively characterized. Thus, as shown in Chapter 2, our 80-gene targeted panel also included genes that are frequently mutated in pediatric/AYA and older adult AML.

In sum, this caused us to hypothesize that, a) unrelated, AYA HSCT donors may harbor hematopoietic clones with mutations <0.02 VAF in genes other than those associated with CHIP, and b) these mutations may confer a growth or survival advantage and are therefore selected and engraft recipients. In this model, prior and ongoing chemotherapy, radiotherapy and immunosuppression can act as potent selective pressures on any cell with a survival or proliferation advantage. In fact, ECS has previously demonstrated a comparable process in therapy-related AML (t-AML)[31] where pre-existing *TP53*-mutated hematopoietic progenitors, as rare as 0.0003 VAF, are selected by treatment of the primary malignancy and lead to t-AML months to years later. To interrogate this hypothesis, our primary goal was to find retrospectively banked, matched unrelated donor:recipient samples with as many longitudinal time points as possible. For each pair, five samples were evaluated: donor pre-HSCT, recipient pre-HSCT, recipient at 30 (D30), 100 (D100) and 360 days (D360 or 1-year) post-HSCT. We asked the following questions: a) what is the clonal hematopoietic spectrum in younger, healthy donors, b) how many donor clones are typically transferred to recipients, and c) what happens to these clones longitudinally in recipients. Given that the presence of clonal hematopoiesis is unexpected in this donor age group and there may have been little to no clonal transfer to recipients, this study was not designed to correlate clinical outcomes with donor clonal hematopoiesis, but the results indicate that such a study is warranted.

4.2 Materials and Methods

4.2.1 Overall study design

This retrospective pilot study was designed to interrogate donor-derived clonal dynamics post-HSCT. From the adult AML specimen repository at Washington University, a total of 25 patients were identified that had banked samples prior to transplant and at days 30, 100 and 360 post-HSCT (Figure 4.1). There were no other selection criteria. From that group, the CIBMTR was able to provide the matched donor pre-HSCT specimens, again without any additional selection. The institutional review board at Washington University in St. Louis approved the study (IRB #201710007), and patient consent had been obtained.

4.2.2 Sample collection

Four longitudinally-collected peripheral blood and/or bone marrow samples per recipient were acquired for 25 recipients with primary hematological malignancies who had undergone matched, unrelated donor allogeneic HSCT at Barnes-Jewish Hospital/ Siteman Cancer

Center/Washington University School of Medicine (Table 4.1). 64% of the patients were transplanted for myeloid malignancies. For each patient, samples were collected prior to HSCT conditioning (pre-HSCT), 30-days (D30), 100-days (D100) and 1-year post-HSCT (D360). In addition, aliquots from 25 corresponding unrelated donor leukocyte samples collected prior to HSCT were obtained from the Center for International Blood and Marrow Transplant Research (CIBMTR) repository. Upon receipt, the samples were stored in -80°C until library preparation. In total, 125 unique samples (100 patient samples from four time points and 25 donor samples) were collected and processed.

4.2.3 ECS library preparation and mutation analysis

Genomic DNA was extracted from the blood/marrow samples using DNease Blood and Tissue Kit (Qiagen) following manufacturer's recommendations. The final DNA elution volume was 50 µl. The concentration of the extracted DNA was determined with the Qubit dsDNA HS Assay (Life Technologies). Following quantification of DNA concentration, 200-250 ng of DNA per sample was used to make ultra-deep error-corrected sequencing libraries via the custom-made Illumina TruSeq Custom Amplicon kit specified in Chapter 2. The library preparation followed the protocol outlined in Chapter 2 closely with several modifications. First, following ddPCR quantification step, each library was then normalized to 6.3 million UMI-tagged molecules, and a second round of PCR (14 cycles) was performed in a 50 µl reaction: 25 µL of Q5 master mix, 2 µL of P5 Primer (1 µM), 2 µL of P7 Primer (1 µM), and 21 µL of DNA molecules. After that, the amplified libraries were purified, and the libraries were normalized. Six purified libraries were pooled and sequenced per lane in an Illumina HiSeq 4000 instrument with the following settings: 2x144 PE, 8 cycles Index 1, 16 cycles Index 2 (account for 16N random bases used as UMI). For each sample, a technical replicate library was prepared via the same protocol. In total, 125 samples were processed, and 250 ECS libraries were prepared.

The bioinformatics processing followed the pipeline outlined in Chapter 2 closely with additional filters. First, a minimum of three raw reads sharing the same UMI were error-corrected to give error-corrected consensus sequence (ECCS). Each library was deep-sequenced to an average ECCS depth of 9200× (Figure 4.2). Second, after variant-calling using binomial error model, several other filters were applied to remove artifacts and to obtain high-confidence variants: 1) variants that were only called in one technical sequencing replicate but not in the other were removed, 2) variants called due to sequencing batch effect were removed, 3) non-hotspot variants identified in more than one donor-recipient matched pair were removed, 4) variants with <0.001 VAF were removed unless the variants were observed in multiple time points in the matched sample set, 5) variants that had a coefficient of variation >15% between 3-reads and 5-reads error-corrections were removed. Following the filters, we retained a set of high-confidence variants by removing false-positive calls and common variants that are observed in the general population. Indels were called with VarScan2[120] mpileup2indel setting using the aligned consensus reads. Third, Two independent replicate sequencing libraries were made and sequenced separately (DNA was extracted from different aliquots of leukocytes from the same sample). Variants that passed the established filters in all available libraries for that sample were retained for further analysis. Variants present in pre-HSCT recipient samples represented the clonal hematopoietic profile of the recipient and, potentially, any remaining primary leukemia. These pre-HSCT germline variants in recipients were used to evaluate the degree of mixed chimerism in the recipient post-HSCT. Engraftment of donor hematopoietic clone(s) was evaluated based on the presence of variants from donor pre-HSCT observed in recipient post-

HSCT samples post-HSCT. Persistent engraftment was further defined as having donor-derived mutation(s) that persist through 1-year (D360) post-HSCT.

4.2.4 Validation of observed mutations via droplet digital PCR and triplicate sequencing

A possibility remained that somatic variants observed in a single time point were false positives. In order to further rule out potential false positives, we performed droplet digital PCR (ddPCR) using Bio-Rad QX200 platform or triplicate sequencing on these observed variants. For ddPCR, a primer/probe set specific to the variant of interest was designed by Bio-Rad according to MIQE guidelines for quantitative PCR (Table 4.2). Probes target both reference and mutated nucleotide at the same genomic position via different fluorophores. All ddPCR reactions were performed in accordance to manufacturer's recommendations using "ddPCR Supermix for Probe (no dUTP)". For triplicate sequencing, we considered only those variants observed in all three independent sequencing runs to be true positives.

4.2.5 Statistical analysis of clinical correlates

Categorical variables (donor gender, recipient gender, primary disease=AML/MDS or others, disease status before transplant=remission, conditioning=myeloablative or reduced intensity, and HLA mismatch=No) are compared using Fisher's exact test. A nonparametric Wilcoxon rank-sum test is used to compare continuous, non-Gaussian variables (duration of cytopenia, age of donor and age of recipient). Cytopenia is defined as white blood cell count $<2\times10^9$/L, hemoglobin <10g/dL and platelets $<100\times10^9$/L. Cumulative incidence of chronic GvHD is modeled because several patients died without chronic GvHD, making death a competing risk for this endpoint. The start time for chronic GvHD is D100 post-transplant. Leukemia free survival is compared using a Kaplan-Meier model. Mixed chimerism is measured repeatedly as a presence/absence, and it is compared using a repeated measures logistic regression. The analysis is intended to be exploratory, so no attempt was made to adjust the p-values for multiple tests.

4.3 Results

4.3.1 Engraftment of pathogenic somatic variants of donor origin

Given that the prevalence of hematopoietic clones at <0.02 VAF in the healthy AYA population has not been quantified, we first characterized the prevalence and genetic spectrum of clonal hematopoietic mutations in donors prior to transplantation. Because clonal hematopoiesis is associated with multiple complex health problems and all cause mortality[61], we are not solely interested in mutations associated with hematologic malignancies, but rather any mutation that would confer a growth or survival advantage to a cell due to altered molecular functions.

The donor pool consisted of 25 individuals with a median age of 26 years (range 20-58). Only one donor, aged 23 (4% of donors), harbored a CHIP clone at >0.02 VAF (*SRCAP* frame shift Indel). In total, we identified 19 somatic mutations in 11 donors, aged 20-58 (44% of donors) (Figure 4.3A; Table 4.3). The median VAF of these somatic mutations was 0.00247 (an order of magnitude rarer than the definition of CHIP) with a range of 0.00058 – 0.0274. Fourteen donors had no clonal mutations in the 80 target genes. Consistent with previous studies, despite a younger cohort, donors possessed mutations most frequently in *DNMT3A* and *TET2* (Figure 4.3B). None of the mutations detected in donors were observed in the pre-HSCT samples of recipients. Each mutation was annotated using the Combined Annotation-Dependent Depletion

(CADD) scoring system. Mutations with a scaled CADD score ≥20 represent the top 1% of mutations expected to be most pathogenic to any cellular function[138], and were thus labeled as “pathogenic” mutations in this study. We found that 84.2% of the detected mutations were pathogenic (Figure 4.3B; Table 4.1) and all detected somatic mutations engrafted in recipients. The most common mutations were cytosine to thymine transitions (Figure 4.3C) as previously seen in healthy, elderly adults[93]. The median age for the donors with clonal hematopoiesis and those without was 36 and 24, respectively, which was a significant difference (Figure 4.3D; p-value=0.03; two-sided Wilcoxon rank-sum test).

Notably, of the 19 engrafted mutations, 14 (74%) clones persisted through D365 post-HSCT, and 13 of these possessed pathogenic mutations (Figure 4.3E; Figure 4.4). The likelihood of persistent engraftment was not dependent on the initial VAF in donors (p-value=0.105; two-sided Wilcoxon rank-sum test). Despite an initially low VAF, 3 recipients (12%) had engrafted clones that expanded beyond the defined CHIP threshold of ≥0.02 VAF post-HSCT at D100 and D360 (Figure 4.5). All mutations that expanded to ≥0.02 VAF were scored as pathogenic, and the mutated genes were *TP53* p.R150W (COSMIC ID: COSM99925; CADD=25.7), *DNMT3A* p.Q222P (CADD=26.1) and *CREBBP* p.R445* (COSMIC ID: COSM255965; CADD=38).

4.3.2 Presence of de novo pathogenic somatic mutations in recipients post-HSCT

Next, we examined longitudinal differences in the mutational spectrum of engrafted clones. By comparing the recipient’s clonal profile pre-HSCT and post-HSCT, we accounted for residual physiologic hematopoietic clones and residual primary disease (Table 4.4 – 4.5). These recipient clones were filtered out accordingly.

As expected, given their high prevalence, *DNMT3A* mutations were most commonly observed after HSCT across all time points (Figure 4.6A), and the majority of these were engrafted from donors. In addition, 39 of the detected mutations in recipients after HSCT were new mutations not previously observed in donors. These newly detected mutations were called in different genes from those observed in donors and in previous CHIP studies[60-62]. For instance, *TET2*, *CREBBP* and *FAT1* were more commonly mutated in recipients after HSCT than in donors before HSCT (Figure 4.6B). The most common type of nucleotide change was cytosine to thymine (Figure 4.6C). We also found that the mutation burden across the entire cohort significantly increased from pre-HSCT (19 total somatic mutations) in donors to D100 (33 total somatic mutations) (p-value = 0.048, one-sided Wilcoxon rank-sum test; Figure 4.6C). The presence of these mutations was not due to differences in sequencing metrics. In addition, when comparing the presence of these mutations in recipients who were transplanted from donors with (n=11) or without (n=14) detectable clonal mutations, we found no difference in this observation (n.s., p-value =0.44, two-sided Wilcoxon rank-sum test).

Potential explanations for the presence of these mutations were either that they were 1) present in donors prior to transplant with a VAF below the limit of ECS detection, or 2) arose *de novo* after engraftment. To distinguish between these two possibilities, we performed ddPCR on a subset of mutations in all five samples from matched pairs. We found that these mutations were a mixture of few extremely rare donor mutations that engrafted in recipients and underwent clonal expansion, and a majority of *de novo* mutations that appear post-HSCT (Figure 4.7A-C; Figure 4.8). Some *de novo* mutations persist or expanded (Figure 4.7A) over time, while some were transient and vanished by later time points (Figure 4.7B). With respect to exceedingly rare pre-existing clones, one recipient (PID_0450) was found to have a *CREBBP* nonsense mutation, which was not detected in the donor pre-HSCT sample via ECS. By ddPCR, the same mutation was detected in the donor pre-HSCT (Figure 4.7C) and underwent an approximate 500-fold expansion with an increase in VAF from 0.000046 to 0.027 by D360 post-HSCT. The prevalence of these mutations was associated with gene length (p-value <0.00001, Pearson correlation = 0.5136; Figure 4.9) suggesting a stochastic mechanism of mutation.

4.3.3 Persistent engraftment of donor-derived mutations and clinical descriptors

While this study was not designed nor powered to establish clinical correlations to clonal hematopoiesis, we nevertheless explored the relationships between engrafted donor-derived mutations and clinical outcomes as a descriptive and exploratory pilot analysis. We were particularly interested in chronic GvHD, which was recently associated with CHIP clones engrafted from older, related donors[92]. Since young, unrelated donors with CHIP are rare (we detected CHIP in one donor), we examined the effect of persistent engraftment (up to one year) of these donor-derived mutations. We found that 75% of recipients who had at least one persistently engrafted, pathogenic mutation developed chronic GvHD versus roughly 50% of those without any persistently engrafted mutated clones. However, given the sample size, the difference is not statistically significant (p-value=0.17, Gray's test; Figure 4.10 – 4.11). Descriptive results for other clinical outcome measures for donors with and without clonal mutations (as well as pathogenic or non-pathogenic) are listed in Table 4.6.

4.4 Discussion

In this pilot study intended to quantify the presence of rare hematopoietic clones in the healthy AYA population and observe the dynamics of these clones over time in an unrelated allogeneic HSCT context, we have made five observations, which address several outstanding questions. First, we showed that clonal hematopoietic mutations ≥0.0005 VAF are common (44%) in the AYA population – an age group where CHIP is virtually non-detectable in previous studies[60-62], but constitutes 86% of eligible unrelated HSPC donors. While not demonstrated here, prior data suggests that these mutations, which were present at 10-fold lesser VAF than CHIP, are likely to occur in hematopoietic progenitors due to their presence in myeloid and lymphoid lineages in comparable frequencies as well as their persistent nature over time[93,139]. A substantial proportion of these clones harbor mutations that could confer a survival or proliferative advantage upon selective pressures. If we simply examine common mutations at or above the defined CHIP threshold of 0.02 VAF without considering rare clones, we would miss most, if not all, of these mutations in unrelated donors that might have as yet unknown clinical impacts, as acknowledged by Frick and colleagues[92]. Given the many indications for unrelated, allogeneic HSCT and recent associations of clonal hematopoiesis with risks for developing leukemia[29,30], atherosclerosis[89] and chronic GvHD after HSCT[92], and that under selective pressures these pre-existing clones can emerge to clinical relevance years after their selection[31], it is crucial to understand how putatively pathogenic clones in this age group can be transferred from healthy donors to recipients who have undergone combinations of radiation, chemotherapy and immunosuppression.

Second, we find that donor hematopoietic clones harbor mutations that are mostly pathogenic (84.2%) and have a seemingly strong predilection for engraftment (100% in this cohort). Third, rare clones with pathogenic mutations are likely to persist/expand for at least one-year post-HSCT, regardless of initial VAF. These two observations support the hypothesis that pathogenic mutations confer a variable fitness advantage to the donor cells[140] and would also explain why these engrafted rare, pathogenic mutations persist/expand in recipients post-HSCT. Fourth, the fact that there was no difference in the pre-HSCT VAF of clones with and without persistent engraftment argues for quantifying the presence of rare clones with mutations conferring a strong effect over time and against recent reports attributing clinical relevance solely to "clone size"[141]. An example of this is the recipient with a rare donor-derived *CREBBP*-mutated clone expanding 500-fold in the recipient one-year post-HSCT. *CREBBP* mutations have been shown to adversely impact hematopoietic development and are associated with malignant lymphoid stem-like properties[142]. Thus, in an advantageous context, rare clones with mutations conferring a strong effect size or selective advantage can expand relatively rapidly regardless of their initial VAF.

Fifth, we found that the clonal hematopoietic spectrum of recipients post-HSCT transiently changes over time, revealing mutations within the first year post-HSCT less commonly seen in physiologic CHIP and apparently developing from *de novo* mutations gained post-HSCT. The positive association between post-HSCT mutations and gene length suggests clonal drift. Under this scenario, the rapid proliferation of donor hematopoietic progenitors would introduce stochastic mutations across the genome, and only clones with an advantage would persist over time. In light of this, we suggest that there are many rare hematopoietic progenitors with pathogenic mutations in unrelated, otherwise-healthy AYA donors that are otherwise neutral in the donor, due to a lack of selective pressure, but in recipients, the selective pressures previously mentioned would enable preferential expansion.

Alternatively, donor cells may experience a transient hypermutative phase upon encountering an unfamiliar microenvironment. Transient hypermutation of cellular subpopulations has been shown to give rise to adaptive mutations that allows new cellular phenotypes to emerge[143,144], and the process selectively mutates epigenetic modifier genes as they promote cell phenotypic heterogeneity[145]. Such a hypothesis would be consistent with the observed increase in clonal mutation burden as a function of time post-HSCT, as well as the observation that some *de novo* mutations disappear and some expand by D365, meaning only the clones with a selective advantage persist. In addition, most of *DNMT3A* mutations observed in recipients were engrafted from donors, supporting the hypothesis that *DNMT3A* mutated clones, or more broadly - clones with mutations in epigenetic modifiers, such as *CREBBP*, or *TET2*, harbor a competitive advantage[146,147].

In summary, we have shown that extremely rare, pre-existing clones with pathogenic mutations are preferentially engrafted over clones with benign mutations. Our sample size and only one-year of post-HSCT follow-up prevented us from establishing clinical correlations. It would stand to reason that our demonstration of engraftment of clones at 10-fold lower VAF than CHIP would require a longer time to manifestation of clinical consequences. Thus, this pilot study interrogating the prevalence of rare clonal hematopoiesis in the AYA population and what happens to these clones in unrelated HSCT recipients merits a much larger study with longer follow up to correlate post-HSCT morbidities with transfer and persistence of donor clones. Such correlations could enable clinicians to survey the clonal hematopoietic profile of potential donors to improve post-HSCT surveillance and mitigate potential long-term morbidity.

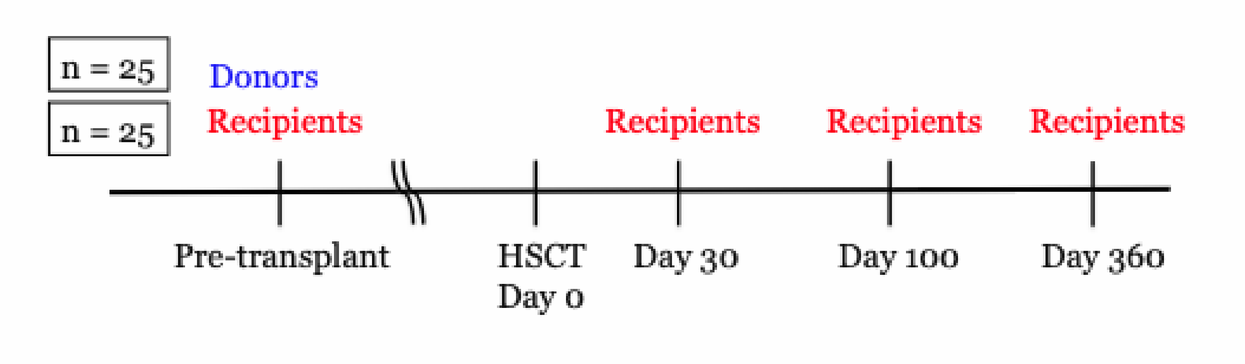


Figure 4.1: Samples collected at different time points in this study. 25 recipient samples were collected 30 days, 100 days and 1 year post-HSCT from the Siteman Cancer Center at Washington University School of Medicine. The disease samples prior to HSCT were also collected. Matched, unrelated donor samples were obtained from the National Marrow Donor Program.

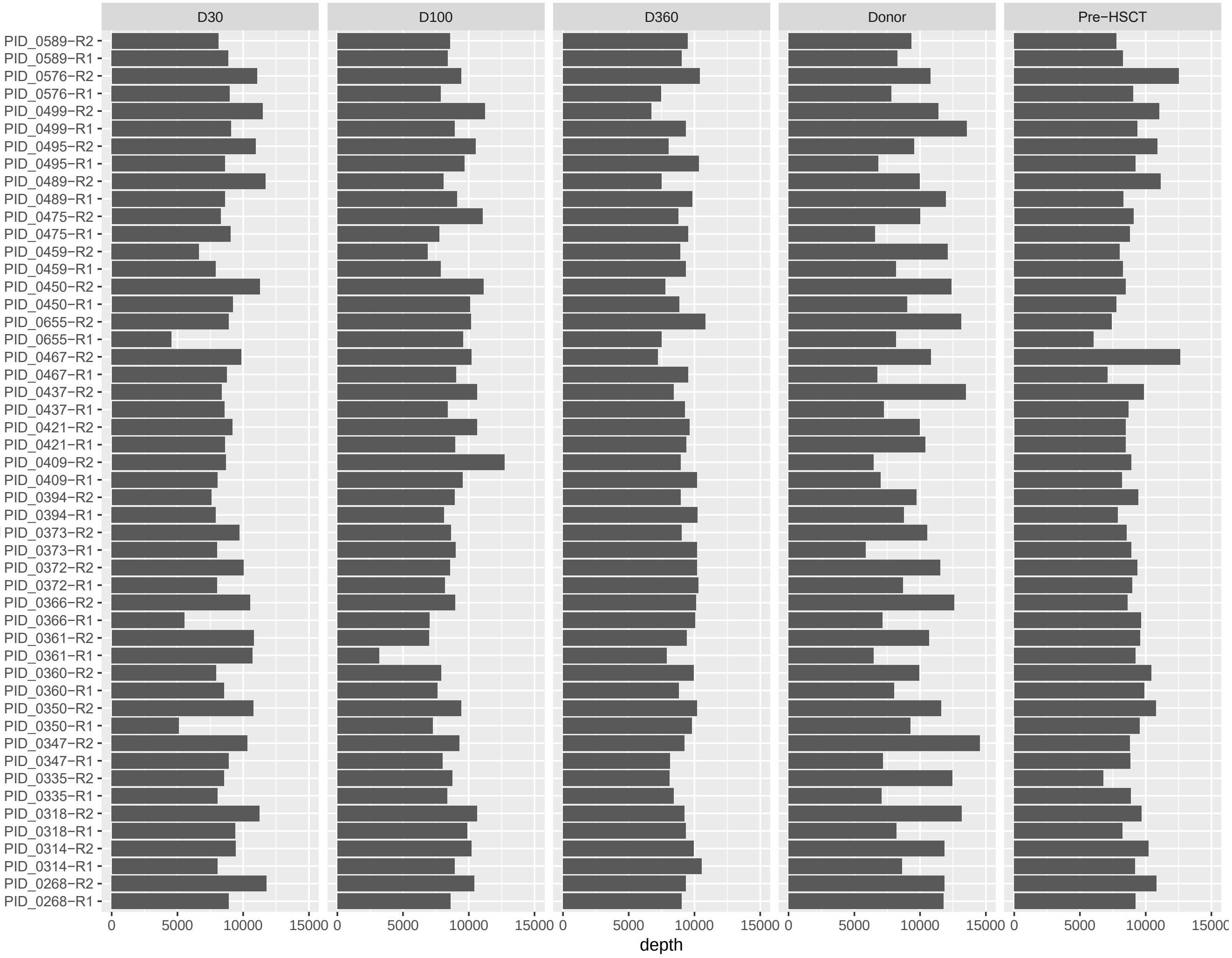


Figure 4.2: Error-corrected consensus sequencing depths (mean 9200x) of individual libraries at all time points (no difference in sequencing depth, p-values >0.05, two-sided Wilcoxon rank-sum test).

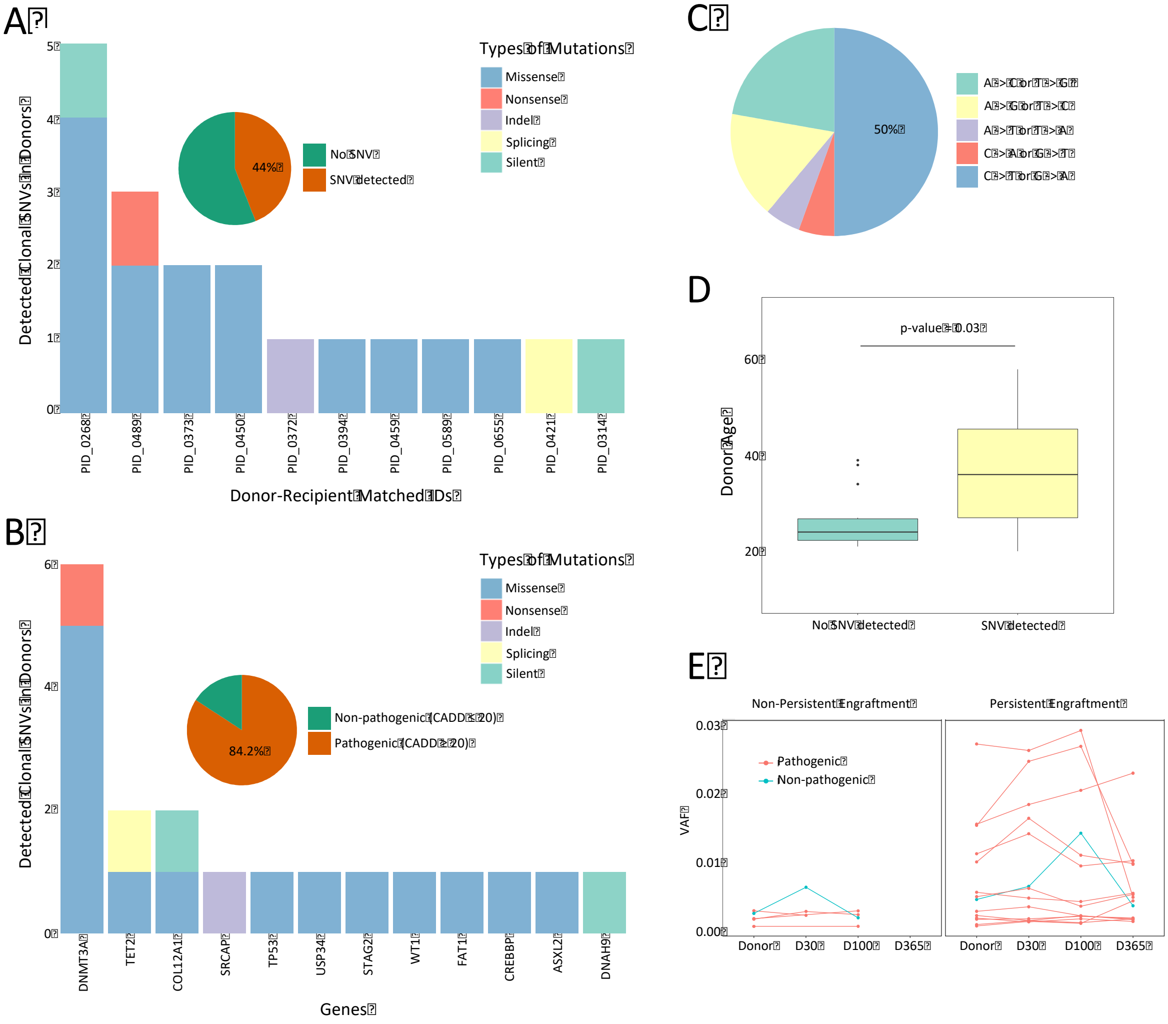


Figure 4.3: Mutation burden and spectrum in donors. (A) Number and types of somatic mutations detected in donors. 44% of the donor cohort harbored at least one somatic mutation. (B) Mutation spectrum of detected mutations in donors. *DNMT3A* and *TET2* are most recurrently mutated. (C) Types of nucleotide changes. (D) Older donors are significantly more likely to harbor detectable mutation(s). Boxes showed the 25th and 75th percentiles, as well as median. (p-value = 0.03, two-sided Wilcoxon rank-sum test). (E) Clonal dynamics of engrafted mutations in recipients post-HSCT.

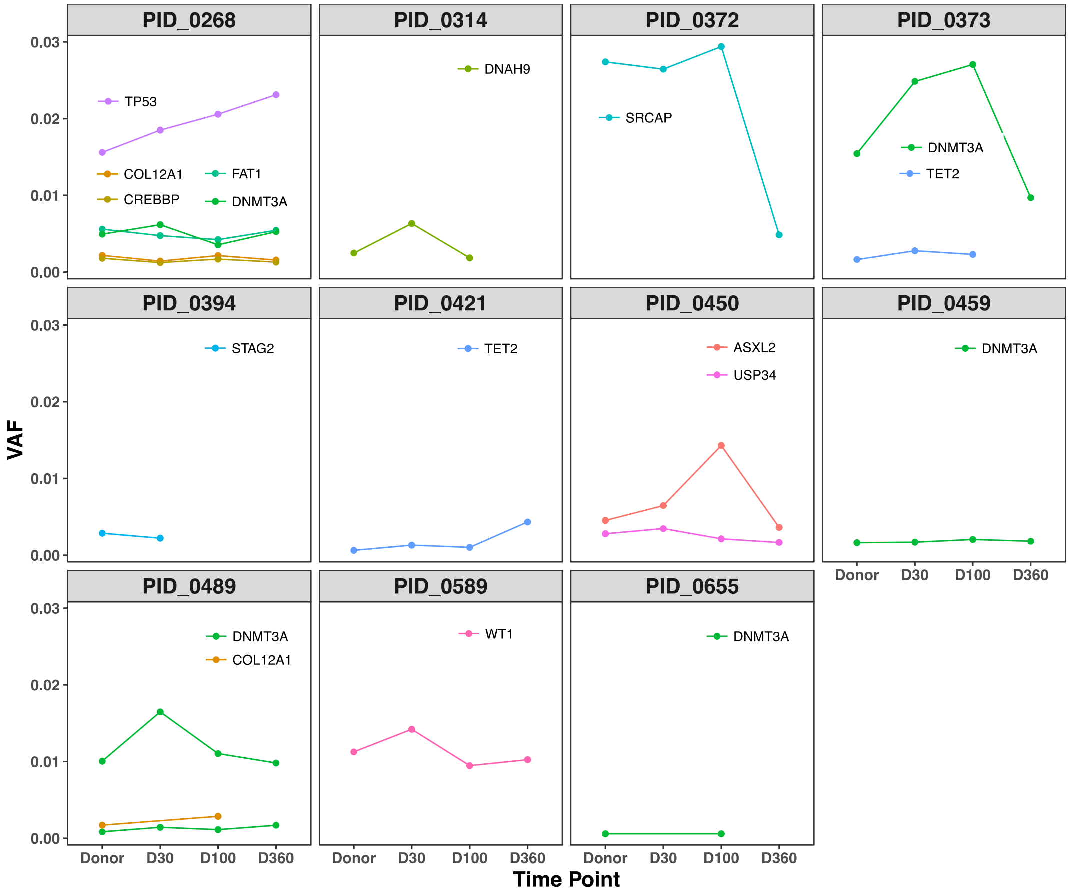


Figure 4.4: Engrafted donor mutations in recipients. Dots connected by a line indicate a mutation that was observed at multiple time points. A total of 19 mutations from 11 donors engrafted in the recipients, with 14 mutations persisted through 1-year post-HSCT in recipients. Four donors harbored more than one somatic mutation in these genes.

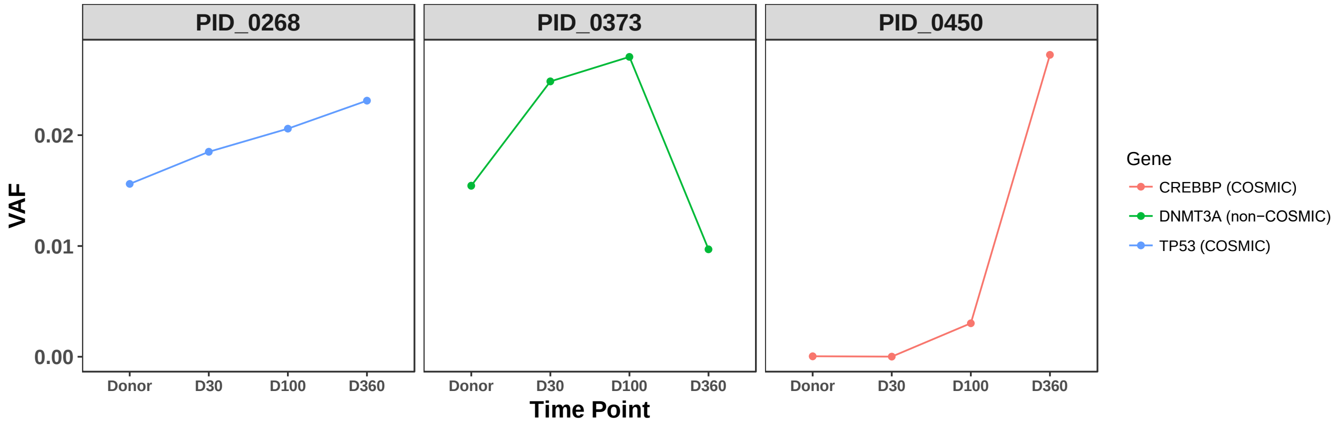


Figure 4.5: Clonal expansion of mutations reaching CHIP level (≥0.02 VAF) in three patients post-HSCT. Engraftment of *DNMT3A* and *TP53* mutations were identified by ECS while engraftment of *CREBBP* mutation was identified by ddPCR that has higher sensitivity compared to ECS to detect rare mutations.

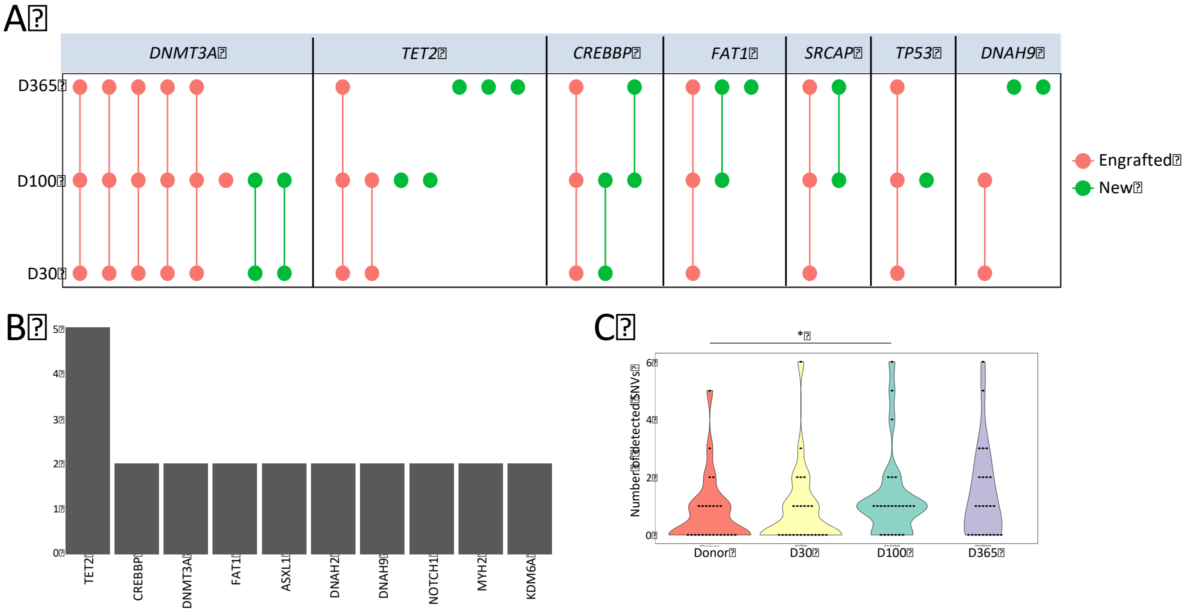


Figure 4.6: Mutation burden and spectrum of donor cells engrafted in the recipients. (A) Mutation spectrum in donor cells in recipients at different time point. (B) New mutations detected in recipients but not in pre-HSCT donors. (C) Violin plot showing mutation burden at different time points post-HSCT. * denotes p-value <0.05 with one-sided Wilcoxon rank-sum test.

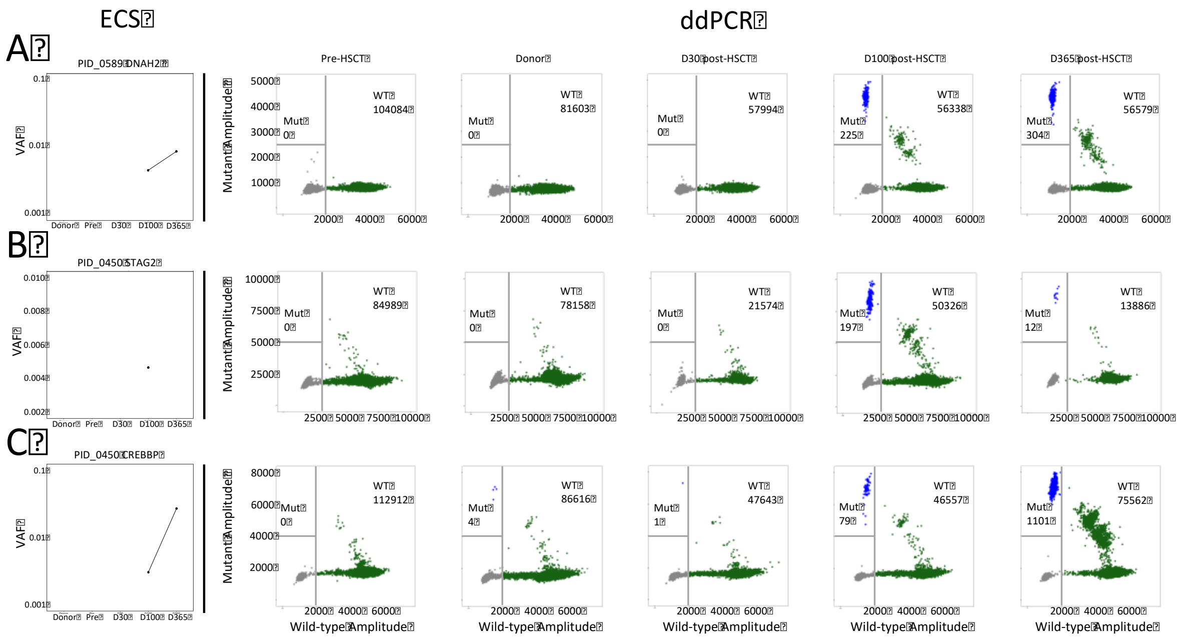


Figure 4.7: Droplet digital PCR validations of identified somatic mutations in donors. (A and B) ddPCR results of *de novo* mutations. (A) the mutation first arose in D100 post-HSCT, and experienced a slight expansion by 1-year post-HSCT. (B) the mutation first arose in D100 post-HSCT, but decreased in frequency by 1-year post-HSCT, and was undetected via ECS. (C) ddPCR results of an engrafted donor-derived mutation that was extremely rare in the donor, and was undetected via ECS. For the panels, blue dots indicate positive mutant droplets, green dots indicate positive wild-type droplets, and gray dots indicate empty droplets.

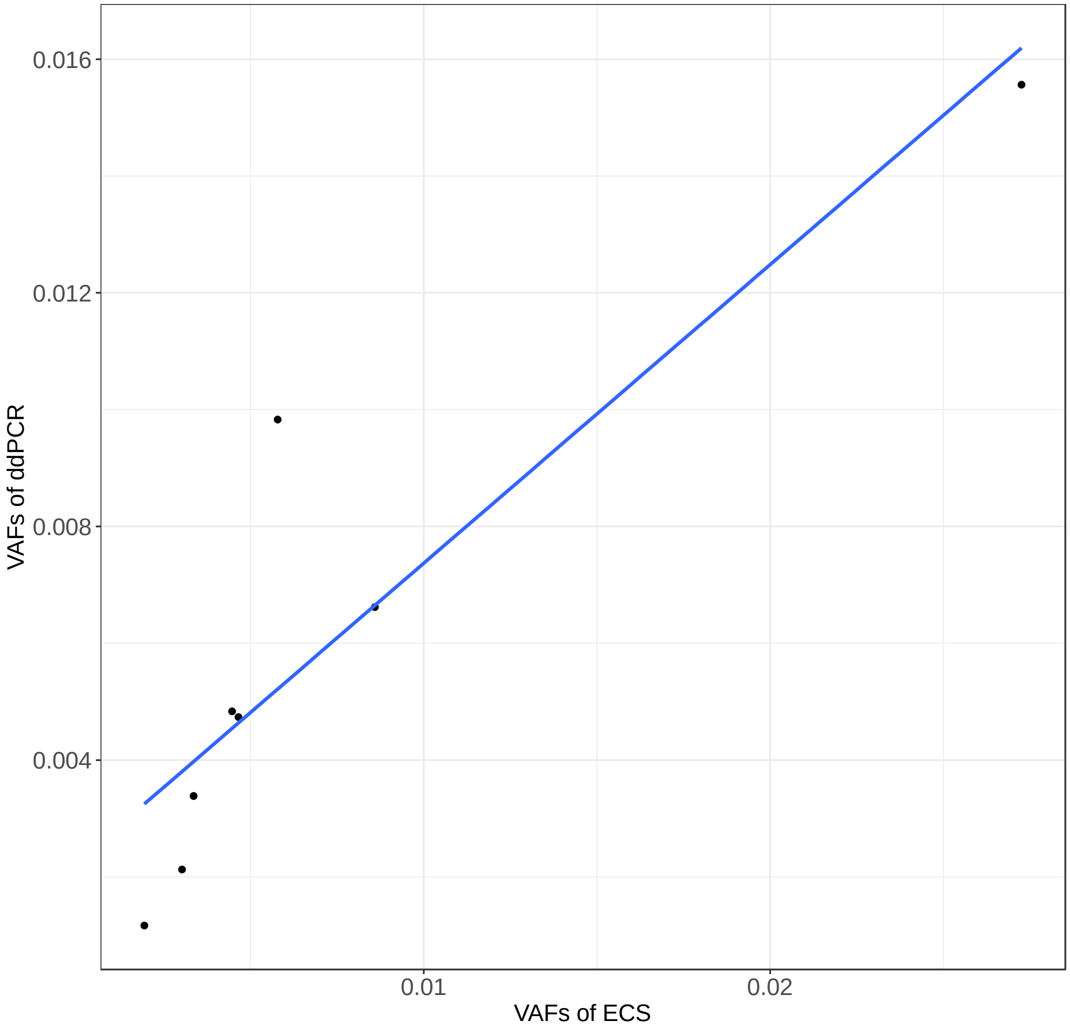


Figure 4.8: ECS calls validated by ddPCR (Positive correlation with Pearson correlation coefficient = 0.9003, p-value = 0.002296).

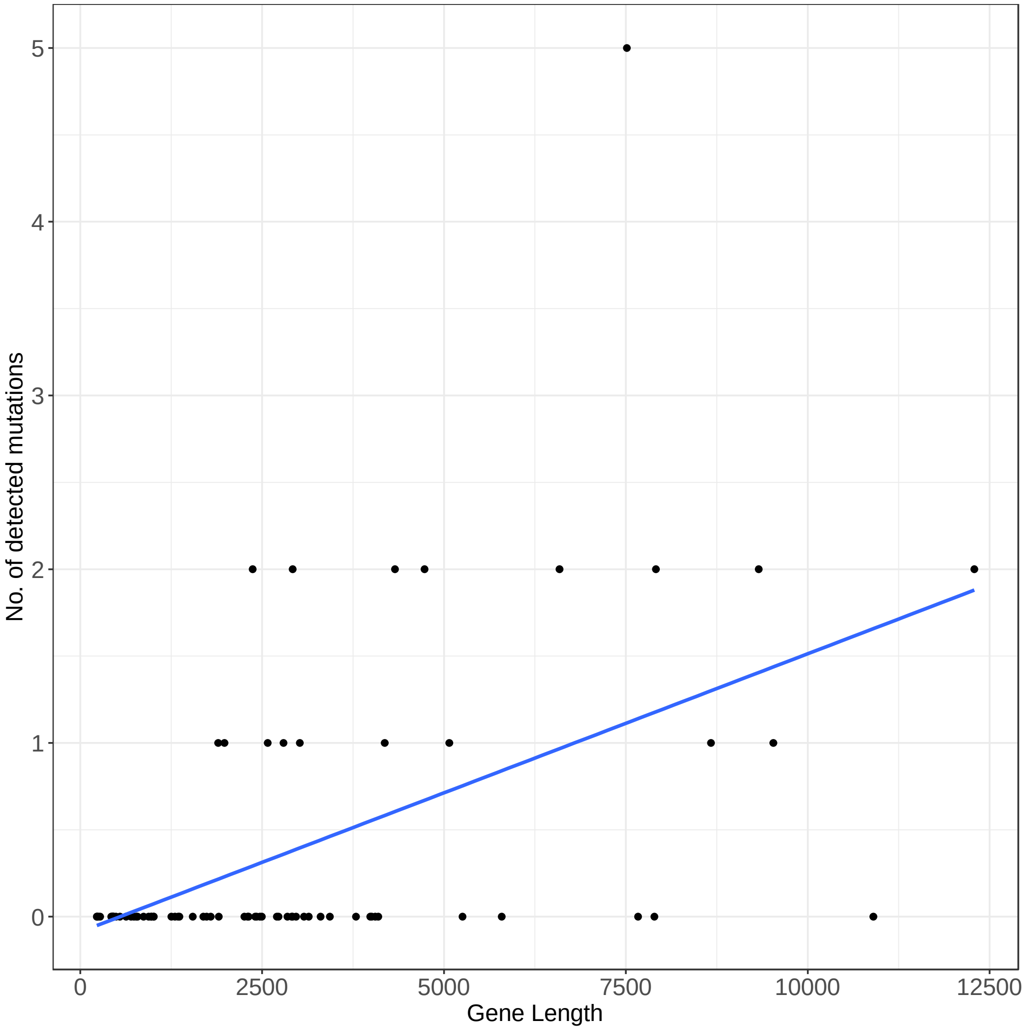
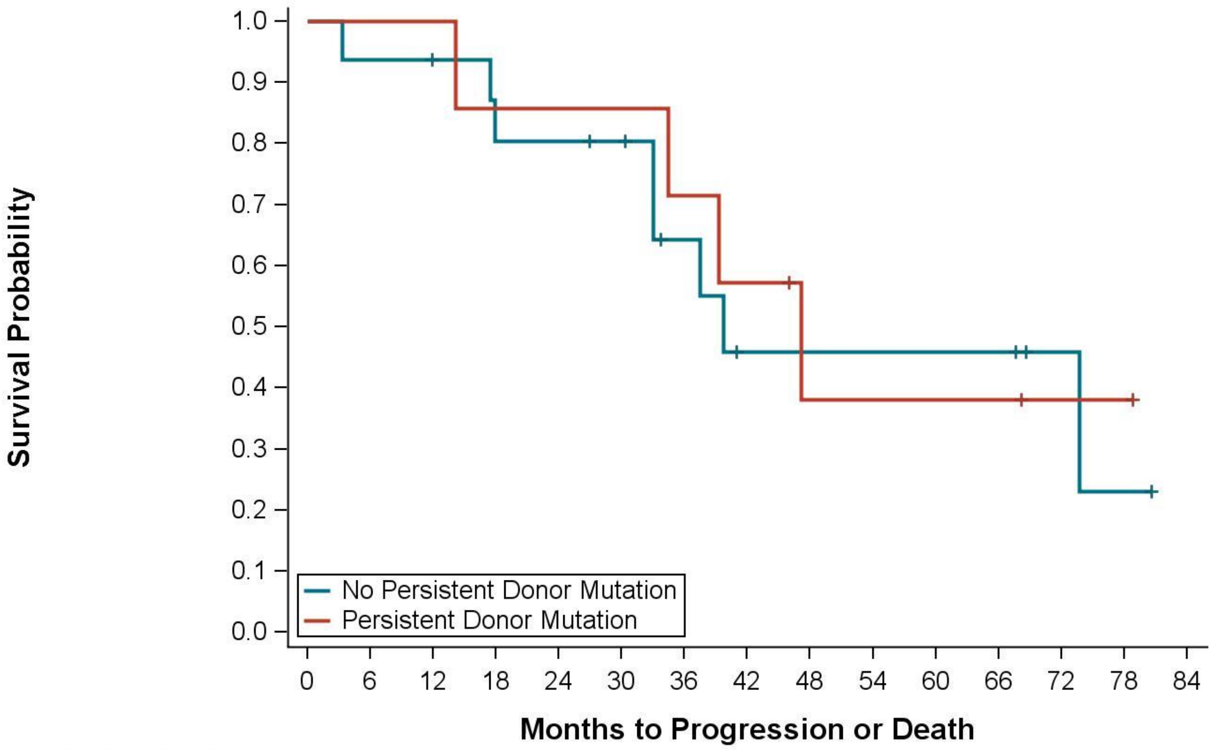
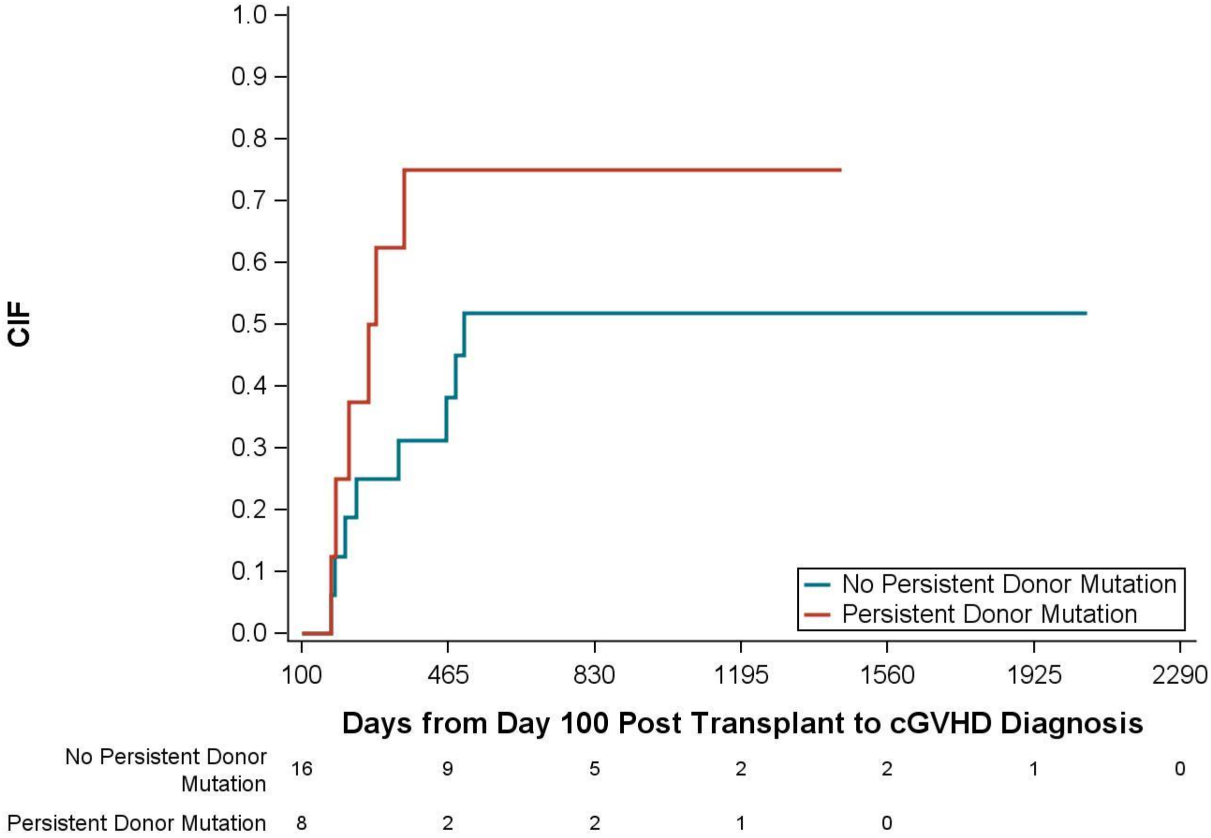


Figure 4.9: Number of detected mutations in genes according to gene length. (Positive correlation with Pearson correlation coefficient = 0.5131, p-value <0.00001).



No Persistent Donor Mutation	16	15	14	12	12	11	7	4	4	4	4	4	2	1	0
Persistent Donor Mutation	7	7	7	6	6	6	5	4	2	2	2	2	1	1	0

Figure 4.10: Leukemia-free survival of recipients with or without persistent engraftment of donor-derived mutations. (n.s., p-value = 0.7636).



No Persistent Donor Mutation	16	9	5	2	2	1	0
Persistent Donor Mutation	8	2	2	1	0		

Figure 4.11: Cumulative incidence of chronic GvHD in recipients with or without persistent engraftment of donor-derived mutations. (n.s., p-value = 0.1755).

Table 4.1: Demographical information of recipients and the corresponding matched donors in relation to engraftment of donor-derived mutations.

Characteristic	Category	No Donor Mutation (n=14)	Mutation Engrafted (n=11)	p-value	Test performed
Donor age	Median (Range)	24 (21-39)	36 (20-58)	0.03	Wilcoxon rank-sum test
Donor Gender	Male	10 (71.4%)	8 (72.7%)	0.99	Fisher's Exact test
	Female	4 (28.6%)	3 (27.3%)		
Recipient Age	Median (Range)	51 (27-65)	55 (19-69)	0.66	Wilcoxon rank-sum test
Recipient Gender	Male	13 (92.9%)	7 (63.6%)	0.13	Fisher's Exact test
	Female	1 (7.1%)	4 (36.4%)		
Primary Disease	AML/MDS	7 (50%)	9 (81.8%)	0.21	Fisher's Exact test
	Non-AML	7 (50%)	2 (18.2%)		
Disease status prior to transplant	CR	7 (50%)	5 (45.4%)	0.99	Fisher's Exact test
	Non-CR	7 (50%)	5 (45.4%)		
	Unknown	0 (0%)	1 (9.1%)		
Conditioning	MAC	8 (57.1%)	7 (63.6%)	0.99	Fisher's Exact test
	Non-MAC	6 (42.9%)	4 (36.4%)		
HLA mismatch	No Mismatch	13 (92.9%)	9 (81.8%)	0.56	Fisher's Exact test
	Mismatch	1 (7.1%)	2 (18.2%)		

Table 4.2: ddPCR probe sequences.

Mutation	Probe sequence
STAG2 p.R259*	CATTAATATGGATAATACACAAAGACAATATGAAGCAGAACGGAATAAAATGATTGGAAAA[C/T]GAGCCAATGAGAGGCTAGAACTCCTGCTACAAAAGCGGAAAGAGGTAAACTTTTATATTGA
DNAH2 p.G1808G	CACTGACCACGGCATTGCACCTGCACCGAGGGGGCTCCCCCAAAGGCCCTGCAGGCACAGG[C/A]AAGACCGAGACCGTCAAGGACCTGGGCAAGGCCCTGGGCATATATGTCATTGTGGTCAACT
CREBBP p.R445X	AGGGCAACAGAATGCCACTTCTTTAAGTAACCCAAATCCCATAGACCCCAGCTCCATGCAG[C/T]GAGCCTATGCTGCTCTCGGACTCCCCTACATGAACCAGCCCCAGACGCAGCTGCAGCCTCA
TET2 p.R1359H	TTCGCATTCACACACACTTTTATTTTTCAGATTGAATATGAACACAGAGCACCAGAGTGCC[G/A]TCTGGGTCTGAAGGAAGGCCGTCCATTCTCAGGGGTCACTGCATGTTTGGACTTCTGTGCT
SRCAP p.L1779P	CCAGTGGGCCCAGCCCCAGCTCACACGCTGACTTTGGCTCCAGCATCGTCATCTGCTTCAC[T/C]CCTGGCCCCAGCTTCAGTGCAGACACTGACCTTGAGCCCTGCCCCAGTTCCTACCCTGGGC
FAT1 p.Q249K	CTGTCCAGTTCTGATGGTGACAATGTCACTGCTGTTATCACCGGAGCACATTCATTGGCCT[G/T]TTCGATGTGCACCGTTAGCTTGGCCATGCTGCTGATGCCACTGCTCCCATACAACTTCATG

Table 4.3: Detected somatic mutations in donors via ECS

Chr	Start	Ref	Alt	Gene	Type	AA	COSMIC ID	CADD	Deleterious	VAF1	VAF2	patientID	Engrafted	Age
chr11	32456672	G	A	WT1	missense	p.R74W	NA	28.7	Yes	0.011304	0.011202	PID_0589	Yes	30
chr2	25469138	C	T	DNMT3A	nonsense	p.W288X	COSM1130818	40	Yes	0.0004	0.001268	PID_0489	Yes	58
chr2	25470498	G	T	DNMT3A	missense	p.R174S	NA	26.5	Yes	0.011218	0.008888	PID_0489	Yes	58
chr6	75893069	T	A	COL12A1	missense	p.I530L	COSM271996	22.1	Yes	0.000843	0.002573	PID_0489	Yes	58
chr2	25467428	C	T	DNMT3A	missense	p.G398R	COSM256035	30	Yes	0.00142	0.00182	PID_0459	Yes	28
chr2	25965982	C	T	ASXL2	missense	p.R1075Q	COSM6494820	10.77	No	0.004005	0.004873	PID_0450	Yes	26
chr2	61492689	T	C	USP34	missense	p.H1874R	NA	22.5	Yes	0.002401	0.002433	PID_0450	Yes	26
chr2	25470544	A	C	DNMT3A	missense	p.I158M	NA	23.6	Yes	0.000541	0.000614	PID_0655	Yes	36
chr4	106180927	G	A	TET2	splicing	NA	COSM87141	34	Yes	0.000185805	0.001074009	PID_0421	Yes	38
chrX	123179113	T	G	STAG2	missense	p.Y188D	NA	26.3	Yes	0.00164	0.004084	PID_0394	Yes	40
chr2	25469921	T	G	DNMT3A	missense	p.Q222P	NA	26.1	Yes	0.012781	0.01809	PID_0373	Yes	51
chr4	106182995	A	G	TET2	missense	p.Y1345C	NA	32	Yes	0.00133	0.001933	PID_0373	Yes	51
chr16	30721283	T	TGCTTCGCC	SRCAP	indel	NA	NA	29	Yes	0.029	0.0258	PID_0372	Yes	23
chr17	11556271	T	C	DNAH9	silent	p.Y849Y	NA	6.417	No	0.003965	0.000986	PID_0314	Yes	20
chr6	75818737	G	A	COL12A1	silent	p.A1535A	NA	1.438	No	0.001792	0.00254	PID_0268	Yes	53
chr16	3801781	G	A	CREBBP	missense	p.T1204I	NA	25.5	Yes	0.002441	0.001158	PID_0268	Yes	53
chr17	7577094	G	A	TP53	missense	p.R150W	COSM99925	25.7	Yes	0.013423	0.017781	PID_0268	Yes	53
chr2	25459821	T	G	DNMT3A	missense	p.H669P	NA	23.3	Yes	0.005147	0.004734	PID_0268	Yes	53
chr4	187549458	C	T	FAT1	missense	p.D1554N	COSM1429043	25.8	Yes	0.007042	0.004139	PID_0268	Yes	53

Table 4.4: Detected somatic mutations in recipients post-HSCT after removing recipient's hematopoietic clones using ECS.

Chr	Start	Ref	Alt	Gene	Type	AA	COSMIC ID	CADD	VAF1	VAF2	Timepoint	patientID
chr16	3801781	G	A	CREBBP	missense	p.T1204I	NA	25.5	0.00088	0.001588	D30	PID_0268
chr16	3801781	G	A	CREBBP	missense	p.T1204I	NA	25.5	0.001788	0.001573	D100	PID_0268
chr16	3801781	G	A	CREBBP	missense	p.T1204I	NA	25.5	0.001285	0.001304	D365	PID_0268
chr17	7577094	G	A	TP53	missense	p.R150W	COSM99925	25.7	0.018132	0.018873	D30	PID_0268
chr17	7577094	G	A	TP53	missense	p.R150W	COSM99925	25.7	0.023088	0.018078	D100	PID_0268
chr17	7577094	G	A	TP53	missense	p.R150W	COSM99925	25.7	0.023448	0.022782	D365	PID_0268
chr2	25459821	T	G	DNMT3A	missense	p.H669P	NA	23.3	0.004792	0.00754	D30	PID_0268
chr2	25459821	T	G	DNMT3A	missense	p.H669P	NA	23.3	0.00522	0.001862	D100	PID_0268
chr2	25459821	T	G	DNMT3A	missense	p.H669P	NA	23.3	0.004636	0.005898	D365	PID_0268
chr2	25462044	G	A	DNMT3A	missense	p.A636V	NA	26.9	0.001243	0.001637	D30	PID_0268
chr2	25462044	G	A	DNMT3A	missense	p.A636V	NA	26.9	0.001319	0.000832	D100	PID_0268
chr6	75818737	G	A	COL12A1	silent	p.A1535A	NA	1.438	0.001169	0.001683	D30	PID_0268
chr6	75818737	G	A	COL12A1	silent	p.A1535A	NA	1.438	0.00164	0.002633	D100	PID_0268
chr6	75818737	G	A	COL12A1	silent	p.A1535A	NA	1.438	0.002188	0.000939	D365	PID_0268
chr4	106157471	ATT	A	TET2	indel	NA	NA	27.2	0.004	0.0027	D365	PID_0268
chr4	187549458	C	T	FAT1	missense	p.D1554N	COSM1429043	25.8	0.003422	0.006051	D30	PID_0268
chr4	187549458	C	T	FAT1	missense	p.D1554N	COSM1429043	25.8	0.004722	0.00372	D100	PID_0268
chr4	187549458	C	T	FAT1	missense	p.D1554N	COSM1429043	25.8	0.005821	0.005068	D365	PID_0268
chr17	11556271	T	C	DNAH9	silent	p.Y849Y	NA	6.417	0.002639	0.001049	D30	PID_0314
chr17	11556271	T	C	DNAH9	silent	p.Y849Y	NA	6.417	0.010758	0.001891	D100	PID_0314
chr17	10427979	T	C	MYH2	missense	p.Q1660R	NA	23.8	0.001193	0.001827	D365	PID_0318
chr4	187630237	G	T	FAT1	missense	p.Q249K	NA	22.7	0.004705	0.006863	D100	PID_0318
chr4	187630237	G	T	FAT1	missense	p.Q249K	NA	22.7	0.002002	0.003591	D365	PID_0318
chr7	50450427	C	T	IKZF1	intronic	NA	NA	0.256	0.003137	0.00269	D365	PID_0335
chr7	135322646	A	G	NUP205	intronic	NA	NA	11.51	0.001635	0.001371	D100	PID_0335
chr9	139391938	C	A	NOTCH1	missense	p.A2085S	NA	25.1	0.010538	0.011864	D100	PID_0335
chr9	139391938	C	A	NOTCH1	missense	p.A2085S	NA	25.1	0.004486	0.002006	D365	PID_0335

chr17	7578190	T	C	TP53	missense	p.Y88C	COSM99718	29.5	0.006901	0.00515	D100	PID_0347
chr20	31022402	TCACCACTGCCATAGAGAGGCGGC	T	ASXL1	indel	NA	NA	35	0.0142	0.0129	D100	PID_0360
chr4	106190797	C	T	TET2	missense	p.R1359C	COSM41649	32	0.004008	0.002231	D100	PID_0361
chr2	25458642	T	G	DNMT3A	missense	p.K692T	NA	28.6	0.003823	0.004353	D30	PID_0366
chr2	25458642	T	G	DNMT3A	missense	p.K692T	NA	28.6	0.001147	0.00074	D100	PID_0366
chrX	44945182	T	C	KDM6A	missense	p.V1090A	NA	27	0.001714	0.002526	D365	PID_0366
chr16	30721283	T	TGCTTCGCC	SRCAP	indel	NA	NA	29	0.0266	0.0263	D30	PID_0372
chr16	30721283	T	TGCTTCGCC	SRCAP	indel	NA	NA	29	0.033	0.0255	D100	PID_0372
chr16	30721283	T	TGCTTCGCC	SRCAP	indel	NA	NA	29	0.006	0.0037	D365	PID_0372
chrX	44896997	C	G	KDM6A	intronic	NA	NA	5.46	0.003722	0.003451	D365	PID_0372
chr2	25469921	T	G	DNMT3A	missense	p.Q222P	NA	26.1	0.024433	0.025272	D30	PID_0373
chr2	25469921	T	G	DNMT3A	missense	p.Q222P	NA	26.1	0.02903	0.025094	D100	PID_0373
chr2	25469921	T	G	DNMT3A	missense	p.Q222P	NA	26.1	0.01125	0.008141	D365	PID_0373
chr4	106182995	A	G	TET2	missense	p.Y1345C	NA	32	0.003003	0.002545	D30	PID_0373
chr4	106182995	A	G	TET2	missense	p.Y1345C	NA	32	0.001785	0.002803	D100	PID_0373
chr17	10432262	A	G	MYH2	silent	p.T1163T	NA	15.78	0.002419	0.002522	D365	PID_0394
chr17	11701030	T	A	DNAH9	silent	p.T2780T	NA	1.371	0.005347	0.005514	D365	PID_0394
chr16	30736081	T	C	SRCAP	missense	p.L1779P	NA	23.7	0.002611	0.001984	D100	PID_0394
chr16	30736081	T	C	SRCAP	missense	p.L1779P	NA	23.7	0.001275	0.002601	D365	PID_0394
chrX	123179113	T	G	STAG2	missense	p.Y188D	NA	26.3	0.003225	0.001184	D30	PID_0394
chr4	24572496	T	G	DHX15	intronic	NA	NA	16.29	0.009922	0.005584	D30	PID_0421
chr4	106156485	TC	T	TET2	indel	NA	NA	26.7	0.0289	0.02	D365	PID_0421
chr4	106180927	G	A	TET2	splicing	NA	COSM87141	34	0.000737075	0.001845444	D30	PID_0421
chr4	106180927	G	A	TET2	splicing	NA	COSM87141	34	0.000895255	0.001129093	D100	PID_0421
chr4	106180927	G	A	TET2	splicing	NA	COSM87141	34	0.004502	0.004145	D365	PID_0421
chr4	106190860	C	T	TET2	missense	p.H1380Y	COSM87161	27.4	0.00169	0.001286	D365	PID_0421
chr16	3832811	G	A	CREBBP	stopgain	p.R445X	COSM255965	38	0.002238	0.003813	D100	PID_0450
chr16	3832811	G	A	CREBBP	stopgain	p.R445X	COSM255965	38	0.031195	0.023312	D365	PID_0450
chr17	11502173	G	T	DNAH9	missense	p.V120L	NA	11.09	0.002636	0.003343	D365	PID_0450
chr2	25965982	C	T	ASXL2	missense	p.R1075Q	COSM6494820	10.77	0.005974	0.007519	D30	PID_0450

chr2	25965982	C	T	ASXL2	missense	p.R1075Q	COSM6494820	10.77	0.007112	0.016319	D100	PID_0450
chr2	25965982	C	T	ASXL2	missense	p.R1075Q	COSM6494820	10.77	0.003058	0.004677	D365	PID_0450
chr2	61492689	T	C	USP34	missense	p.H1874R	NA	22.5	0.003149	0.003222	D30	PID_0450
chr2	61492689	T	C	USP34	missense	p.H1874R	NA	22.5	0.002382	0.001837	D100	PID_0450
chr2	61492689	T	C	USP34	missense	p.H1874R	NA	22.5	0.001734	0.001308	D365	PID_0450
chrX	123181311	C	T	STAG2	stopgain	p.R259X	COSM1598816	36	0.004316	0.004992	D100	PID_0450
chr9	139399647	G	A	NOTCH1	intronic	NA	NA	1.322	0.001	0.001415	D100	PID_0450
chr9	139399647	G	A	NOTCH1	intronic	NA	NA	1.322	0.001125	0.001466	D365	PID_0450
chr2	25467428	C	T	DNMT3A	missense	p.G398R	COSM256035	30	0.001465	0.001914	D30	PID_0459
chr2	25467428	C	T	DNMT3A	missense	p.G398R	COSM256035	30	0.001199	0.002875	D100	PID_0459
chr2	25467428	C	T	DNMT3A	missense	p.G398R	COSM256035	30	0.000982	0.002632	D365	PID_0459
chr4	106190798	G	A	TET2	missense	p.R1359H	COSM42055	33	0.001979	0.004812	D100	PID_0467
chr4	187517780	G	A	FAT1	missense	p.A4305V	COSM6056356	33	0.006958	0.008668	D365	PID_0467
chr17	10355270	G	T	MYH4	silent	p.V1242V	NA	15.61	0.018793	0.019914	D365	PID_0475
chr16	3778533	T	G	CREBBP	missense	p.N2134T	NA	22.7	0.000779	0.001114	D30	PID_0489
chr16	3778533	T	G	CREBBP	missense	p.N2134T	NA	22.7	0.000768	0.000833	D100	PID_0489
chr2	25469138	C	T	DNMT3A	stopgain	p.W288X	COSM1130818	40	0.001723	0.001138	D30	PID_0489
chr2	25469138	C	T	DNMT3A	stopgain	p.W288X	COSM1130818	40	0.000715	0.00153	D100	PID_0489
chr2	25469138	C	T	DNMT3A	stopgain	p.W288X	COSM1130818	40	0.001764	0.001629	D365	PID_0489
chr2	25470498	G	T	DNMT3A	missense	p.R174S	NA	26.5	0.015894	0.017054	D30	PID_0489
chr2	25470498	G	T	DNMT3A	missense	p.R174S	NA	26.5	0.011207	0.01087	D100	PID_0489
chr2	25470498	G	T	DNMT3A	missense	p.R174S	NA	26.5	0.011979	0.007639	D365	PID_0489
chr6	75893069	T	A	COL12A1	missense	p.I530L	COSM271996	22.1	0.002051	0.003665	D100	PID_0489
chr20	31022592	CG	C	ASXL1	indel	NA	NA	34	0.0062	0.0053	D365	PID_0495
chr17	7681670	C	A	DNAH2	silent	p.G1808G	NA	5.337	0.004119	0.004819	D100	PID_0589
chr17	7681670	C	A	DNAH2	silent	p.G1808G	NA	5.337	0.008886	0.008291	D365	PID_0589
chr19	13445165	C	T	CACNA1A	intronic	NA	NA	12.07	0.004625	0.001711	D365	PID_0589
chr11	32456672	G	A	WT1	missense	p.R74W	NA	28.7	0.013294	0.015119	D30	PID_0589
chr11	32456672	G	A	WT1	missense	p.R74W	NA	28.7	0.009875	0.009044	D100	PID_0589
chr11	32456672	G	A	WT1	missense	p.R74W	NA	28.7	0.01129	0.009187	D365	PID_0589

chr2	25470544	A	C	DNMT3A	missense	p.I158M	NA	23.6	0.00058	0.000579	D100	PID_0655

Table 4.5: Shared variants in pre-HSCT and post-HSCT recipient samples due to incomplete clearance of recipient's hematopoietic clones post-HSCT via ECS.

Chr	Start	Ref	Alt	Gene	Type	AA	COSMIC ID	VAF1	VAF2	Timepoint	patientID
chr18	7002407	A	C	LAMA1	intronic	NA	NA	0.434713	0.356371	Pre	PID_0576
chr18	7002407	A	C	LAMA1	intronic	NA	NA	0.021635	0.024166	D30	PID_0576
chr18	7002407	A	C	LAMA1	intronic	NA	NA	0.013414	0.018904	D100	PID_0576
chr7	135285707	G	A	NUP205	silent	p.V430V	COSM40403	0.505976	0.512114	Pre	PID_0499
chr7	135285707	G	A	NUP205	silent	p.V430V	COSM40403	0.019954	0.021033	D30	PID_0499
chr7	135285707	G	A	NUP205	silent	p.V430V	COSM40403	0.022706	0.025451	D100	PID_0499
chr11	32456726	G	A	WT1	missense	p.R56W	NA	0.404366	0.403446	Pre	PID_0495
chr11	32456726	G	A	WT1	missense	p.R56W	NA	0.000908	0.000609	D30	PID_0495
chr16	30736222	C	T	SRCAP	missense	p.S1826L	COSM5850742	0.507346	0.496049	Pre	PID_0495
chr16	30736222	C	T	SRCAP	missense	p.S1826L	COSM5850742	0.04389	0.039401	D30	PID_0495
chr16	30736222	C	T	SRCAP	missense	p.S1826L	COSM5850742	0.006037	0.003924	D100	PID_0495
chr17	11550421	G	A	DNAH9	missense	p.R668Q	COSM1236009	0.910698	0.90111	Pre	PID_0495
chr17	11550421	G	A	DNAH9	missense	p.R668Q	COSM1236009	0.042462	0.055456	D30	PID_0495
chr17	7669761	G	A	DNAH2	missense	p.E1213K	NA	0.912955	0.915971	Pre	PID_0495
chr17	7669761	G	A	DNAH2	missense	p.E1213K	NA	0.052973	0.051185	D30	PID_0495
chr8	128750945	C	T	MYC	missense	p.S161L	COSM1454792	0.40811	0.402829	Pre	PID_0495
chr8	128750945	C	T	MYC	missense	p.S161L	COSM1454792	0.002216	0.001834	D30	PID_0495
chr9	134072817	C	G	NUP214	silent	p.T138T	NA	0.507812	0.506112	Pre	PID_0495
chr9	134072817	C	G	NUP214	silent	p.T138T	NA	0.042121	0.039321	D30	PID_0495
chr9	134072817	C	G	NUP214	silent	p.T138T	NA	0.001785	0.004946	D100	PID_0495
chr17	7697727	G	A	DNAH2	intronic	NA	NA	0.491119	0.492356	Pre	PID_0489
chr17	7697727	G	A	DNAH2	intronic	NA	NA	0.001426	0.00106	D30	PID_0489
chr9	134074118	G	C	NUP214	missense	p.S572T	NA	0.507816	0.517495	Pre	PID_0489
chr9	134074118	G	C	NUP214	missense	p.S572T	NA	0.001387	0.001787	D30	PID_0489
chr17	7577644	C	G	TP53	intronic	NA	COSN26958754	0.507881	0.508418	Pre	PID_0475
chr17	7577644	C	G	TP53	intronic	NA	COSN26958754	0.000881	0.000709	D30	PID_0475
chr5	180057843	A	G	FLT4	intronic	NA	NA	0.519392	0.502796	Pre	PID_0475
chr5	180057843	A	G	FLT4	intronic	NA	NA	0.001027	0.001216	D30	PID_0475
chr13	28608020	G	C	FLT3	intronic	NA	NA	0.505227	0.483484	Pre	PID_0459
chr13	28608020	G	C	FLT3	intronic	NA	NA	0.032958	0.028085	D30	PID_0459
chr13	28608020	G	C	FLT3	intronic	NA	NA	0.056743	0.058917	D100	PID_0459
chr4	106180845	G	T	TET2	missense	p.W1291C	COSM4383925	0.175306	0.185252	Pre	PID_0459
chr4	106180845	G	T	TET2	missense	p.W1291C	COSM4383925	0.007234	0.00947	D30	PID_0459
chr4	106180845	G	T	TET2	missense	p.W1291C	COSM4383925	0.004535	0.003971	D100	PID_0459
chr4	106157845	C	T	TET2	nonsense	p.Q916X	COSM3733079	0.171096	0.191542	Pre	PID_0459
chr4	106157845	C	T	TET2	nonsense	p.Q916X	COSM3733079	0.004234	0.003035	D30	PID_0459
chr4	106157845	C	T	TET2	nonsense	p.Q916X	COSM3733079	0.007722	0.004722	D100	PID_0459

chr4	55575669	G	A	KIT	missense	p.V399I	COSM51494	0.485625	0.495795	Pre	PID_0459
chr4	55575669	G	A	KIT	missense	p.V399I	COSM51494	0.029492	0.03654	D30	PID_0459
chr4	55575669	G	A	KIT	missense	p.V399I	COSM51494	0.05834	0.047988	D100	PID_0459
chr5	180057876	C	T	FLT4	intronic	NA	NA	0.509449	0.501632	Pre	PID_0459
chr5	180057876	C	T	FLT4	intronic	NA	NA	0.033549	0.029598	D30	PID_0459
chr5	180057876	C	T	FLT4	intronic	NA	NA	0.063978	0.057464	D100	PID_0459
chr5	180046495	A	C	FLT4	intronic	NA	NA	0.530321	0.566294	Pre	PID_0459
chr5	180046495	A	C	FLT4	intronic	NA	NA	0.028811	0.03513	D30	PID_0459
chr5	180046495	A	C	FLT4	intronic	NA	NA	0.067174	0.070913	D100	PID_0459
chr3	128200072	C	T	GATA2	silent	p.A397A	COSM5019736	0.501648	0.504831	Pre	PID_0467
chr3	128200072	C	T	GATA2	silent	p.A397A	COSM5019736	0.150707	0.155821	D30	PID_0467
chr6	138632652	G	A	ARFGEF3	intronic	NA	NA	0.001567	0.001827	Pre	PID_0467
chr6	138632652	G	A	ARFGEF3	intronic	NA	NA	0.001096	0.000721	D30	PID_0467
chr16	15814717	C	T	MYH11	silent	p.K1590K	COSM5020106	0.499401	0.496203	Pre	PID_0437
chr16	15814717	C	T	MYH11	silent	p.K1590K	COSM5020106	0.002947	0.003508	D100	PID_0437
chr16	15814717	C	T	MYH11	silent	p.K1590K	COSM5020106	0.00276	0.003484	D360	PID_0437
chrX	44922982	C	G	KDM6A	missense	p.L536V	NA	0.509905	0.497262	Pre	PID_0437
chrX	44922982	C	G	KDM6A	missense	p.L536V	NA	0.001176	0.002106	D30	PID_0437
chrX	44922982	C	G	KDM6A	missense	p.L536V	NA	0.002725	0.001689	D100	PID_0437
chr4	106156384	G	A	TET2	missense	p.G429R	COSM219042	0.493444	0.49112	Pre	PID_0409
chr4	106156384	G	A	TET2	missense	p.G429R	COSM219042	0.00807	0.008017	D30	PID_0409
chr4	106156384	G	A	TET2	missense	p.G429R	COSM219042	0.013121	0.011373	D100	PID_0409
chr16	3778424	T	G	CREBBP	missense	p.Q2170H	COSM96470	0.476022	0.503146	Pre	PID_0409
chr16	3778424	T	G	CREBBP	missense	p.Q2170H	COSM96470	0.008232	0.008722	D30	PID_0409
chr16	3778424	T	G	CREBBP	missense	p.Q2170H	COSM96470	0.01268	0.015275	D100	PID_0409
chr16	3779594	C	T	CREBBP	silent	p.V1780V	COSM5019155	0.489053	0.493702	Pre	PID_0409
chr16	3779594	C	T	CREBBP	silent	p.V1780V	COSM5019155	0.008906	0.010031	D30	PID_0409
chr16	3779594	C	T	CREBBP	silent	p.V1780V	COSM5019155	0.01582	0.015028	D100	PID_0409
chr16	30727853	C	G	SRCAP	intronic	NA	NA	0.48963	0.496588	Pre	PID_0394
chr16	30727853	C	G	SRCAP	intronic	NA	NA	0.03362	0.032133	D30	PID_0394
chr17	7684131	C	T	DNAH2	intronic	NA	NA	0.014319	0.012846	D30	PID_0361
chr17	7684131	C	T	DNAH2	intronic	NA	NA	0.503446	0.507958	Pre	PID_0361
chr20	31024704	G	A	ASXL1	missense	p.G1397S	COSM133033	0.016931	0.018797	D30	PID_0361
chr20	31024704	G	A	ASXL1	missense	p.G1397S	COSM133033	0.498835	0.498557	Pre	PID_0361
chrX	39937244	A	C	BCOR	intronic	NA	NA	0.011781	0.015437	D30	PID_0361
chrX	39937244	A	C	BCOR	intronic	NA	NA	1	0.998739	Pre	PID_0361
chr2	209113113	G	A	IDH1	missense	p.R132C	COSM28747	0.009563	0.005514	Pre	PID_0360
chr2	209113113	G	A	IDH1	missense	p.R132C	COSM28747	0.032294	0.034172	D100	PID_0360
chr20	31024260	A	G	ASXL1	missense	p.M1249V	COSM6498418	0.513097	0.490742	Pre	PID_0360
chr20	31024260	A	G	ASXL1	missense	p.M1249V	COSM6498418	0.039568	0.033524	D30	PID_0360
chr20	31024260	A	G	ASXL1	missense	p.M1249V	COSM6498418	0.057228	0.074445	D100	PID_0360
chr20	31022444	G	C	ASXL1	silent	p.G643G	NA	0.609968	0.650678	Pre	PID_0360

chr20	31022444	G	C	ASXL1	silent	p.G643G	NA	0.064838	0.070318	D30	PID_0360
chr20	31022444	G	C	ASXL1	silent	p.G643G	NA	0.094106	0.115806	D100	PID_0360
chr21	36252865	C	T	RUNX1	missense	p.R139Q	COSM36055	0.005559	0.003582	Pre	PID_0360
chr21	36252865	C	T	RUNX1	missense	p.R139Q	COSM36055	0.028548	0.027494	D100	PID_0360
chr3	47125604	A	G	SETD2	missense	p.M1845T	NA	0.501668	0.513251	Pre	PID_0360
chr3	47125604	A	G	SETD2	missense	p.M1845T	NA	0.03326	0.031476	D30	PID_0360
chr3	47125604	A	G	SETD2	missense	p.M1845T	NA	0.066519	0.069635	D100	PID_0360
chr7	148523590	C	T	EZH2	missense	p.R249Q	COSM1449004	0.00326	0.004129	Pre	PID_0360
chr7	148523590	C	T	EZH2	missense	p.R249Q	COSM1449004	0.014485	0.015988	D100	PID_0360
chr16	30735348	C	G	SRCAP	missense	p.P1535A	NA	0.491278	0.505079	Pre	PID_0347
chr16	30735348	C	G	SRCAP	missense	p.P1535A	NA	0.013502	0.012958	D30	PID_0347
chr16	30735348	C	G	SRCAP	missense	p.P1535A	NA	0.02768	0.030115	D100	PID_0347
chr17	7578535	T	C	TP53	missense	p.K93R	COSM3388223	0.002619	0.002549	Pre	PID_0347
chr17	7578535	T	C	TP53	missense	p.K93R	COSM3388223	0.003423	0.003649	D30	PID_0347
chr17	7578406	C	T	TP53	missense	p.R43H	COSM10648	0.001863	0.000881	Pre	PID_0347
chr17	7578406	C	T	TP53	missense	p.R43H	COSM10648	0.002412	0.001305	D30	PID_0347
chr17	7578406	C	T	TP53	missense	p.R43H	COSM10648	0.001066	0.002532	D100	PID_0347
chr17	7578263	G	A	TP53	nonsense	p.R64X	COSM3378446	0.040574	0.037321	Pre	PID_0347
chr17	7578263	G	A	TP53	nonsense	p.R64X	COSM3378446	0.016537	0.012518	D30	PID_0347
chr17	7578263	G	A	TP53	nonsense	p.R64X	COSM3378446	0.052481	0.039326	D100	PID_0347
chr2	198266834	T	C	SF3B1	missense	p.K700E	COSM84677	0.029823	0.030639	Pre	PID_0347
chr2	198266834	T	C	SF3B1	missense	p.K700E	COSM84677	0.01055	0.007808	D30	PID_0347
chr2	198266834	T	C	SF3B1	missense	p.K700E	COSM84677	0.043792	0.031544	D100	PID_0347
chr17	29677184	C	T	NF1	intronic	NA	NA	0.014122	0.013542	D30	PID_0335
chr17	29677184	C	T	NF1	intronic	NA	NA	0.501188	0.487533	Pre	PID_0335
chr11	32417962	A	G	WT1	intronic	NA	COSM6494031	0.506841	0.483758	Pre	PID_0268
chr11	32417962	A	G	WT1	intronic	NA	COSM6494031	0.00479	0.005205	D30	PID_0268
chr11	32417962	A	G	WT1	intronic	NA	COSM6494031	0.00144	0.001023	D100	PID_0268
chr17	7727547	C	T	DNAH2	missense	p.R3863C	NA	0.497982	0.518192	Pre	PID_0268
chr17	7727547	C	T	DNAH2	missense	p.R3863C	NA	0.004903	0.00291	D30	PID_0268
chr17	7727547	C	T	DNAH2	missense	p.R3863C	NA	0.002456	0.002404	D100	PID_0268
chr2	25991798	A	G	ASXL2	intronic	NA	NA	0.559502	0.508356	Pre	PID_0268
chr2	25991798	A	G	ASXL2	intronic	NA	NA	0.004492	0.003673	D30	PID_0268
chr2	25991798	A	G	ASXL2	intronic	NA	NA	0.003618	0.000983	D100	PID_0268
chr9	139399343	C	T	NOTCH1	silent	p.L1600L	NA	0.506623	0.51215	Pre	PID_0268
chr9	139399343	C	T	NOTCH1	silent	p.L1600L	NA	0.005394	0.003518	D30	PID_0268
chr9	139399343	C	T	NOTCH1	silent	p.L1600L	NA	0.003461	0.002481	D100	PID_0268

Table 4.6: Analysis of recipient clinical outcomes in relation to engraftment of donor-derived mutations.

Mutation engraftment variable	Disease characteristic	Test	p-value
Pathogenic donor mutation	Acute GvHD	Fisher's Exact test	0.54
	Cytopenia	Fisher's Exact test	0.24
	Duration of cytopenia	Wilcoxon rank-sum test	0.28
	Cumulative incidence of of chronic GvHD	Fine-Gray subdistribution hazard model, Gray's test	0.22
	Leukemia Free Survival	Kaplan-Meier model, log-rank test	0.57
	CMV reactivation	Fisher's Exact test	0.092
	Cardiac event	Fisher's Exact test	0.99
	Mixed chimerism	Repeated measure logistic regression	0.23
	Neutrophil engraftment	No test, all engrafted	-
Persistent donor engraftment	Acute GvHD	Fisher's Exact test	0.23
	Cytopenia	Fisher's Exact test	0.21
	Duration of cytopenia	Wilcoxon rank-sum test	0.054
	Cumulative incidence of of chronic GvHD	Fine-Gray subdistribution hazard model, Gray's test	0.23
	Leukemia Free Survival	Kaplan-Meier model, log-rank test	0.72
	CMV reactivation	Fisher's Exact test	0.099
	Cardiac event	Fisher's Exact test	0.57
	Mixed chimerism	Repeated measure logistic regression	0.33
	Neutrophil engraftment	No test, all engrafted	-
Persistent engraftment of COSMIC-related donor mutation	Acute GvHD	Fisher's Exact test	0.5
	Cytopenia	Fisher's Exact test	0.99
	Duration of cytopenia	Wilcoxon rank-sum test	0.2
	Cumulative incidence of of chronic GvHD	Fine-Gray subdistribution hazard model, Gray's test	0.14
	Leukemia Free Survival	Kaplan-Meier model, log-rank test	0.43
	CMV reactivation	Fisher's Exact test	0.33
	Cardiac event	Fisher's Exact test	0.55
	Mixed chimerism	Repeated measure logistic regression	0.65
	Neutrophil engraftment	No test, all engrafted	-

Chapter 5: The clonal hematopoietic spectrum of Down syndrome with and without myeloid leukemia

5.1 Introduction

Down syndrome (DS) is a genetic disorder characterized by the presence of an extra copy of chromosome 21 (Trisomy 21). Approximately 1 in 800 newborns are affected by this disorder[148], and these newborns suffer from several cognitive and physical abnormalities during development such as having a flattened face, almond-shape eyes, poor muscle tone and intellectual disability[149]. Children with Down syndrome are also associated with several perinatal clinical conditions such as gastrointestinal malfunctions[150], spinal defects[151], endocrine disorders[152], and heart defects[153]. In addition, children with Down syndrome (DS) have a 150-fold higher risk of developing myeloid leukemia (ML-DS), often of the megakaryocyte lineage almost exclusively <4 years of age[154], despite having lower incidence of other solid malignancies[155]. It has been postulated that the higher risk of ML-DS is the consequence stemming from genomic instability in DS blood cells caused by a disturbance in the metabolism of oxygen radicals[156]. This in turn increases the cells' sensitivity to various forms of somatic mutagenesis[157,158]. Expectedly, the blasts of ML-DS are typically characterized by truncating somatic mutations in the X-linked transcription factor, *GATA1*, which is involved in regulating normal blood formation and differentiation[159,160], and the affected children with *GATA1* mutations often display an inhibited ability to drive terminal erythroid and megakaryocytic differentiation in definitive hematopoiesis[161].

ML-DS is subsequently found to be preceded by transient abnormal myelopoiesis (TAM) which is a pre-leukemic condition that is self-remitting by 6 months of age in most cases[162,163]. However, approximately 20% of DS children with TAM would eventually develop ML-DS, and these children are characterized by a positive mutation status for *GATA1*[164,165]. Recently, Roberts and colleagues[166] showed that approximately 30% of healthy DS (HDS) neonates (i.e. without leukemia or any discernable hematologic features) would also acquire truncated *GATA1* mutations at low VAFs in the fetal hematopoietic stem and progenitor cells (HSPCs) consistent with clonal events, suggesting inherently perturbed hematopoiesis in these children as a prerequisite for subsequent ML-DS development. Along with *GATA1* mutations, children with TAM and ML-DS also have a distinct mutation landscape implicating putatively driver events in cohesin complex genes such as *RAD21* and *STAG2*[167]. However, very little is known about the physiologic clonal hematopoiesis (CH) in HDS or even the leukemia-associated clonal hematopoiesis in ML-DS beyond these genes. We hypothesize that CH in ML-DS differs from HDS, and that the differences would not only inform risk stratification in disease surveillance, but also add insight into how trisomy 21 impacts the acquisition of selection of various mutations in DS hematopoietic progenitors.

Given recent observations that rare hematopoietic clonal mutations with variant allele frequency (VAF) ≥0.005 are clinically informative[30,99,], we opted to use targeted error-corrected sequencing (ECS) enriching for 80 genes frequently mutated in adult and pediatric AML as previously described in Chapter 2[102] in order to detect and compare clonal mutations at ≥0.0005 variant allele frequency (VAF) in ML-DS and HDS cohorts. We found that ML-DS children in our study had higher mutation burden than HDS, and mutations in genes such as *GATA1*, *EZH2*, *RAD21* and *STAG2* were enriched in ML-DS, supporting previous observations. Interestingly, despite their prevalence, many *GATA1* mutations in ML-DS had low VAFs relative to the blast counts such that these *GATA1*-mutated clones would be considered subclonal[168]. In addition, we found that the mutation landscape between younger and older HDS children was different (demarcated at age 4 – the age at which the risk of developing ML-DS decreases significantly[155]), and older HDS children were shown to display a more 'adult-like' signature as demonstrated in clonal hematopoiesis in adults at age 50s – 60s[93].

5.2 Materials and Methods

5.2.1 Study Population

A total of 46 ML-DS patients from the Children's Oncology Group (COG) AAML1531 trial (all <4 years old), and 63 non-leukemia HDS patients (23 <4 years old and 40 ≥4 years old) were included in this study. The ML-DS samples were collected at diagnosis of leukemia, and the white blood cell DNA were shipped to the laboratory. Upon receipt, the DNA samples were kept in -20°C freezer until library preparation. The HDS peripheral blood samples were collected at St. Louis Children's Hospital with informed consent from the parents of the patients. Upon collection, genomic DNA was extracted from the white blood cells of HDS samples using Qiagen DNeasy Blood and Tissue Kit following manufacturer's recommendations. The final DNA elution volume was 50 μl, and the DNA were stored in -20°C until library preparation.

5.2.2 ECS library preparation and mutation analysis

Approximately 200-250 ng of DNA per sample was used to make ultra-deep error-corrected sequencing libraries via the custom-made Illumina TruSeq Custom Amplicon kit described in Chapter 2. The library preparation followed the protocol outlined in Chapter 2 closely with several modifications. First, following ddPCR quantification step, each library was then normalized to 6.3 million UMI-tagged molecules, and a second round of PCR (14 cycles) was performed in a 50 µl reaction: 25 µL of Q5 master mix, 2 µL of P5 Primer (1 µM), 2 µL of P7 Primer (1 µM), and 21 µL of DNA molecules. After that, the amplified libraries were purified, and the libraries were normalized. Six purified libraries were pooled and sequenced in an Illumina NextSeq 550 High-Output Kit with the following settings: 2x144 PE, 8 cycles Index 1, 16 cycles Index 2 (account for 16N random bases used as UMI). For each sample, a technical replicate library was prepared via the same protocol. In total, 109 samples were processed, and 218 ECS libraries were prepared.

The bioinformatics processing followed the pipeline outlined in Chapter 2 closely with additional filters. The retained single nucleotide variants (SNVs) after Bonferroni correction[102] were further curated in a stepwise manner to remove potential false positive calls: 1) SNVs called due to batch effect were removed, 2) non-hotspot variants identified in more than one sample were removed, 3) SNVs that had a coefficient of variation >15% between 5-reads and 3-reads error-correction were removed. Indels were called via VarScan2 with the aligned consensus reads using mpileup2indel setting, and were further curated as described above. Additional adjustment was to survey the genomic region of *GATA1*. Some samples had low coverage at the *GATA1* region (~300x consensus coverage), and the coverage filter was lowered from 700x to 200x specifically at this locus in order to more accurately capture the signals. The SNVs and/or Indels identified would still be required to pass the statistical frameworks outlined above. Since majority of *GATA1* variants were Indels spanning from exon 2 to exon 3 (including the intronic region)[160], we further customized a single-amplicon ECS approach to survey these regions for additional variants that were possibly missed by the gene panel approach.

5.2.3 Sequencing of ML-DS at end-of-induction treatment to verify somatic nature of *GATA1* mutations detected at diagnosis.

GATA1 mutations with VAFs at approximately 0.3 – 0.6 could be germline variants inherited from either parent. In order to ascertain the somatic nature of these high VAF variants, we sequenced the DNA of corresponding ML-DS children collected at end-of-induction using our ECS approach described previously. We reasoned that somatic *GATA1* mutations present in blast cells would display a decrease in VAF after end-of-induction treatment.

5.3 Results

5.3.1 Difference in mutation landscape between ML-DS and age-matched HDS

First, given the age-related risk of ML-DS (<4 years)[155], we compared CH in 46 ML-DS patients (median: 1.77 years; range: 0.78 – 3.54 years) from the Children's Oncology Group AAML1531 trial to 23 age-matched HDS children from St. Louis Children's Hospital (median: 1.42 years; range: 0.18 – 3.15). We detected CH in 84.8% of ML-DS patients and 52.2% of age-matched HDS children at median VAFs of 0.0023 and 0.002 (range: 0.0005 – 0.61175 and 0.0005 – 0.0716) respectively by ECS gene panel sequencing (Table 5.1 – 5.2). The prevalence of somatic mutations was significantly higher in ML-DS (p-value=0.0075, two-sided Fisher's exact test; Figure 5.1A). *GATA1* mutations were detected in 80.5% of the ML-DS cohort (Figure 5.1B), and a majority of these mutations were frameshift Indels (Figure 5.1C), supporting previous observations in ML-DS[160]. In addition to *GATA1* mutations, ML-DS was also enriched for mutations in histone modifiers, particularly in *EZH2* (p-value=0.025, two-sided Fisher's exact test) and in cohesin complex genes such as *STAG2* and *RAD21* (p-value=0.0059, two-sided Fisher's exact test) (Figure 5.1C). There was a possibility that *GATA1* mutations detected at

VAFs ranging approximately from 0.3 – 0.6 were germline variants. Sequencing of end-of-induction samples showed that these *GATA1* mutations displayed a marked decrease in VAF (Figure 5.2), definitively proving that these were somatic mutations. We also observed that mutations in tumor suppressors were only present in ML-DS and not in the age-matched HDS<4 group. The mutation burden per individual and the mutation VAFs were also significantly higher in ML-DS (Figure 5.3). In contrast, the epigenetic modifiers *DNMT3A* and *TET2*, commonly mutated in adult CH, were rarely mutated in either cohort.

5.3.2 Subclonal nature of *GATA1*-mutated clones in ML-DS

When we examined the VAFs in further detail, we noted that the *GATA1* mutations had a wide VAF range (0.0013 – 0.6117 VAF), with 57.1% of all detected variants being <0.1 VAF (median: 0.046). In this ML-DS cohort that consisted entirely of diagnosis samples, a cognate driver mutation in the founding clone (i.e. heterozygous in female and hemizygous in male for a X-linked *GATA1* mutation)[159,160] would be expected to have a hemizgosity-adjusted VAF at ≥0.8 of the total blast count in a patient[168] if the disease progression follows a linear clonal acquisition model. Our data, however, suggests that many *GATA1* mutations were maintained at low frequency as minor clones after they were acquired (Figure 5.4A, green dots). When we examined if there could be another major clone besides these *GATA1*-mutated minor clones, we found that a majority of patients with minor *GATA1* clone (70.6%) did not have any other co-occurring somatic mutations with high VAF in other candidate genes included in our sequencing panel. However, they were significantly enriched for specific germline variants associated with hematologic malignancies in genes such as *TP53* and *CSF3R* (p-value=0.035, one-sided Fisher's exact test; Figure 5.4B; Table 5.3).

5.3.3 Difference in mutation landscape between younger and older HDS

Next, we compared CH between younger and older HDS children (HDS<4yo; n=23 and HDS≥4yo; n=40) to characterize physiologic CH in DS as well as delineation of ML-DS risk at 4 years of age. We detected CH in 70% of HDS≥4 versus 52.2% of HDS<4 at median VAFs of 0.001 and 0.002 (range: 0.0005 – 0.01047 and 0.0005 – 0.0716) respectively (Table 5.4), a difference that was not statistically different. However, the genes implicated in CH in HDS≥4 were distinct from those in HDS<4. Although statistically insignificant (p-value=0.08, one-sided Fisher's exact test), the HDS≥4 cohort possessed more mutations in canonical adult CH genes[64,169] (Figure 5.5). These adult CH genes, such as *ASXL1* and *TET2*, were also not recurrently mutated in ML-DS. Overall, our results also revealed that clonal mutations with >0.0005 VAF are common (an overall prevalence of 63.5%) amongst HDS children <20s (median age 6.49), an age group where even CH ≥0.02 VAF was virtually non-detectable in non-DS individuals in previous studies using standard NGS[60-62].

5.4 Discussion

Our data corroborated with the findings in previous studies where *GATA1*, other genes involved in epigenetic modifications and cohesin complex were recurrently mutated in the blood of ML-DS[167,170]. Importantly, the same set of genes was not recurrently mutated in age-matched HDS individuals, demonstrating a divergence in mutation landscape of CH between ML-DS and HDS. This supports the hypothesis that neonatal hematopoietic clones must acquire additional somatic mutations beyond *GATA1* in a defined set of putative driver genes for ML-DS transformation during a strict postnatal developmental window before 4 years of age[155], such that clones not harboring mutations that are contextually important within the critical developmental window would not be sustained[171]. This is further substantiated by a previous finding that showed approximately 95% of HDS (17 of 18) neonates with low VAF *GATA1* mutations did not eventually developed ML-DS[166], suggesting that additional mutations must be acquired for clonal expansion and malignant transformation.

We noted that many previous studies rarely discuss the actual VAF of *GATA1* mutations in ML-DS patients due to the fact that enrichment of *GATA1* mutations was done by denaturing high-performance liquid chromatography followed by direct Sanger sequencing, which reveals the presence or absence of a mutation with a detection limit of ~5%, but offers relatively low sensitivity to the true diagnostic VAF estimation[159,160,166]. Other studies used whole-genome or whole-exome sequencing approaches[167,172], that have a limit of detection at around 0.02 VAF, therefore also precluding comprehensive examination of low VAF mutations. Our ECS approach enabled us to estimate VAF of rare subclonal mutations in an unbiased fashion. With this, we showed that more than half of the *GATA1*-mutated clones were minor subclones. This is consistent with prior data suggesting that minor *GATA1* mutated clones might exert their impact in malignant transformation via non-cell-autonomous regulation[173]. Specifically, we posit that the minor *GATA1* mutated clones create a perturbed hematopoietic clonal architecture by modulating the production and secretion of cytokines[174]. This would in turn promote oligoclonal cooperation[175,176]. In fact, these *GATA1*-mutated minor clones were significantly enriched for germline variants associated with hematologic malignancies in genes such as *TP53* and *CSF3R*, suggesting a potential mechanism of functional cooperativity in malignant transformation. Recently, Labuhn and colleages (2019)[172] demonstrated the cooperative nature between *Gata1* mutations and several other clonal variants in genes such as *Trp53* in murine models to promote oncogenic transformation from transient abnormal myelopoiesis to ML-DS.

When we examined the mutation landscape of CH between younger and older HDS individuals, we found that older HDS individuals displayed a more ‘adult-like’ CH signature. Interestingly, this shift in mutation landscape seemed to coincide with the age-related change in ML-DS risk in DS children, suggesting that the hematopoietic clones present during the risk window of ML-DS has either become quiescent or been exhausted by age 4. These results also support the notion that certain genes exert varying effect sizes during different developmental time windows[177], and mutations in genes that are functionally relevant during a specific context would be more prone to selection pressures. In addition, it has been previously demonstrated DS patients exhibit an accelerated aging phenotype in blood cells[178]. A separate study that examined CH in healthy blood donors (median age 26) using the same gene panel found that 44% of their cohort harbored at least one somatic mutation[129]. Despite a much younger cohort, CH was found in approximately 20% higher number of individuals in the HDS group. Therefore, it stands to reason that the higher-than-expected prevalence of CH could partially be the result of this accelerated aging where mutations would be more frequent.

In summary, we presented a comprehensive characterization of the leukemia-associated and physiologic CH in DS children with or without ML-DS. Our data provide insight into the role of *GATA1* mutated clones in ML-DS and age-related leukemia risk in DS based on CH. As such, detailed molecular and mechanistic studies are warranted to assess the utility of detecting ML-DS specific mutations as early detection biomarkers, and to elucidate the pathogenicity of oligoclonal interactions.

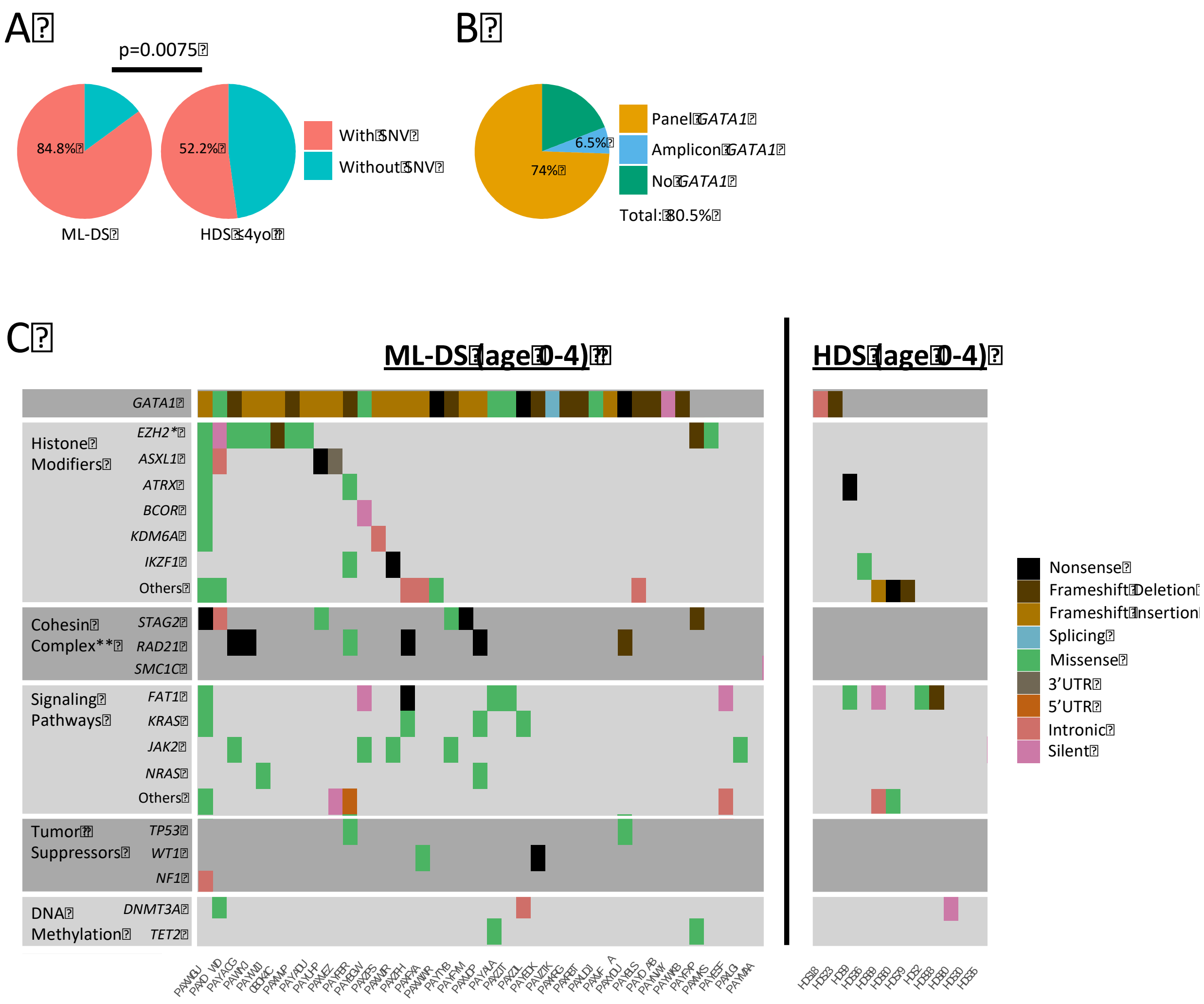


Figure 5.1: Characteristics of clonal mutations detected by ECS in ML-DS and HDS ≤4yo. (A) Proportion of individuals with detected SNVs. ML-DS cohort was significantly more likely to harbor somatic mutations(s) compared to HDS ≤4yo. (B) Prevalence of *GATA1* mutations in ML-DS using different ECS methods. (C) Mutation spectrum of ML-DS and HDS ≤4yo as detected by ECS-panel. *EZH2* mutations were significantly enriched in ML-DS (denoted as *; p-value=0.025; two-sided Fisher's exact test). Mutations in cohesin complex genes also were significantly enriched in ML-DS compared to HDS ≤4yo (denoted as **; p-value=0.0059; two-sided Fisher's exact Test).

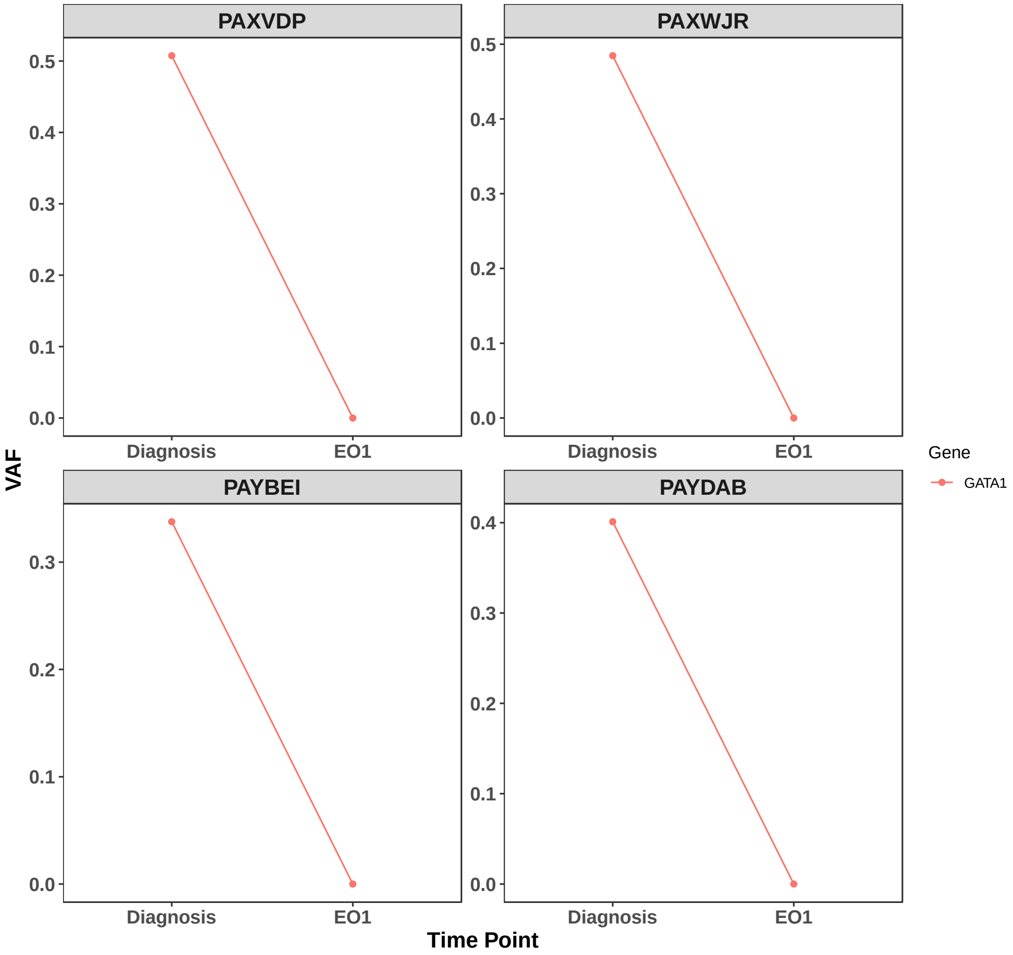


Figure 5.2: *GATA1* mutations detected at diagnosis with 0.3 – 0.6 VAF decreased in frequency at end-of-induction (EO1).

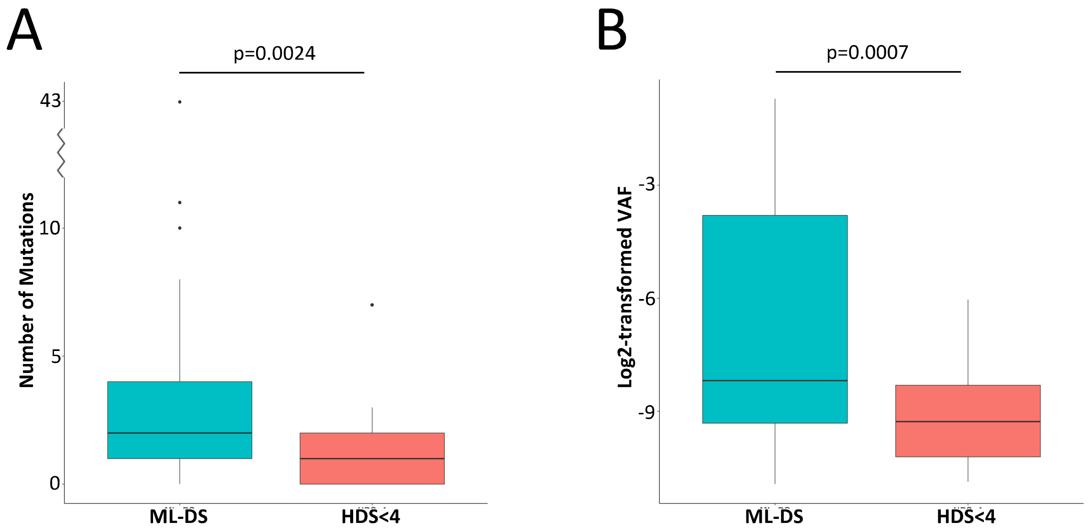


Figure 5.3: Mutation burden and mutation VAFs were significantly higher in ML-DS relative to age-matched HDS. (A) Number of somatic mutations per individual in ML-DS and HDS<4. ML-DS patients had significantly higher number of somatic mutations compared to HDS<4 patients (p-value=0.0024, two-sided Wilcoxon rank-sum test). (B) Log2-tranformed VAFs of somatic mutations in ML-DS and HDS<4. Mutations in ML-DS had significantly higher VAFs compared to those in HDS<4 (p-value=0.0007, two-sided Wilcoxon rank-sum test).

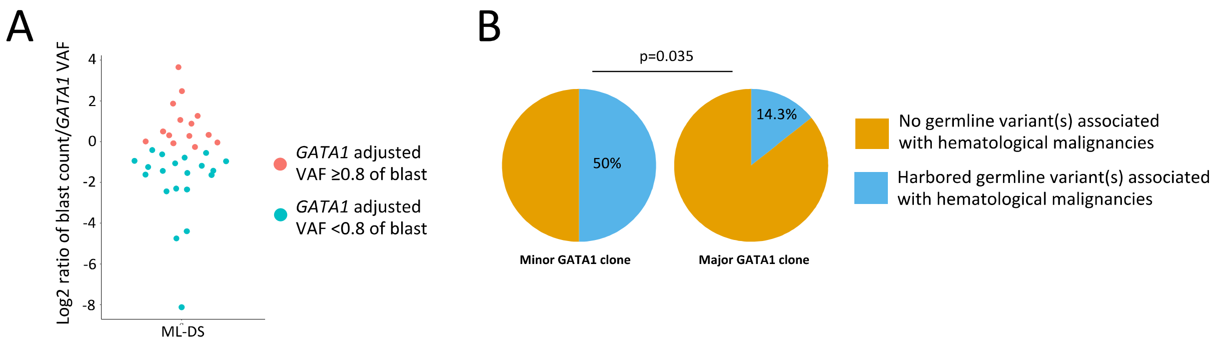


Figure 5.4: Characterization of *GATA1*-mutated clones in ML-DS. (A) Log2 ratio of blast count to VAF of detected *GATA1* mutations in ML-DS. *GATA1* mutations with hemizygosity-adjusted VAF at ≥0.8 of blast count were considered somatic drivers. (B) Presence or absence of germline variants associated hematological malignancies in ML-DS patients with either minor or major *GATA1* clones. Patients with minor *GATA1* clone were significantly enriched for pathological germline variants in hematological malignancies (one-sided Fisher's exact test).

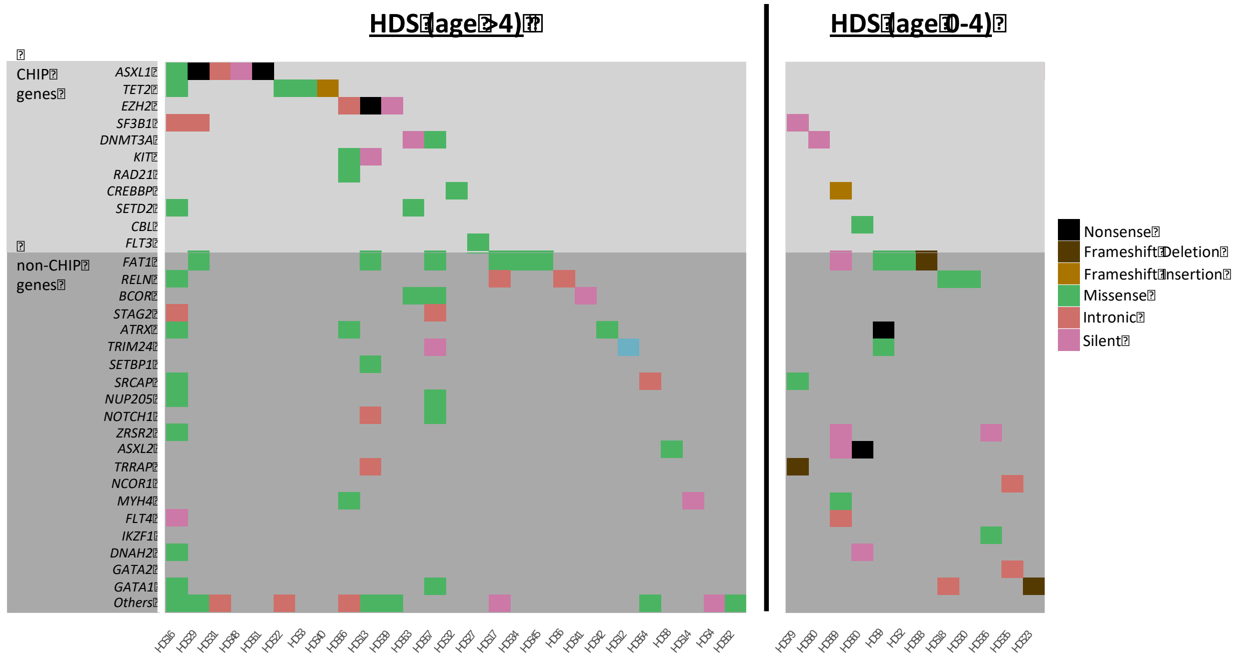


Figure 5.5: Characteristics of clonal mutations detected by ECS in HDS >4yo and HDS ≤4yo.

Table 5.1: Detected somatic mutations in ML-DS via ECS in the 80-gene panel.

Chr	Start	Ref	Alt	Gene	Type	AA_change	cosmic84	VAF1	VAF2	SampleID
chrX	15838398	G	T	ZRSR2	missense	p.C299F	NA	0.003407	0.001749	PAYMAA
chrX	15826455	G	T	ZRSR2	intronic	NA	NA	0.001391	0.001754	PAXDWD
chr11	32414263	G	A	WT1	nonsense	p.R413X	COSM5879212	0.02517	0.024555	PAYEDK
chr11	32414268	C	T	WT1	missense	p.C411Y	NA	0.019362	0.018562	PAXFXA
chr2	61456161	T	A	USP34	intronic	NA	NA	0.001348	0.001288	PAYESF
chr2	61434005	T	G	USP34	missense	p.H2979P	NA	0.000655	0.001632	PAXWGU
chr2	61416091	C	T	USP34	silent	p.Q3329Q	NA	0.002099	0.002442	PAXWGU
chr7	98555680	C	T	TRRAP	missense	p.T2078I	NA	0.001249	0.001712	PAXWWR
chr7	138263867	A	G	TRIM24	intronic	NA	NA	0.016851	0.008269	PAYDAB2
chr17	7577058	C	T	TP53	missense	p.E162K	COSM44127	0.001689	0.001571	PAYFBR
chr17	7577556	C	G	TP53	missense	p.C110S	COSM1610838	0.001573	0.003146	PAXVDU
chr4	106155325	C	T	TET2	missense	p.P76S	NA	0.00084	0.000639	PAYFXP
chr4	106157266	C	A	TET2	missense	p.P723T	NA	0.000515	0.00051	PAYALA
chrX	123210285	G	T	STAG2	missense	p.M879I	NA	0.002242	0.001761	PAYUHP
chrX	123197045	G	C	STAG2	missense	p.R604P	NA	0.08064	0.090841	PAYTYB
chrX	123179197	C	T	STAG2	nonsense	p.R216X	COSM1315170	0.224804	0.242605	PAYFYM
chrX	123196997	T	A	STAG2	nonsense	p.L588X	COSM164630	0.001765	0.001812	PAXWGU
chrX	123220692	T	A	STAG2	intronic	NA	NA	0.002255	0.002548	PAXDWD
chr17	74732959	G	C	SRSF2	missense	p.P95R	COSM211661	0.057782	0.05022	PAYDAB1
chr16	30734891	C	T	SRCAP	intronic	NA	NA	0.001657	0.001249	PAXWGU
chr16	30734639	T	A	SRCAP	intronic	NA	NA	0.001169	0.001539	PAXWGU
chr16	30735969	C	A	SRCAP	missense	p.P1742T	NA	0.00061	0.001069	PAXWGU
chr16	30735315	C	A	SRCAP	missense	p.P1524T	NA	0.000541	0.000626	PAXDWD
chr11	47381625	T	A	SPI1	intronic	NA	NA	0.001436	0.002198	PAYESF
chr3	47144958	A	T	SETD2	intronic	NA	NA	0.001299	0.001402	PAXWGU
chr3	47158212	C	T	SETD2	missense	p.R1452Q	COSM1045469	0.00136	0.001805	PAXWGU
chr3	47125361	A	T	SETD2	missense	p.L1926Q	NA	0.002687	0.0039	PAXWGU

chr3	47166117	C	T	SETD2	intronic	NA	NA	0.311767	0.333929	PAXFXA
chr7	103214598	G	T	RELN	nonsense	p.Y1484X	NA	0.00133	0.001489	PAYFBR
chr7	103205946	A	G	RELN	silent	p.F1663F	NA	0.000694	0.000674	PAYFBR
chr7	103124180	G	T	RELN	missense	p.N3367K	NA	0.000743	0.000961	PAXDWD
chr8	117878847	A	C	RAD21	missense	p.V41G	NA	0.192318	0.194406	PAYFBR
chr8	117878960	G	T	RAD21	nonsense	p.Y3X	COSM3663527	0.215709	0.219335	PAYACG
chr8	117878960	G	T	RAD21	nonsense	p.Y3X	COSM3663527	0.213149	0.232356	PAXZFH
chr8	117875483	G	A	RAD21	missense	p.R54W	COSM3412707	0.012888	0.011492	PAXVDU
chr8	117878960	G	C	RAD21	nonsense	p.Y3X	COSM1738123	0.243786	0.270337	PAXVDP
chr8	117878960	G	C	RAD21	nonsense	p.Y3X	COSM1738123	0.027805	0.030065	PAWIYJ
chr9	134072839	C	A	NUP214	missense	p.L146I	NA	0.001625	0.001925	PAXDWD
chr1	115258748	C	T	NRAS	missense	p.G12S	COSM563	0.017605	0.018377	PAYWJJ
chr1	115258748	C	A	NRAS	missense	p.G12C	COSM562	0.263815	0.260224	PAXVDP
chr9	139399532	C	T	NOTCH1	silent	p.K1537K	COSM4548968	0.001781	0.003386	PAXWGU
chr17	29533417	C	T	NF1	intronic	NA	NA	0.005576	0.002833	PAXWGU
chr17	15971283	C	T	NCOR1	missense	p.A1572T	NA	0.002482	0.002886	PAXWGU
chr17	10432291	C	A	MYH2	nonsense	p.E1154X	NA	0.001122	0.000918	PAYESF
chr17	10427952	C	T	MYH2	missense	p.S1669N	NA	0.002043	0.002242	PAXWGU
chr16	15835583	C	A	MYH11	intronic	NA	NA	0.000755	0.000809	PAYFBR
chr16	15814795	C	T	MYH11	silent	p.R1564R	NA	0.001065	0.000978	PAXWGU
chr8	128752960	A	G	MYC	missense	p.H374R	COSM1096012	0.000928	0.001102	PAYTYB
chr1	43814979	G	A	MPL	missense	p.S505N	COSM27286	0.015627	0.016379	PAYTYB
chr1	43814979	G	A	MPL	missense	p.S505N	COSM27286	0.311336	0.287269	0BOK4C
chr18	7012089	C	T	LAMA1	missense	p.A1138T	COSM4749391	0.001919	0.001714	PAXWGU
chr18	6980579	G	T	LAMA1	missense	p.A1983D	NA	0.002077	0.00377	PAXWGU
chr12	25378562	C	T	KRAS	missense	p.A146T	COSM1165198	0.014571	0.023685	PAXZJL
chr12	25378561	G	A	KRAS	missense	p.A146V	COSM19900	0.017878	0.013241	PAXZFH
chr12	25380279	C	A	KRAS	missense	p.G60V	COSM5879374	0.007211	0.002402	PAXWGU
chr12	25380285	G	A	KRAS	missense	p.T58I	COSM87288	0.002398	0.00213	PAXVDP
chr11	118353218	G	C	KMT2A	intronic	NA	COSM6493997	0.328403	0.323526	PAYLBLS
chr4	55575675	G	A	KIT	missense	p.A401T	NA	0.000843	0.00182	PAXWGU

chrX	44913303	T	C	KDM6A	intronic	NA	NA	0.0008	0.001696	PAXZPS
chrX	44920614	C	T	KDM6A	missense	p.P163S	NA	0.001507	0.00193	PAXWGU
chr9	5073770	G	T	JAK2	missense	p.V617F	COSM12600	0.023441	0.027417	PAYTYB
chr9	5073770	G	A	JAK2	missense	p.V617I	COSM29117	0.003066	0.004521	PAYBGW
chr9	5073770	G	T	JAK2	missense	p.V617F	COSM12600	0.095274	0.072727	PAYACG
chr9	5073770	G	T	JAK2	missense	p.V617F	COSM12600	0.287436	0.263685	PAXWJR
chr9	5073785	A	T	JAK2	missense	p.N473Y	NA	0.008395	0.014773	PAXLGI
chr7	50450300	C	T	IKZF1	missense	p.R75W	COSM3265410	0.186535	0.189081	PAYFBR
chr7	50455075	C	T	IKZF1	nonsense	p.R65X	NA	0.019345	0.023815	PAXWJR
chr2	209113452	G	A	IDH1	intronic	NA	NA	0.00476	0.003457	PAXWGU
chr2	209113116	C	T	IDH1	missense	p.G131S	NA	0.002084	0.001889	PAXDWD
chr11	534346	G	A	HRAS	UTR5	NA	COSM308584	0.010766	0.012829	PAYFBR
chr4	153245418	G	A	FBXW7	silent	p.D473D	NA	0.001302	0.000626	PAYDAB1
chr4	153249482	G	A	FBXW7	silent	p.N314N	NA	0.002594	0.003157	PAXWGU
chr4	187558030	G	A	FAT1	silent	p.T1227T	NA	0.003808	0.002717	PAYESF
chr4	187629067	A	G	FAT1	missense	p.S639P	NA	0.000651	0.000661	PAYDAB1
chr4	187539568	G	A	FAT1	silent	p.L2724L	NA	0.002348	0.001678	PAYBGW
chr4	187541686	G	T	FAT1	missense	p.N2018K	NA	0.316843	0.309846	PAYALA
chr4	187540173	T	A	FAT1	missense	p.I2523F	NA	0.000666	0.000773	PAXZJT
chr4	187629967	T	A	FAT1	nonsense	p.K339X	COSM3714775	0.00061	0.000598	PAXZFH
chr4	187542704	G	A	FAT1	missense	p.T1679I	NA	0.001515	0.002328	PAXWGU
chr4	187542648	C	A	FAT1	missense	p.V1698L	NA	0.001406	0.002016	PAXWGU
chr7	148515265	A	G	EZH2	intronic	NA	NA	0.000694	0.000782	PAYWJJ
chr7	148508788	C	T	EZH2	missense	p.V570M	COSM3942110	0.029021	0.031481	PAYWJJ
chr7	148523605	G	A	EZH2	missense	p.T244M	NA	0.002312	0.001706	PAYADU
chr7	148506468	C	T	EZH2	missense	p.A626T	NA	0.132196	0.10567	PAYACG
chr7	148512098	G	T	EZH2	missense	p.P483H	NA	0.035277	0.038133	PAXWGU
chr7	148526910	G	A	EZH2	missense	p.P93S	COSM133047	0.032791	0.032848	PAXMVP
chr7	148543620	C	T	EZH2	missense	p.R63Q	NA	0.317843	0.24761	PAXMKS
chr7	148526908	A	G	EZH2	silent	p.P93P	COSM5020984	0.313955	0.333721	PAXDWD
chr7	148508731	C	T	EZH2	missense	p.E589K	COSM5945116	0.02569	0.02914	PAWIYJ

chr12	12022740	G	A	ETV6	silent	p.R282R	NA	0.001606	0.002029	PAXWGU
chr2	25459912	G	T	DNMT3A	intronic	NA	NA	0.001196	0.001842	PAXZJL
chr2	25505501	T	A	DNMT3A	missense	p.E86V	NA	0.000785	0.000924	PAXDWD
chr17	11543680	A	T	DNAH9	missense	p.N627I	COSM6023401	0.001185	0.001041	PAYDAB1
chr17	11797799	T	C	DNAH9	missense	p.W110R	NA	0.000869	0.001681	PAXWGU
chr17	7637813	G	T	DNAH2	missense	p.Q255H	NA	0.00253	0.001387	PAXWGU
chr17	7681745	G	A	DNAH2	silent	p.L1833L	NA	0.00061	0.000773	PAXDWD
chr16	3801831	G	T	CREBBP	intronic	NA	NA	0.000811	0.000607	PAXZFH
chr16	3779441	C	T	CREBBP	missense	p.M1831I	NA	0.001363	0.001352	PAXWGU
chr6	75800115	C	A	COL12A1	intronic	NA	NA	0.000938	0.000768	PAYWJJ
chr11	119169169	C	T	CBL	silent	p.L785L	NA	0.001791	0.002023	PAXWGU
chr11	119148908	C	T	CBL	silent	p.S376S	NA	0.001393	0.004236	PAXVEZ
chr19	13414472	A	T	CACNA1A	intronic	NA	NA	0.001136	0.002163	PAXWGU
chr19	13318208	G	A	CACNA1A	silent	p.H2480H	NA	0.002445	0.001595	0BOK4C
chr7	140453154	T	C	BRAF	missense	p.D594G	COSM467	0.002146	0.002107	PAYDAB1
chrX	39932820	G	A	BCOR	silent	p.S593S	NA	0.000548	0.000813	PAYBGW
chrX	39913340	C	A	BCOR	intronic	NA	NA	0.001741	0.004016	PAXWGU
chrX	39921436	C	T	BCOR	missense	p.A1428T	NA	0.003394	0.003547	PAXWGU
chrX	39911383	C	T	BCOR	silent	p.L1715L	NA	0.002499	0.001863	PAXWGU
chrX	76939296	T	A	ATRX	missense	p.R484S	NA	0.000988	0.000765	PAYFBR
chrX	76849194	G	T	ATRX	silent	p.R2028R	NA	0.001486	0.001366	PAYFBR
chrX	76937612	T	C	ATRX	missense	p.S1046G	NA	0.003139	0.003737	PAYDAB1
chrX	76849477	A	G	ATRX	intronic	NA	NA	0.001139	0.001017	PAXWGU
chrX	76813215	A	T	ATRX	intronic	NA	NA	0.001874	0.00286	PAXWGU
chrX	76972693	C	T	ATRX	silent	p.Q16Q	NA	0.002149	0.002663	PAXWGU
chrX	76938691	T	A	ATRX	missense	p.K686M	NA	0.00068	0.000833	PAXWGU
chrX	76920212	C	T	ATRX	missense	p.D1289N	NA	0.001158	0.001881	PAXWGU
chr2	25991694	C	T	ASXL2	missense	p.R183K	NA	0.001604	0.003019	PAXWGU
chr20	31024371	C	T	ASXL1	nonsense	p.Q1286X	NA	0.008887	0.010263	PAYUHP
chr20	31023745	C	T	ASXL1	missense	p.P1077L	NA	0.001328	0.002097	PAXWGU
chr20	31025204	A	C	ASXL1	UTR3	NA	NA	0.213333	0.181901	PAXVEZ

chr20	31022224	T	C	ASXL1	intronic	NA	NA	0.000914	0.001112	PAXDWD
chr6	138599685	G	T	ARFGEF3	silent	p.L742L	NA	0.001141	0.000811	PAYWJJ
chr6	138551013	C	T	ARFGEF3	intronic	NA	NA	0.002654	0.003427	PAXWGU
chr6	138584112	G	A	ARFGEF3	missense	p.E498K	COSM21854	0.001304	0.002385	PAXWGU
chrX	48649738	T	A	GATA1	splicing	NA	NA	0.023797	0.019476	PAVZTK
chrX	48649605	C	T	GATA1	missense	p.S30L	NA	0.002046	0.009449	PAWIYJ
chrX	48649517	A	G	GATA1	missense	p.M1V	COSM17819	0.03604	0.046159	PAXDWD
chrX	48649519	G	A	GATA1	missense	p.M1I	COSM17822	0.215989	0.196763	PAXUDJ
chrX	48649565	C	T	GATA1	nonsense	p.Q17X	COSM13211	0.041977	0.036555	PAXWWR
chrX	48649553	G	T	GATA1	nonsense	p.E13X	COSM17828	0.216194	0.229826	PAXYDU
chrX	48649715	G	T	GATA1	nonsense	p.E67X	COSM13203	0.19341	0.207524	PAXZJL
chrX	48649519	G	T	GATA1	missense	p.M1I	COSM87863	0.174048	0.192021	PAXZJT
chrX	48649517	A	T	GATA1	missense	p.M1L	NA	0.069693	0.06516	PAYALA
chrX	48649736	G	A	GATA1	missense	p.V74I	COSM17833	0.604794	0.618704	PAYBGW
chrX	48649689	C	A	GATA1	missense	p.A58E	NA	0.239645	0.221273	PAYFYM
chrX	48649706	A	C	GATA1	silent	p.R64R	NA	0.001552	0.000955	PAYNWY
chrX	48649596	C	G	GATA1	missense	p.T27R	NA	0.085195	0.099618	PAYTYB
chrX	48649598	C	G	GATA1	missense	p.P28A	NA	0.085388	0.098456	PAYTYB
chrX	48649684	A	AGCTGCGT	GATA1	frame_shift_ins	NA	NA	0.0446	0.0649	PAXWGU
chrX	48649565	C	CA	GATA1	frame_shift_ins	NA	NA	0.4686	0.5006	PAXWJR
chrX	48649666	G	GAGCACAGCCAC	GATA1	frame_shift_ins	NA	NA	0.1349	0.1421	PAXZFH
chrX	48649665	C	CT	GATA1	frame_shift_ins	NA	NA	0.0273	0.037	PAXZPS
chrX	48649700	TACTAC	T	GATA1	frame_shift_del	NA	NA	0.0554	0.0407	PAYACG
chrX	48649677	C	CCGCTGCAGCGG	GATA1	frame_shift_ins	NA	NA	0.0142	0.0101	PAYADU
chrX	48649686	C	CT	GATA1	frame_shift_ins	NA	NA	0.3334	0.3418	0BOK4C
chrX	48649674	C	CCACCGCTGCAGCT	GATA1	frame_shift_ins	NA	NA	0.4191	0.4459	PAYBHY
chrX	48649655	T	TCCACTGCC	GATA1	frame_shift_ins	NA	NA	0.0821	0.0684	PAYBHY
chrX	48649551	CAG	C	GATA1	frame_shift_del	NA	NA	0.1719	0.1577	PAYBLS
chrX	48649551	CAG	C	GATA1	frame_shift_del	NA	NA	0.4071	0.3949	PAYDAB
chrX	48649574	GA	G	GATA1	frame_shift_del	NA	NA	0.0723	0.0821	PAYEDK
chrX	48649697	GCCTACTACAGGGA	G	GATA1	frame_shift_del	NA	NA	0.1669	0.1963	PAYFBR

chrX	48649688	G	GTAGT	GATA1	frame_shift_ins	NA	NA	0.2358	0.2204	PAYFYM
chrX	48649605	CAG	C	GATA1	frame_shift_del	NA	NA	0.0338	0.0535	PAXMVP
chrX	48649653	CCTCCACTGCCCCGAGCA	C	GATA1	frame_shift_del	NA	NA	0.5139	0.5088	PAXRBT
chrX	48649617	T	TC	GATA1	frame_shift_ins	NA	NA	0.5042	0.5108	PAXVDP
chrX	48649702	C	CT	GATA1	frame_shift_ins	NA	NA	0.1335	0.1791	PAXVEZ
chrX	48649666	G	GAGCACAGC	GATA1	frame_shift_ins	NA	NA	0.1357	0.129	PAXVFA
chrX	48649689	C	CGGCACTGG	GATA1	frame_shift_ins	NA	NA	0.0533	0.0412	PAWIYJ
chrX	48649736	G	GGT	GATA1	frame_shift_ins	NA	NA	0.0443	0.0452	PAXFXA
chrX	48649525	CCCTGG	C	GATA1	frame_shift_del	NA	NA	0.0703	0.0734	PAXKRG
chrX	48649589	TCCTC	T	GATA1	frame_shift_del	NA	NA	0.0839	0.0988	PAYTYB
chrX	48649601	GA	G	GATA1	frame_shift_del	NA	NA	0.0448	0.0524	PAYTYB
chrX	48649556	C	CCCCT	GATA1	frame_shift_ins	NA	NA	0.0619	0.0793	PAYUHP
chrX	48649689	C	CGGCACTGGCCTACTACAGG	GATA1	frame_shift_ins	NA	NA	0.0226	0.0276	PAYWJJ
chrX	48649705	CAGGGACGCTG	C	GATA1	frame_shift_del	NA	NA	0.0194	0.0186	PAYWKB

Table 5.2: Detected somatic mutations in HDS ≤4yo via ECS in the 80-gene panel.

Chr	Start	Ref	Alt	Gene	Type	AA_change	cosmic84	VAF1	VAF2	SampleID
chr7	103131160	C	T	RELN	missense	p.S3187N	NA	0.001188	0.001468	HDS18
chr4	187538996	G	A	FAT1	missense	p.P2915L	NA	0.000924	0.000776	HDS2
chr7	103389876	A	G	RELN	missense	p.I218T	NA	0.004488	0.005849	HDS30
chr3	128206002	G	A	GATA2	intronic	NA	NA	0.005278	0.007939	HDS35
chr17	15971493	T	A	NCOR1	intronic	NA	NA	0.000926	0.000732	HDS35
chr7	50468074	C	A	IKZF1	missense	p.L437M	NA	0.00117	0.00273	HDS36
chrX	15818060	C	T	ZRSR2	silent	p.L63L	NA	0.002199	0.002697	HDS36
chr4	187542884	A	T	FAT1	missense	p.I1619N	NA	0.000636	0.001024	HDS9
chr7	138252230	C	T	TRIM24	missense	p.P512L	COSM2859869	0.016165	0.014275	HDS9
chrX	76888771	A	T	ATRX	nonsense	p.Y1686X	NA	0.001206	0.001461	HDS9
chr2	25965069	G	A	ASXL2	silent	p.G1379G	NA	0.001173	0.000978	HDS69
chr4	187541362	G	A	FAT1	silent	p.H2126H	NA	0.001148	0.000902	HDS69
chr2	25467492	G	A	DNMT3A	silent	p.Y528Y	NA	0.003418	0.008104	HDS80
chr5	180056186	G	A	FLT4	intronic	NA	NA	0.000741	0.000621	HDS69
chr11	119169197	C	A	CBL	missense	p.S794Y	NA	0.002097	0.002268	HDS60
chr2	198267442	G	A	SF3B1	silent	p.L639L	NA	0.000992	0.007003	HDS79
chr16	30750051	G	A	SRCAP	missense	p.G2897E	NA	0.001582	0.005455	HDS79
chrX	15833881	G	A	ZRSR2	silent	p.Q213Q	NA	0.001338	0.001003	HDS69
chr17	10354146	G	A	MYH4	missense	p.A1311V	NA	0.000727	0.000683	HDS69
chr17	7689552	C	T	DNAH2	silent	p.T2080T	NA	0.001261	0.002762	HDS60
chr2	25966728	A	G	ASXL2	silent	p.S826S	NA	0.000502	0.00057	HDS69
chr2	25965659	C	A	ASXL2	nonsense	p.E1183X	NA	0.002808	0.003299	HDS60
chrX	48649831	T	C	GATA1	intronic	NA	NA	0.000632	0.00082	HDS18
chrX	48649706	AG	A	GATA1	frame_shift_del	NA	NA	0.002	0.0025	HDS23

Table 5.3: Germline variants in ML-DS that were associated with hematologic malignancies.

Chr	Start	Ref	Alt	Gene	Type	AA_change	cosmic84	VAF1	VAF2	SampleID
chr13	28608283	G	A	FLT3	silent	p.Y591Y	COSM6494228	0.500339	0.490453	PAVZTK
chr1	36932047	C	T	CSF3R	missense	p.E835K	COSM5762855	0.481907	0.478406	PAXCGV
chr12	11905432	G	C	ETV6	missense	p.A28P	COSM5945248	0.492006	0.507824	PAXFXA
chr2	25470960	G	A	DNMT3A	silent	p.S267S	COSM6495452	0.503858	0.487982	PAXUDJ
chr17	7579579	C	T	TP53	silent	p.P36P	COSM6474190	0.528612	0.470927	PAXUTK
chr6	41908122	G	A	CCND3	missense	p.P134S	COSM6495114	0.999773	1	PAXVDP
chr21	36259308	C	T	RUNX1	silent	p.P61P	COSM6494711	0.325572	0.290361	PAXVDP
chr6	41903706	G	A	CCND3	missense	p.P284L	COSM220537	0.422444	0.432896	PAXVEZ
chr6	41903706	G	A	CCND3	missense	p.P284L	COSM220537	0.422444	0.432896	PAXVEZ
chr12	25378562	C	T	KRAS	missense	p.A146T	COSM1165198	0.377026	0.374139	PAXVEZ
chr4	187628248	C	T	FAT1	missense	p.V912I	COSM1717643	0.485451	0.48827	PAXWWR
chr12	25378561	G	A	KRAS	missense	p.A146V	COSM19900	0.416401	0.424791	PAXZJL
chr6	41903798	C	A	CCND3	missense	p.E253D	COSM5019335	0.463646	0.521831	PAYACG
chr11	32456784	C	G	WT1	silent	p.P36P	COSM6494045	0.492644	0.506854	PAYFXP
chr9	139399132	C	T	NOTCH1	missense	p.V1671I	COSM33750	0.564286	0.553021	PAYFXP
chr1	36932047	C	T	CSF3R	missense	p.E835K	COSM5762855	0.459792	0.470022	PAYLBLS
chr3	128200072	C	T	GATA2	silent	p.A411A	COSM5019736	0.527926	0.502423	PAYTYB
chr16	3779115	T	C	CREBBP	missense	p.N1978S	COSM96469	0.482854	0.506611	PAYWJJ
chr2	25470960	G	A	DNMT3A	silent	p.S267S	COSM6495452	0.496782	0.493492	PAYWJJ
chr17	29483108	C	T	NF1	silent	p.S56S	COSM6494460	0.505213	0.494312	PAYWKB
chr17	7578210	T	C	TP53	silent	p.R174R	COSM1741225	0.502452	0.484384	PAYWKB
chr17	29483108	C	T	NF1	silent	p.S56S	COSM6494460	0.486188	0.500268	PAYXGX
chr6	41903798	C	A	CCND3	missense	p.E253D	COSM5019335	0.495353	0.478406	PAYXGX

Table 5.4: Detected somatic mutations in HDS >4yo via ECS in the 80-gene panel.

Chr	Start	Ref	Alt	Gene	Type	AA_change	cosmic84	VAF1	VAF2	SampleID
chr7	138263948	G	A	TRIM24	splicing	NA	NA	0.000804	0.000768	HDS12
chr4	187541859	C	T	FAT1	missense	p.G1961S	COSM6476268	0.002297	0.001839	HDS13
chr4	187539902	G	A	FAT1	missense	p.T2613I	NA	0.001547	0.001794	HDS13
chr7	148523620	A	T	EZH2	nonsense	p.L278X	NA	0.000623	0.001059	HDS13
chr9	139399592	C	T	NOTCH1	intronic	NA	NA	0.001253	0.001624	HDS13
chr11	118343104	C	T	KMT2A	silent	p.I410I	NA	0.000552	0.000653	HDS13
chr7	98548005	T	A	TRRAP	intronic	NA	NA	0.002158	0.001269	HDS13
chr4	55575653	A	T	KIT	silent	p.L393L	NA	0.000838	0.001198	HDS13
chr18	42531941	C	A	SETBP1	missense	p.S879Y	NA	0.001055	0.001616	HDS13
chr1	36932392	T	A	CSF3R	missense	p.I693F	NA	0.001723	0.002951	HDS13
chr19	13414459	C	T	CACNA1A	intronic	NA	NA	0.000923	0.001184	HDS13
chr17	10353805	G	A	MYH4	silent	p.D1382D	COSM6719421	0.002608	0.00205	HDS14
chr2	198267794	A	G	SF3B1	intronic	NA	NA	0.000589	0.001035	HDS16
chr5	180056395	G	A	FLT4	silent	p.R283R	NA	0.000603	0.000984	HDS16
chr7	135333248	T	C	NUP205	missense	p.F1995L	NA	0.000568	0.000862	HDS16
chrX	123179279	T	A	STAG2	intronic	NA	NA	0.000972	0.001371	HDS16
chr12	112915556	G	C	PTPN11	intronic	NA	NA	0.001572	0.002603	HDS16
chr4	106155246	T	A	TET2	missense	p.N49K	NA	0.001049	0.001938	HDS16
chr7	103205782	G	A	RELN	missense	p.T1718I	NA	0.000717	0.000902	HDS16
chrX	76939099	C	T	ATRX	missense	p.S550N	NA	0.000994	0.001585	HDS16
chrX	76938808	A	T	ATRX	missense	p.L647H	NA	0.000516	0.001844	HDS16
chrX	76938295	T	C	ATRX	missense	p.D818G	NA	0.000645	0.000682	HDS16
chr3	47147554	G	C	SETD2	missense	p.A1591G	NA	0.000514	0.001179	HDS16
chr6	41908193	G	C	CCND3	missense	p.A110G	NA	0.000618	0.001078	HDS16
chr20	31023553	G	T	ASXL1	missense	p.S1013I	NA	0.000524	0.001676	HDS16
chr16	30750189	C	A	SRCAP	missense	p.P2943H	NA	0.000503	0.001026	HDS16
chr16	15835387	A	T	MYH11	missense	p.L938Q	NA	0.000536	0.000865	HDS16
chrX	15827423	C	T	ZRSR2	missense	p.A180V	NA	0.000745	0.001205	HDS16
chr12	11803214	G	A	ETV6	intronic	NA	NA	0.000701	0.001298	HDS16
chr17	11592972	C	T	DNAH9	missense	p.S1278F	COSM3514081	0.003148	0.010101	HDS16
chr17	7667524	C	T	DNAH2	missense	p.T1090I	NA	0.000506	0.001007	HDS16
chr4	187540946	G	A	FAT1	missense	p.A2265V	NA	0.000558	0.000591	HDS17
chr4	187540339	C	T	FAT1	silent	p.V2467V	NA	0.001098	0.001416	HDS17
chr4	153249452	T	C	FBXW7	silent	p.T442T	NA	0.000705	0.00079	HDS17
chr7	103214814	C	G	RELN	intronic	NA	NA	0.00074	0.001029	HDS17
chr4	106162534	C	T	TET2	missense	p.H1150Y	NA	0.000933	0.001044	HDS22
chrX	44921863	A	T	KDM6A	intronic	NA	NA	0.001962	0.001361	HDS22
chr13	28592669	T	C	FLT3	missense	p.K826E	NA	0.000685	0.000521	HDS27
chr2	198267820	G	A	SF3B1	intronic	NA	NA	0.001187	0.001046	HDS29

chr4	187541772	C	T	FAT1	missense	p.E1990K	NA	0.001152	0.000778	HDS29
chr20	31022598	C	T	ASXL1	nonsense	p.Q695X	COSM5879660	0.001055	0.001428	HDS29
chr4	24534689	G	T	DHX15	intronic	NA	NA	0.001455	0.001198	HDS29
chr17	11726351	G	T	DNAH9	missense	p.Q3082H	NA	0.000821	0.001135	HDS29
chr4	106158251	A	T	TET2	missense	p.Q1051L	NA	0.000767	0.00093	HDS3
chr6	131913622	G	C	MED23	intronic	NA	NA	0.000745	0.001215	HDS31
chrX	44937746	G	A	KDM6A	silent	p.K1030K	NA	0.001829	0.001155	HDS31
chr20	31021753	A	G	ASXL1	intronic	NA	NA	0.00098	0.001554	HDS31
chr16	3777724	A	T	CREBBP	missense	p.L2442M	NA	0.000676	0.000876	HDS32
chr4	187629186	A	T	FAT1	missense	p.L599H	NA	0.000541	0.000717	HDS34
chr7	148512620	G	A	EZH2	silent	p.C508C	NA	0.004579	0.003008	HDS39
chr17	11501883	G	A	DNAH9	missense	p.R23Q	NA	0.000551	0.001009	HDS39
chr19	13409870	C	T	CACNA1A	silent	p.R859R	COSM5580480	0.00218	0.002072	HDS4
chrX	39933012	G	A	BCOR	silent	p.N529N	NA	0.001355	0.001455	HDS41
chrX	76938688	T	A	ATRX	missense	p.E687V	NA	0.000837	0.001733	HDS42
chr4	187524744	G	A	FAT1	missense	p.R3646C	NA	0.003142	0.003236	HDS45
chr20	31023605	C	T	ASXL1	silent	p.D1030D	NA	0.000624	0.000659	HDS48
chr20	31024758	C	T	ASXL1	nonsense	p.R1415X	COSM166446	0.01263	0.008307	HDS51
chr2	61416069	C	T	USP34	missense	p.E3337K	NA	0.001655	0.00176	HDS52
chr3	47162771	T	A	SETD2	missense	p.S1119C	NA	0.001904	0.001056	HDS53
chrX	39934258	A	G	BCOR	missense	p.F114S	NA	0.00111	0.000996	HDS53
chr2	25497876	C	T	DNMT3A	silent	p.Q191Q	NA	0.001131	0.000988	HDS53
chr5	170819810	A	G	NPM1	missense	p.K150R	NA	0.000529	0.000782	HDS54
chr16	30727892	C	T	SRCAP	intronic	NA	NA	0.000517	0.00063	HDS54
chr7	148511337	C	G	EZH2	intronic	NA	NA	0.001174	0.000606	HDS56
chr8	117874128	G	T	RAD21	missense	p.T109N	NA	0.001054	0.0012	HDS56
chrX	76940035	G	A	ATRX	missense	p.A238V	NA	0.003925	0.00193	HDS56
chr4	55593704	T	A	KIT	missense	p.S586R	NA	0.000712	0.00076	HDS56
chr1	36933154	C	T	CSF3R	intronic	NA	NA	0.001361	0.000612	HDS56
chr17	10348009	T	C	MYH4	missense	p.E1860G	NA	0.000574	0.000938	HDS56
chr4	187542500	G	A	FAT1	missense	p.T1747I	NA	0.000749	0.00116	HDS57
chr9	139391187	A	G	NOTCH1	missense	p.L2335P	NA	0.001037	0.001052	HDS57
chr7	138266477	T	C	TRIM24	silent	p.H918H	NA	0.000931	0.000989	HDS57
chr7	135329722	T	C	NUP205	missense	p.V1880A	NA	0.000687	0.00066	HDS57
chrX	123196847	A	T	STAG2	intronic	NA	NA	0.003511	0.00411	HDS57
chrX	48649608	G	T	GATA1	missense	p.G31V	NA	0.001583	0.001465	HDS57
chrX	39933479	T	C	BCOR	missense	p.K374E	NA	0.001258	0.001826	HDS57
chr2	25458681	C	T	DNMT3A	missense	p.R831K	NA	0.000866	0.000846	HDS57
chr7	103180952	C	A	RELN	intronic	NA	NA	0.002039	0.001548	HDS6
chr2	25965982	C	T	ASXL2	missense	p.R1075Q	COSM6494820	0.008255	0.007256	HDS8
chrX	48649535	G	A	GATA1	missense	p.G7R	NA	0.000611	0.001222	HDS16
chrX	48649608	G	T	GATA1	missense	p.G31V	NA	0.001581	0.001461	HDS57

Chapter 6: NF1-glioblastoma multi-region clonal profiling

Note: This chapter is published at *Neurology* as of thesis submission (Wong *et al.* 2019)[179]. I conceptualized the manuscript together with Drs. David Gutmann and Todd Druley. I generated, processed, and analyzed the data. I wrote the published manuscript in its entirety with comments from co-authors, and generated the figures. Materials from the manuscript were re-formatted and re-used in writing this chapter.

6.1 Introduction

Neurofibromatosis Type 1 (NF1) is an inherited autosomal dominant cancer predisposition syndrome[180], and it is generally diagnosed in children by age 8 years[181]. Patients with NF1 are characterized genetically by a germline mutation in the *NF1* tumor suppressor gene which encodes a cytoplasmic protein called neurofibromin. This protein acts as a negative regulator of RAS oncogene[182,183]. The germline loss of this gene leads to an increase in oncogenic RAS activity, thereby causing unopposed cell proliferation via activation of downstream MAPK and mTOR signaling pathways[184,185]. Affected children typically present with multiple café-au-lait macules, intertriginous freckling, osseous lesions and Lisch nodules[182,183]. When these children reach adulthood, nearly all of them would develop benign neurofibromas on the peripheral nerve sheath[186]. Children and adults with NF1 are also more prone to developing malignant low-grade neoplasms such as optic and brainstem gliomas[187,188]. Among these NF1-associated neoplasms, high-grade glioblastoma (GBM) is rare, as it constitutes only approximately 2% of all malignant gliomas reported in NF1 patients[189,190].

While thus rare, the incidence of GBM is at least 5-folds higher in NF1 patients than in individuals without NF1 in the general population[191]. Notably, NF1-GBM tends to arise decades earlier relative to its sporadic counterparts[180]. In addition, in contrast to its non-syndromic GBM, NF1-GBM is presumed to arise in the setting of bi-allelic *NF1* inactivation (germline variant and a subsequent somatic mutation)[192,193].

Since GBM are rare in patients with NF1, the limited number of cases reported has precluded a comprehensive analysis of tumor ontogeny and evolution. Most cases were reported as a single case study detailing the clinical features, but these reports generally lacked disease ontogeny analysis based on genetics or genomics[194-196]. The largest cohort study with genomics analysis to date included 13 cases of NF1-GBM[192], and the authors performed whole-exome sequencing on a single neoplastic sample for each of these cases. However, single-region sequencing generally lacks the resolution needed to uncover intra-tumor genetic diversity on a spatial scale in solid cancers[197]. Herein, we leveraged a unique opportunity in which intra-tumoral samples from multiple brain lesions were obtained at autopsy from a 27-year-old young adult male with a clinical diagnosis of generalized NF1 (established at 3 years of age) who subsequently died from GBM. Specifically, we performed multi-region samplings within the same lesions in order to facilitate investigation into the spatial heterogeneity and genetic diversity. Contrary to the “2-hit hypothesis” of malignant gliomagenesis in the *NF1* gene[182,193], we found that the *NF1* somatic mutation was acquired subclonally at a later stage of disease progression. On the other hand, mutations in *KMT2B* were implicated in the founding clone.

6.2 Materials and Methods

6.2.1 Patient information

The patient was a 27-year-old male who had his NF1 diagnosed at the age of 3. When this patient initially presented with behavioral changes, neuroimaging revealed a bi-thalamic tumor (Figure 6.1), which was pathologically classified as a World Health Organization (WHO) grade IV astrocytoma (GBM) by stereotactic biopsy. This patient was then treated with concurrent temozolomide and cranial radiation, followed by temozolomide and Bevacizumab chemotherapy. Within two years, he exhibited progressive tumor growth, and developed a new enhancing lesion in the right parietal lobe and increased tumor infiltration in the cerebellum, and subsequently in the cerebral hemispheres and brainstem. The patient died 39 months after initial diagnosis.

6.2.2 Whole-exome sequencing analysis

Multi-region intra-tumoral samples were obtained wherever possible at autopsy, from different brain regions with neoplastic involvement, including thalamus (four independent samples), anterior commissure (two independent samples), septum pellucidum (two independent samples), amygdala (two independent samples), and cerebellum (one sample). One representative non-malignant brain sample was also harvested to serve as a non-neoplastic tissue control to distinguish somatic mutations from germline variants. Following DNA extraction, whole-exome sequencing libraries were generated using Agilent SureSelect Clinical Research Exome V2 chemistry and sequenced on an Illumina HiSeq 3000 platform.

Both somatic and germline variants were independently called using two programs, VarScan2[120] and Strelka[198], using default parameters. Genomic positions with less than 20x coverage were filtered. High-confidence variants independently called by both callers were retained for further analysis. For variants identified by both methods only in subsets of sequenced samples, further manual curation was performed to check for the presence of these variants in all other samples using bam files and filtered calls (Figure 6.2). This secondary screen was performed in order to ascertain whether the absence of variants in some samples was not due to technical issues in variant calling. In many cases, variants were excluded by one program, but not by the other, due to different statistical frameworks applied, as well as the relatively low coverage of the genomic position. These variants were further checked against the data from the normal tissue sample for absence of alternate supporting reads. Variants curated in this fashion were added to the final dataset for downstream analysis. Copy number analysis was performed with VarScan2 and titanCNA[199] using recommended parameters. Tumor purity was estimated with titanCNA, and the variant allele frequency (VAF) of somatic mutations was adjusted for tumor purity. Evolutionary history of tumor subpopulations was reconstructed using PhyloWGS[200] with the variant outputs in Strelka and titanCNA VCF format. A total of 1000 trees per iteration were generated in four separate iterations, and the best tree identified by the program was selected for downstream analysis. Evolutionary trajectories of tumor subpopulations with cellular contribution ≥0.05 were examined[201]. The reconstructed trees were then heuristically processed, such that the intermediate parent/daughter nodes containing the same exact tumor sites with cellular contribution ≥ 0.05 were merged. Putative pathogenicity of variants was assessed using CADD score[138]. Mutations with CADD score >20 are considered to be top 1% most deleterious across the genome.

6.3 Results

Whole-exome sequencing (mean depth 131.7x after data processing) was performed, revealing pathogenic germline variants in the *IDH1* (c.G532A; p.V178I; COSM97131; CADD score = 25.1) and *NF1* (c.C7285T; p.R2429*) genes, the latter confirming a diagnosis of NF1. Additionally, non-synonymous somatic mutations, not seen in non-neoplastic tissue, were identified in six pan-cancer driver genes[202] (Figure 6.3; Table 6.1). Of these, only the nonsense and missense *KMT2B* mutations (NM_014727; p.E1799*, CADD score = 47; and p.Q1683H, CADD score = 22.2) were shared across all tumor samples at estimated mean variant allele frequencies (VAF) of 0.31 and 0.25, respectively, after adjusting for tumor purity (Figure 6.3). In addition, the *KMT2B* locus was also amplified in all neoplastic samples (Figure 6.4). Importantly, the somatic *NF1* mutation (p.Y2141H) was only found in samples from the anterior commissure, septum pellucidum, amygdala and one specific thalamic site (T4) (Figure 6.3).

To interrogate the temporal sequence of somatic mutations acquisition, the evolutionary history of tumor subpopulations was reconstructed using PhyloWGS[200]. The *KMT2B* mutations were inferred to be the founding somatic events in this individual (Figure 6.5). Consistent with the patient's initial clinical presentation, the thalamus was inferred to be the site of origin. We identified intra-tumoral heterogeneity amongst the four thalamic samples, with only thalamus site 4 (T4) being mainly responsible for the spread of tumor cells to the amygdala, anterior commissure and septum pellucidum. The tumor at T4 secondarily acquired additional missense somatic mutations in other pan-cancer driver genes, including *NF1* and *PIK3R1* (Figure 6.5), demonstrating that *NF1* somatic mutation occur late in the progression of this individual's GBM.

6.4 Discussion

Our data diverged from the commonly held "2-hit hypothesis" in NF1 gliomagenesis[182,193]. Since NF1 patients already have a germline loss of function variant in the *NF1* gene, the second hit in the "2-hit hypothesis" refers to an acquired somatic mutation in *NF1* on the other allele that leads to a *NF1*-null clone. This loss of heterozygosity leads to increased RAS oncogenic activity, which in turn drives malignant transformation. However, we demonstrated that somatic *NF1* mutation was only present in a subset of neoplastic samples at relatively low VAFs, suggesting that *NF1* bi-allelic loss was not the primary molecular alteration, but rather a late subclonal event. This observation was further substantiated by the tumor phylogenetic analysis that showed the thalamic tumor at site T4 had secondarily acquired the somatic *NF1* mutation, demonstrating that the somatic *NF1* mutation was acquired subclonally during disease progression of this patient's GBM, and was likely responsible for tumor spreading rather than initial transformation. This molecular ontology analysis provides a proof-of-concept demonstration that some gliomagenesis-associated events (*i.e.*, *KMT2B* mutation/amplification) occur prior to *NF1* bi-allelic inactivation, and may be sufficient to drive gliomagenesis in a *NF1* heterozygous background. Consistent with this conclusion, mouse modeling experiments have suggested that *Tp53* loss precedes *Nf1* loss, such that *Nf1* loss before *Tp53* inactivation does not result in malignant glioma formation[203].

KMT2B (Lysine Methyltransferase 2B) is a chromatin remodeling protein that catalyzes H3K4 mono-methylation primarily at promoters. While recurrently mutated in several different tumor types, including grade II and III gliomas[204], only eleven *KMT2B* mutations have been reported in TCGA GBM samples (~2% of cases), scattered across the coding region[205] (Figure 6.6), and none were detected in the largest series of GBM from adults with NF1. Recently, there have been discussions about the implications of chromatin remodeling defects in tumorigenesis of glioblastoma[176,206,207]. It has been suggested that even a small subpopulation of cells (<1%) with inactivating defect in another histone methyltransferase *KMT5B* could exert profound effects on tumorigenesis of pediatric glioblastoma by modulating non-cell autonomous chemokine releases via epigenetic regulation of certain gene expressions[176]. It therefore stands to reason that a loss of *KMT2B* as presented in this case may have similar effects by altering the chromatin landscape and changing gene expression patterns. It is also important to note that this patient had an *IDH1* germline variant that is associated with malignancies. Although the precise role of *IDH1* variants in tumorigenesis remains uncertain[208], it has been previously shown that some *IDH1* mutations promote the production of the oncogenic metabolite R-2-hydroxyglutarate, which is involved in chromatin remodeling[209]. Taken together, this case alludes to the importance of chromatin dysregulation in gliomagenesis, and that a loss of chromatin remodeling protein (i.e. *KMT2B*) could drive initial transformation and promote tumor instability. Future mechanistic studies will be required to determine whether *KMT2B* alterations drive gliomagenesis by themselves or in the setting of heterozygous *NF1* genetic background.

We acknowledge that our study was limited to a single anecdotal case, although the results were compelling and relevant to those working on NF1-associated tumors. In all likelihood, this patient's tumor ontogeny pathway may be a rare exception rather than a general rule. However, we posit that the fraction of NF1-associated gliomas driven by non-*NF1* somatic alterations is, at the very least, not low. It is generally difficult to obtain post-mortem brain autopsy samples especially from rare diseases[210], and there has not been any comprehensive and systematic study examining spatial heterogeneity of these GBM brain lesions due to a lack of multi-region intra-tumoral samples. When these obstacles are overcome, we believe we would begin to observe alternative route to NF1 gliomagenesis at a relatively high frequency.

In summary, we demonstrated that somatic *KMT2B* alterations were likely the primary oncogenic events in a young man with NF1-GBM, and that *NF1* bi-allelic loss was not required for initial malignant transformation. This case also highlights the importance of establishing tumor ontogeny by multi-region clonal profiling, because the temporal sequence and spatial divergence of somatic events could influence gliomagenesis and progression respectively.

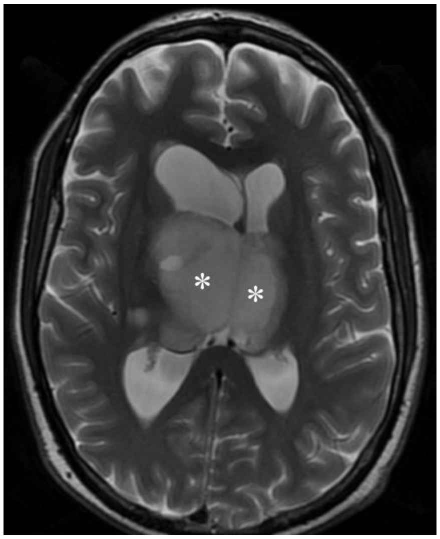


Figure 6.1: MRI neuroimaging of the 27 years old patient revealed bi-thalamic tumors which were subsequently classified as glioblastoma upon biopsy.

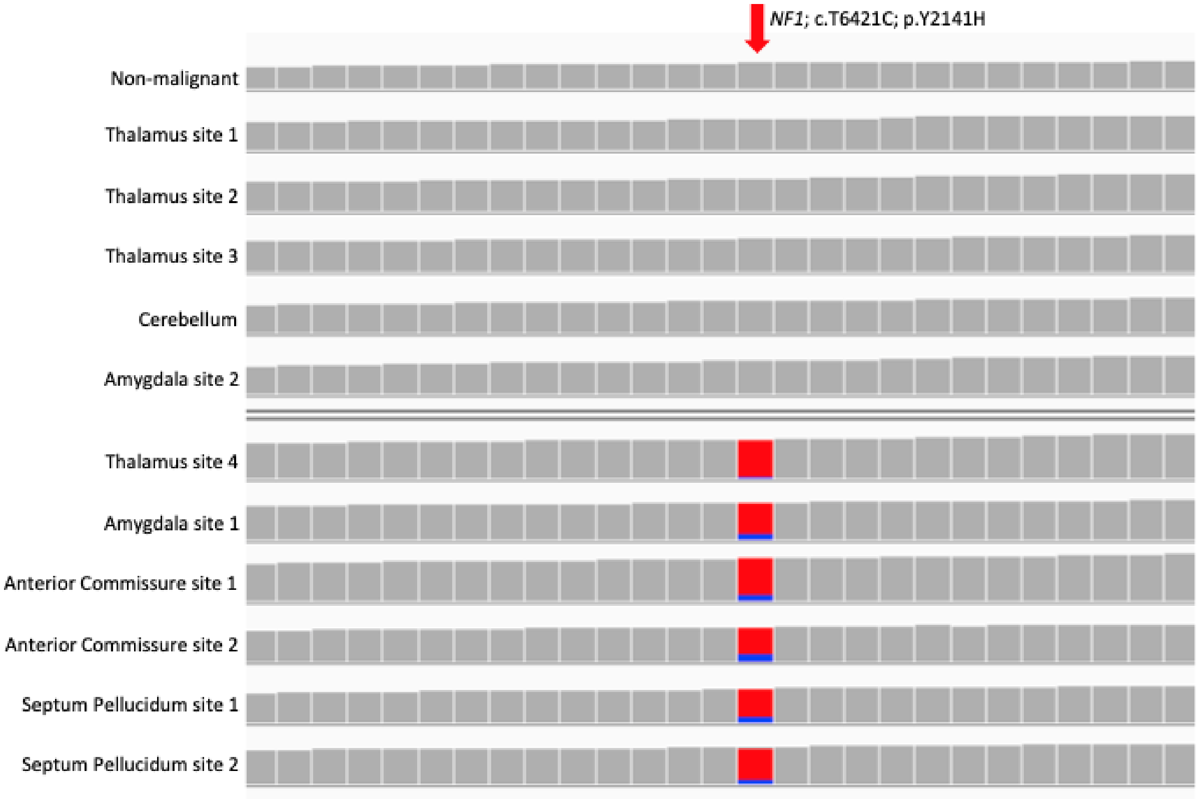


Figure 6.2: The presence and absence of *NF1* somatic mutations in the sequenced samples. The aligned bam file for each sample was manually checked, and viewed in the IGV browser to confirm the status of the mutation in all samples. Gray-colored bars indicate wild type (without mutation) alleles. Red- and blue-colored bars indicate the presence of mutant allele, with red representing wild type thymine and blue representing mutant cytosine.

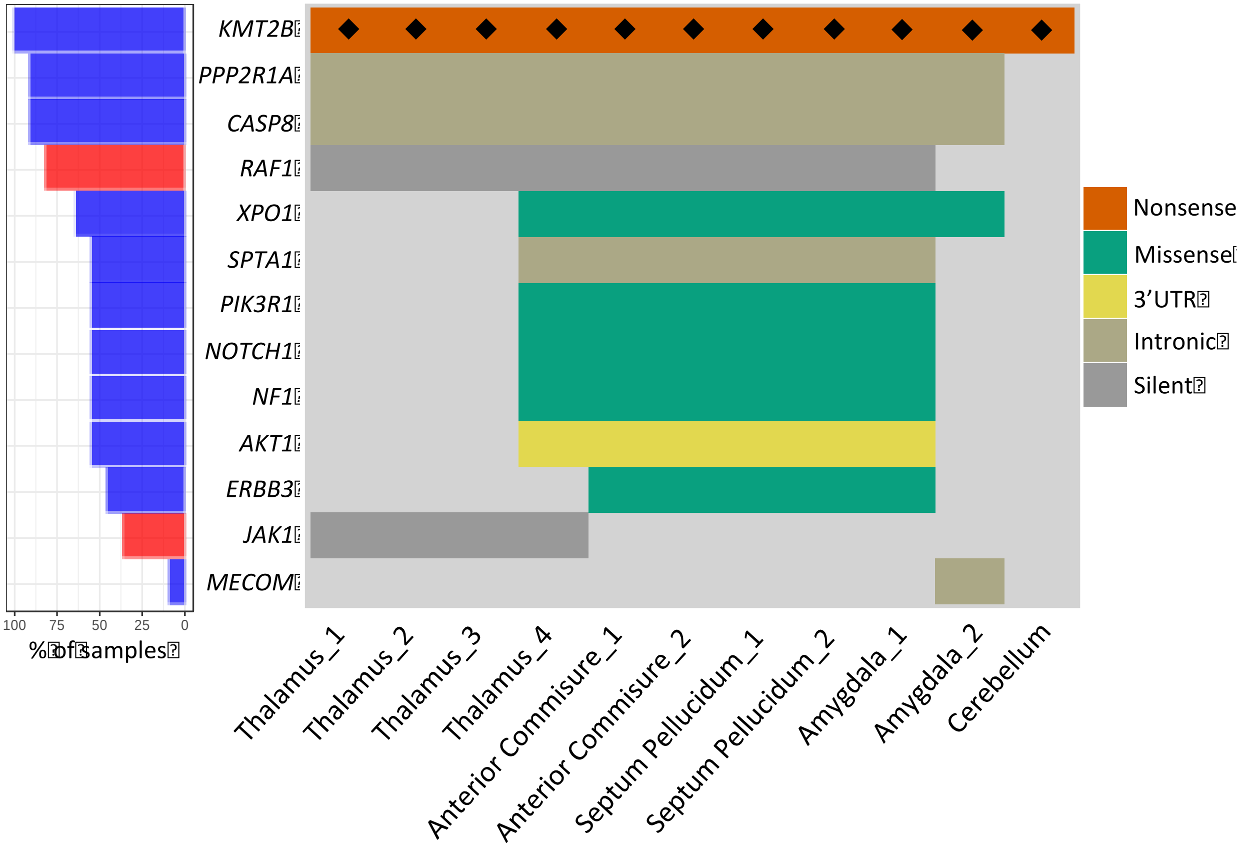


Figure 6.3: Somatic mutations identified in pan-cancer driver genes. The *KMT2B* nonsense somatic mutation was present in all tumor samples, but not in the non-neoplastic brain control. All sequenced tumor samples also harbored a missense (p.Q1683H) and a silent (p.L1645L) *KMT2B* mutation (◆). Somatic non-synonymous mutations in other known GBM driver genes (i.e., *NF1* and *PIK3R1* mutations), were present only in some of the tumor samples.

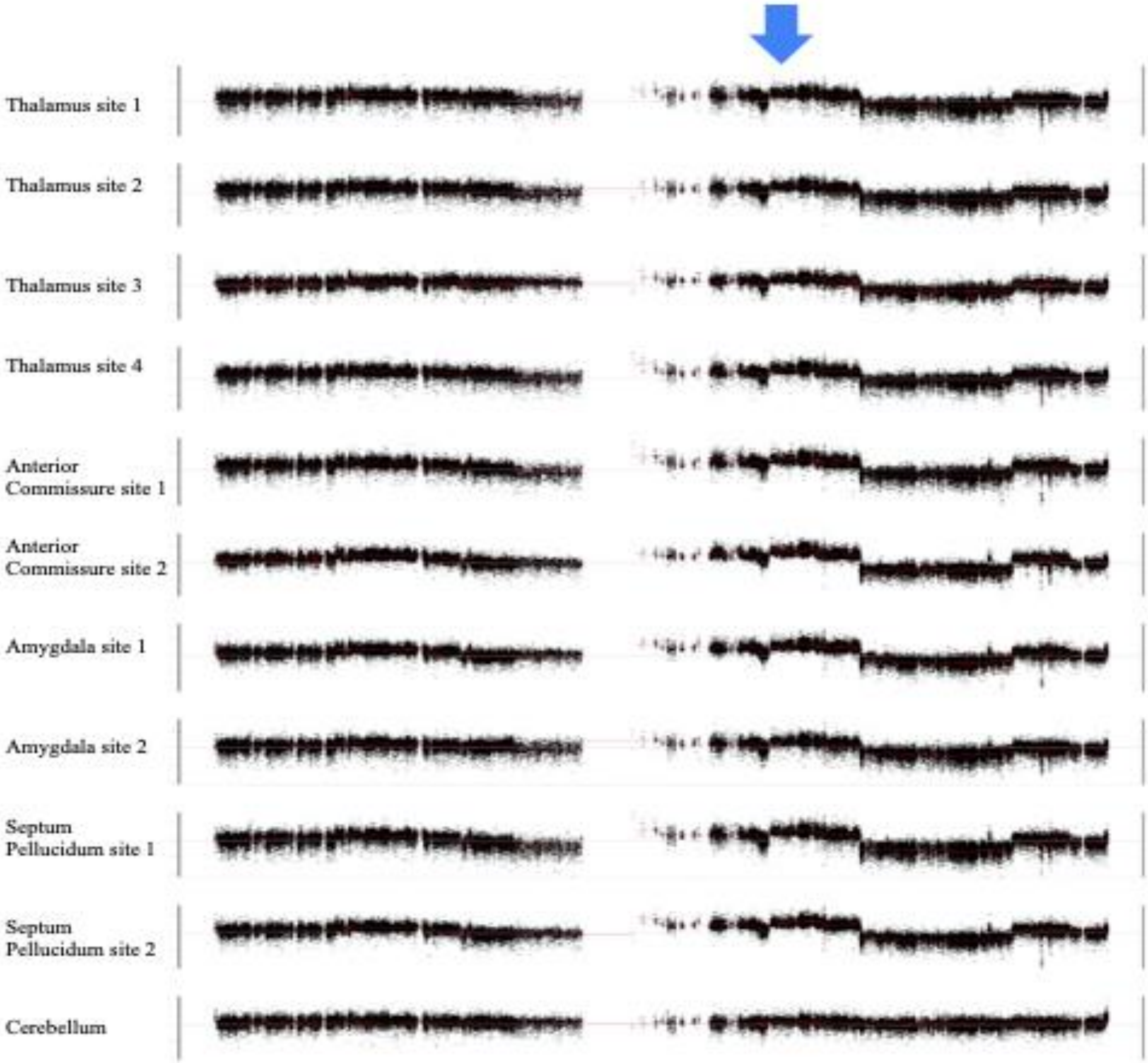


Figure 6.4: Copy number status of *KMT2B* gene locus. The gene locus has been amplified in all malignant samples (blue arrow). The y-axis represents copy number log2-ratio of tumor over normal samples. The x-axis represents genomic positions along chromosome 19. Note that the copy number gain status is not obvious in cerebellum sample given the resizing of the images, but the gain status is indicated by the copy number segment median above a log2-ratio of zero (red color lines in the plot). Cerebellum sample also has a normal cell contamination rate of 77% (as measured by titanCNA), which would have diluted the copy number gain signal.

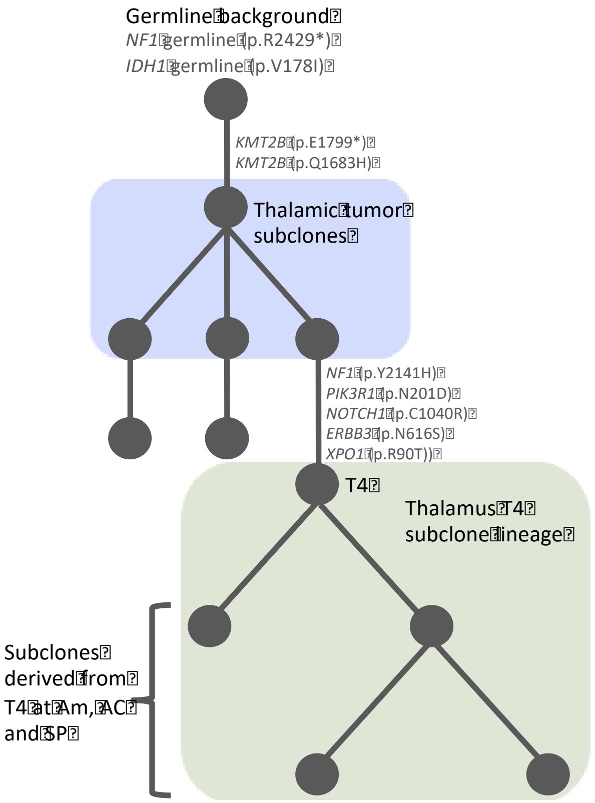


Figure 6.5: Phylogenetic tree depicting the evolutionary trajectory of different tumor subpopulations. Each node denotes a tumor subpopulation. Thalamic tumors contained only subpopulations that form the higher nodes, indicating an ancestral relationship with other subpopulations. The nonsense and missense *KMT2B* mutations were inferred to be the founding events, while somatic missense mutations in other pan-cancer driver genes were subclonally acquired in thalamic site 4 (T4) lineage. Tumor subpopulations at amygdala (Am), anterior commissure (AC) and septum pellucidum (SP) were similar in terms of mutation profiles, suggesting spread from T4. Only amino acid-changing mutations are shown.

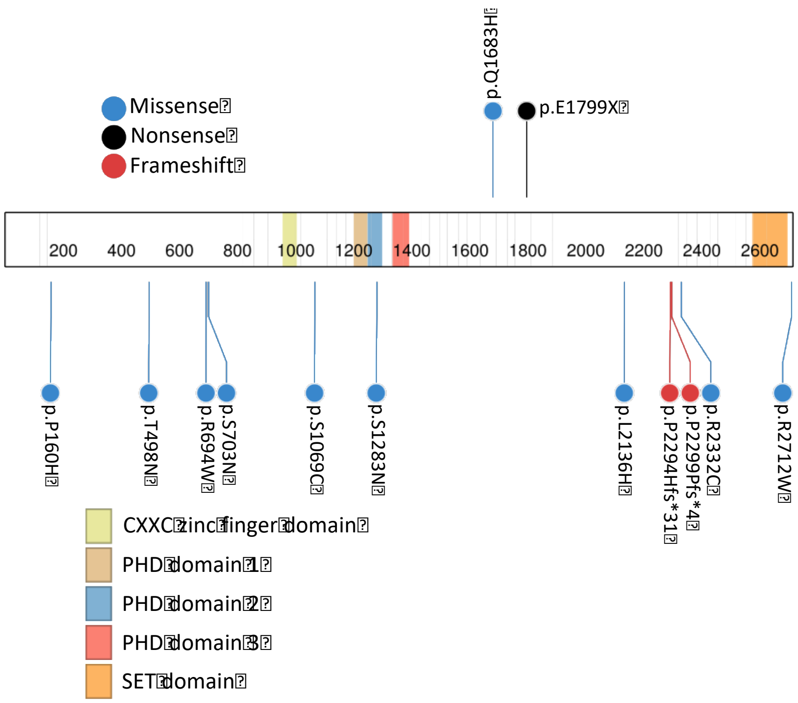


Figure 6.6: Non-synonymous *KMT2B* mutations present in this patient (top) and in non-*NF1*-mutant TCGA GBM (bottom) are depicted. This individual harbored a premature stop codon before the SET domain of *KMT2B*, presumably inactivating its histone methyltransferase activity, and resulting in epigenetic dysregulation.

Table 6.1: Pan-cancer driver somatic mutations observed in this patient at different tumor sites.

Chr	Start	End	Ref	Mut	Gene	Type	AA_change	VAF	ID
chr14	105236497	105236497	C	G	*AKT1*	UTR3	NA	0.379029079	Amygdala_1
chr14	105236497	105236497	C	G	*AKT1*	UTR3	NA	0.425139086	Septum Pellucidum_2
chr14	105236497	105236497	C	G	*AKT1*	UTR3	NA	0.125372536	Thalamus_4
chr14	105236497	105236497	C	G	*AKT1*	UTR3	NA	0.466999918	Anterior Commissure_2
chr14	105236497	105236497	C	G	*AKT1*	UTR3	NA	0.395225112	Septum Pellucidum_1
chr14	105236497	105236497	C	G	*AKT1*	UTR3	NA	0.347559265	Anterior Commissure_1
chr2	202134155	202134155	T	G	*CASP8*	intronic	NA	0.379029079	Amygdala_1
chr2	202134155	202134155	T	G	*CASP8*	intronic	NA	0.383123717	Thalamus_1
chr2	202134155	202134155	T	G	*CASP8*	intronic	NA	0.19221086	Amygdala_2
chr2	202134155	202134155	T	G	*CASP8*	intronic	NA	0.259084761	Septum Pellucidum_2
chr2	202134155	202134155	T	G	*CASP8*	intronic	NA	0.297439944	Thalamus_4
chr2	202134155	202134155	T	G	*CASP8*	intronic	NA	0.09271428	Thalamus_3
chr2	202134155	202134155	T	G	*CASP8*	intronic	NA	0.337543196	Anterior Commissure_2
chr2	202134155	202134155	T	G	*CASP8*	intronic	NA	0.238882378	Septum Pellucidum_1
chr2	202134155	202134155	T	G	*CASP8*	intronic	NA	0.264211938	Thalamus_2
chr2	202134155	202134155	T	G	*CASP8*	intronic	NA	0.233496823	Anterior Commissure_1
chr12	56488328	56488328	A	G	*ERBB3*	missense	p.N616S	0.262999769	Amygdala_1
chr12	56488328	56488328	A	G	*ERBB3*	missense	p.N616S	0.164011003	Septum Pellucidum_2
chr12	56488328	56488328	A	G	*ERBB3*	missense	p.N616S	0.248443956	Anterior Commissure_2
chr12	56488328	56488328	A	G	*ERBB3*	missense	p.N616S	0.169931538	Septum Pellucidum_1
chr12	56488328	56488328	A	G	*ERBB3*	missense	p.N616S	0.187212214	Anterior Commissure_1
chr1	65321304	65321304	C	T	*JAK1*	silent	NA	0.140956747	Thalamus_1
chr1	65321304	65321304	C	T	*JAK1*	silent	NA	0.039800805	Thalamus_4
chr1	65321304	65321304	C	T	*JAK1*	silent	NA	0.096436386	Thalamus_3
chr1	65321304	65321304	C	T	*JAK1*	silent	NA	0.324822355	Thalamus_2
chr19	36220885	36220885	G	C	*KMT2B*	silent	NA	0.285290705	Amygdala_1
chr19	36220885	36220885	G	C	*KMT2B*	silent	NA	0.351452156	Thalamus_1
chr19	36220885	36220885	G	C	*KMT2B*	silent	NA	0.281962212	Amygdala_2
chr19	36220885	36220885	G	C	*KMT2B*	silent	NA	0.210816392	Septum Pellucidum_2
chr19	36220885	36220885	G	C	*KMT2B*	silent	NA	0.192690794	Thalamus_4
chr19	36220885	36220885	G	C	*KMT2B*	silent	NA	0.656354699	Thalamus_3
chr19	36220885	36220885	G	C	*KMT2B*	silent	NA	0.237745413	Anterior Commissure_2
chr19	36220885	36220885	G	C	*KMT2B*	silent	NA	0.709044692	Cerebellum
chr19	36220885	36220885	G	C	*KMT2B*	silent	NA	0.22609042	Septum Pellucidum_1
chr19	36220885	36220885	G	C	*KMT2B*	silent	NA	0.271607183	Thalamus_2
chr19	36220885	36220885	G	C	*KMT2B*	silent	NA	0.179337617	Anterior Commissure_1
chr19	36220999	36220999	G	C	*KMT2B*	missense	p.Q1683H	0.29156083	Amygdala_1
chr19	36220999	36220999	G	C	*KMT2B*	missense	p.Q1683H	0.422049533	Thalamus_1
chr19	36220999	36220999	G	C	*KMT2B*	missense	p.Q1683H	0.26817192	Amygdala_2

chr19	36220999	36220999	G	C	*KMT2B*	missense	p.Q1683H	0.15929262	Septum Pellucidum_2
chr19	36220999	36220999	G	C	*KMT2B*	missense	p.Q1683H	0.301959826	Thalamus_4
chr19	36220999	36220999	G	C	*KMT2B*	missense	p.Q1683H	0.512626575	Thalamus_3
chr19	36220999	36220999	G	C	*KMT2B*	missense	p.Q1683H	0.317768659	Anterior Commissure_2
chr19	36220999	36220999	G	C	*KMT2B*	missense	p.Q1683H	0.718257191	Cerebellum
chr19	36220999	36220999	G	C	*KMT2B*	missense	p.Q1683H	0.189708054	Septum Pellucidum_1
chr19	36220999	36220999	G	C	*KMT2B*	missense	p.Q1683H	0.313840457	Thalamus_2
chr19	36220999	36220999	G	C	*KMT2B*	missense	p.Q1683H	0.163477827	Anterior Commissure_1
chr19	36221726	36221726	G	T	*KMT2B*	stopgain	p.E1799X	0.253814115	Amygdala_1
chr19	36221726	36221726	G	T	*KMT2B*	stopgain	p.E1799X	0.450118028	Thalamus_1
chr19	36221726	36221726	G	T	*KMT2B*	stopgain	p.E1799X	0.287072207	Amygdala_2
chr19	36221726	36221726	G	T	*KMT2B*	stopgain	p.E1799X	0.262169103	Septum Pellucidum_2
chr19	36221726	36221726	G	T	*KMT2B*	stopgain	p.E1799X	0.097692885	Thalamus_4
chr19	36221726	36221726	G	T	*KMT2B*	stopgain	p.E1799X	0.363961017	Thalamus_3
chr19	36221726	36221726	G	T	*KMT2B*	stopgain	p.E1799X	0.277720305	Anterior Commissure_2
chr19	36221726	36221726	G	T	*KMT2B*	stopgain	p.E1799X	0.585946655	Cerebellum
chr19	36221726	36221726	G	T	*KMT2B*	stopgain	p.E1799X	0.269683017	Septum Pellucidum_1
chr19	36221726	36221726	G	T	*KMT2B*	stopgain	p.E1799X	0.400681768	Thalamus_2
chr19	36221726	36221726	G	T	*KMT2B*	stopgain	p.E1799X	0.182557767	Anterior Commissure_1
chr3	168813068	168813068	A	T	*MECOM*	intronic	NA	0.279937868	Amygdala_2
chr17	29664442	29664442	T	C	*NF1*	missense	p.Y2141H	0.311977252	Amygdala_1
chr17	29664442	29664442	T	C	*NF1*	missense	p.Y2141H	0.234778301	Septum Pellucidum_2
chr17	29664442	29664442	T	C	*NF1*	missense	p.Y2141H	0.125833927	Thalamus_4
chr17	29664442	29664442	T	C	*NF1*	missense	p.Y2141H	0.498133246	Anterior Commissure_2
chr17	29664442	29664442	T	C	*NF1*	missense	p.Y2141H	0.263051468	Septum Pellucidum_1
chr17	29664442	29664442	T	C	*NF1*	missense	p.Y2141H	0.220499415	Anterior Commissure_1
chr9	139403375	139403375	A	G	*NOTCH1*	missense	p.C1040R	0.294074286	Amygdala_1
chr9	139403375	139403375	A	G	*NOTCH1*	missense	p.C1040R	0.222072652	Septum Pellucidum_2
chr9	139403375	139403375	A	G	*NOTCH1*	missense	p.C1040R	0.118462238	Thalamus_4
chr9	139403375	139403375	A	G	*NOTCH1*	missense	p.C1040R	0.280199951	Anterior Commissure_2
chr9	139403375	139403375	A	G	*NOTCH1*	missense	p.C1040R	0.154137793	Septum Pellucidum_1
chr9	139403375	139403375	A	G	*NOTCH1*	missense	p.C1040R	0.186288687	Anterior Commissure_1
chr5	67591097	67591097	A	G	*PIK3R1*	missense	p.N201D	0.300524674	Amygdala_1
chr5	67591097	67591097	A	G	*PIK3R1*	missense	p.N201D	0.262169103	Septum Pellucidum_2
chr5	67591097	67591097	A	G	*PIK3R1*	missense	p.N201D	0.107462173	Thalamus_4
chr5	67591097	67591097	A	G	*PIK3R1*	missense	p.N201D	0.392279931	Anterior Commissure_2
chr5	67591097	67591097	A	G	*PIK3R1*	missense	p.N201D	0.267435659	Septum Pellucidum_1
chr5	67591097	67591097	A	G	*PIK3R1*	missense	p.N201D	0.241158577	Anterior Commissure_1
chr19	52705314	52705314	G	C	*PPP2R1A*	intronic	NA	0.34457189	Amygdala_1
chr19	52705314	52705314	G	C	*PPP2R1A*	intronic	NA	0.15975098	Thalamus_1
chr19	52705314	52705314	G	C	*PPP2R1A*	intronic	NA	0.685958863	Amygdala_2
chr19	52705314	52705314	G	C	*PPP2R1A*	intronic	NA	0.151251406	Septum Pellucidum_2
chr19	52705314	52705314	G	C	*PPP2R1A*	intronic	NA	0.676166484	Thalamus_4

chr19	52705314	52705314	G	C	*PPP2R1A*	intronic	NA	0.283109677	Thalamus_3
chr19	52705314	52705314	G	C	*PPP2R1A*	intronic	NA	0.512981448	Anterior Commissure_2
chr19	52705314	52705314	G	C	*PPP2R1A*	intronic	NA	0.267435659	Septum Pellucidum_1
chr19	52705314	52705314	G	C	*PPP2R1A*	intronic	NA	0.452295013	Thalamus_2
chr19	52705314	52705314	G	C	*PPP2R1A*	intronic	NA	0.214564648	Anterior Commissure_1
chr3	12641768	12641768	T	A	*RAF1*	silent	NA	0.351955573	Amygdala_1
chr3	12641768	12641768	T	A	*RAF1*	silent	NA	0.366095996	Thalamus_1
chr3	12641768	12641768	T	A	*RAF1*	silent	NA	0.286002658	Septum Pellucidum_2
chr3	12641768	12641768	T	A	*RAF1*	silent	NA	0.222335531	Thalamus_4
chr3	12641768	12641768	T	A	*RAF1*	silent	NA	0.109008126	Thalamus_3
chr3	12641768	12641768	T	A	*RAF1*	silent	NA	0.503602614	Anterior Commissure_2
chr3	12641768	12641768	T	A	*RAF1*	silent	NA	0.360756299	Septum Pellucidum_1
chr3	12641768	12641768	T	A	*RAF1*	silent	NA	0.342120587	Thalamus_2
chr3	12641768	12641768	T	A	*RAF1*	silent	NA	0.340442575	Anterior Commissure_1
chr1	158589932	158589932	A	C	*SPTA1*	intronic	NA	0.47757664	Amygdala_1
chr1	158589932	158589932	A	C	*SPTA1*	intronic	NA	0.379208882	Septum Pellucidum_2
chr1	158589932	158589932	A	C	*SPTA1*	intronic	NA	0.119957775	Thalamus_4
chr1	158589932	158589932	A	C	*SPTA1*	intronic	NA	0.658371912	Anterior Commissure_2
chr1	158589932	158589932	A	C	*SPTA1*	intronic	NA	0.597992777	Septum Pellucidum_1
chr1	158589932	158589932	A	C	*SPTA1*	intronic	NA	0.368900623	Anterior Commissure_1
chr2	61749778	61749778	C	G	*XPO1*	missense	p.R90T	0.34457189	Amygdala_1
chr2	61749778	61749778	C	G	*XPO1*	missense	p.R90T	0.133793834	Amygdala_2
chr2	61749778	61749778	C	G	*XPO1*	missense	p.R90T	0.188761754	Septum Pellucidum_2
chr2	61749778	61749778	C	G	*XPO1*	missense	p.R90T	0.207991303	Thalamus_4
chr2	61749778	61749778	C	G	*XPO1*	missense	p.R90T	0.392279931	Anterior Commissure_2
chr2	61749778	61749778	C	G	*XPO1*	missense	p.R90T	0.360219459	Septum Pellucidum_1
chr2	61749778	61749778	C	G	*XPO1*	missense	p.R90T	0.261209137	Anterior Commissure_1

Chapter 7: Summary and conclusion

7.1 Future explorations in clonality of disease

The bulk of this thesis concerns clonal hematopoiesis in malignant and non-malignant settings. We demonstrated that clonal profiling; especially via longitudinal sampling is a powerful and informative way to quantitatively characterize the evolutionary trajectories of relevant somatic mutations implicated in various hematopoietic disorders and complex aging phenotypes. By correlating clonal profiles with clinical outcomes, we would gain further insights into disease ontogenies and associated pathogenic mechanisms. However, beyond what we have presented in this thesis, the causes and effects of clonal hematopoiesis in many other scenarios remain unclear.

An intriguing example for future investigation into hematopoietic clonality is sickle cell disease. Over the past few decades, improvements in treatment have led to an increased life expectancy among sickle cell anemia patients, such that many patients could now survive to an age where they are at risk for malignancies[211]. Recently, it has been demonstrated that AYA and, in general, female sickle cell patients have a 3-fold higher risk of developing myeloid leukemia relative to age-matched, non-sickle cell populations[212]. The clonality in these cases as well as the functional mechanism implicating a female-biased risk profile is not currently known. It is very likely that these sickle cell patients have vastly different clonal profiles such that certain clonal mutations could serve as biomarkers for early detection and risk stratification. The female-biased leukemia risk profile in sickle cell patients could also be associated with a X-linked genetic factor. It is therefore fascinating to think that there could be multiple seemingly different mechanisms converging into common pathways that drive leukemogenesis in patients with different hematologic pre-conditions (i.e. dysregulated hematopoietic differentiation in DS versus sickle cell anemia).

In addition, our candidate gene data suggests that clonal hematopoiesis is present in approximately 30% of newborns, indicating that mutagenesis also occurs *in utero*. Given the limited number of genes studied in these experiments, the absolute number of newborns with CH is likely much higher. It is unknown if the DNA mutations were caused by mothers' exposure to environmental mutagens or are simply a by-product of stochastic errors during rapid developmental growth. More importantly, we do not know if newborns with clonal hematopoiesis are associated with a higher life-long risk of developing hematologic disorders relative to those born without clonal hematopoiesis. Collecting enough longitudinal samples to build a case control study would prove to be difficult because many hematologic diseases implicated by clonal hematopoiesis develop in individuals in their 50s – 70s, but it is essential as it would allow us to uncover the discrete steps that transform benign hematopoietic clones into pathogenic entities. One interesting observation made in our studies was that somatic mutations in *DNMT3A* and *TET2* seem to first arise in blood cells after 20 years of age. It is unclear why this is the case. One possibility is that the DNA changes in these epigenetic modifiers are the consequences stemming from exposure to specific external and environmental mutagens that are relevant to individuals during age 20s – 40s[213]. The mutated epigenetic modifiers would then change the epigenetic landscape, thus allowing an adaptive long-term phenotype to be established as a form of cellular plasticity without the cells having to incur too many permanent somatic alterations across the entire genome[214]. This would also explain why there seems to be a positive selection for mutations in the epigenetic modifiers in older individuals. Identifying what these mutagens are would enable the population to reduce the exposure and the risk of further DNA alterations, and ultimately, the risk of leukemia development.

So far, although our approach enables us to characterize clonality with high sensitivity, it does not allow us to examine if multiple mutations present within the same individuals are co-localized within the same cell. Given that individuals without or without leukemia have similar mutation burden[27], and that some individuals with *DNMT3A* mutation at leukemogenic R882 amino acid residue do not develop leukemia[30], it is reasonable to conclude that multiple pathogenic mutations have to co-localize within the same cell to drive leukemogenesis. In order to do this, single-cell sequencing approaches have to be used. Despite, recent advances in single-cell sequencing technologies, RNA expression profiles are generally limited to sequencing of a small amount of the 5' or 3' end of the mRNA molecule, making allele-specific expression quantification untenable. For single cell DNA sequencing, the current number of target genes that can be sequenced simultaneously are quite limited, and single cell profiling of epigenetic changes are still nascent technologies[215,216]. Thus, using a single cell approach to establish cancer ontology is not yet mature enough to address the necessary questions.

The hematopoietic system is ideal for studying clonal evolution since blood samples are routinely acquired from even healthy individuals during medical checkups. Conversely, it is more difficult to study clonality in other solid organs, and hence our understanding of clonality in normal solid tissues remains obscure. Following the concept of clonal hematopoiesis and with the advances in sequencing technologies, it has been shown that normal breast tissues in healthy individuals serially acquire somatic mutations, but the effects of these seemingly benign mutations in malignant transformation remained unclear[217]. In another recent study, normal uterine endometrial epithelium samples were shown to share somatic mutations in *KRAS* and

ARID1A with endometriotic epithelium that is widely considered to be the precursor of endometrial carcinoma[218]. It is postulated that the cell of origin or the site of origin with these mutations determines the evolutionary trajectory of malignant transformation. Therefore, the characterization of clonality in solid organ systems is an important undertaking, and there are much more remain to be discovered.

7.2 Mathematical modeling of clonal trajectories

The democratization of sequencing has resulted in an explosion of sequencing data available to the general public and research community. Ironically, this did not immediately lead to cures for many human diseases, but instead it exposed the complexity of these diseases. Sequencing generally provides a static ‘snapshot’ perspective[125], and it is difficult to infer the evolutionary history of certain disease relevant genetic alterations without examining longitudinally collected samples, which is a challenge in itself. In light of this, mathematical modeling of disease pathogenicity has increasingly become relevant, and it enables us to examine complex biological and evolutionary processes when used in tandem with patient outcome data[35,219]. Recently, a model incorporating VAFs of somatic mutations in blood cells as a way to measure clonal fitness found that clonal hematopoiesis is shaped by positive selection, not drift[125]. This model will be a powerful starting point to investigate evolutionary trajectory of mutations found normal hematopoiesis over a longitudinal time window, and to possibly predict the risk future risk of leukemia with improved accuracy decades before diagnosis. Models that examined metastatic patterns in solid tumors via circulating tumor cell seeding pressure have also been attempted although these models did not take non-uniform fitness landscape into account[220,221]. However, challenges remain with regards to modeling clonality in diseases. For example, many models assume that tumors are composed of well-mixed populations of cells[222]. This is generally true in the hematopoietic system, but the solid malignancies are complicated by spatial intra-tumoral heterogeneity that has been demonstrated to influence tumor clonal dynamics[223]. As such, mathematical models that account for this feature would be particularly informative, especially if they are derived in tandem with multi-region intra-tumoral sequencing[197]. In addition to the fitness landscape of genetic alterations, future models should also account for extrinsic and microenvironmental factors such as hypnosia and angiogenesis that are implicated in tumor progression[224]. Moreover, as many recent studies have shown, oligoclonal cooperation plays an important role in tumorigenesis in many cancer types. Modeling oligoclonal cooperation at this juncture might be difficult because we do not yet know a lot about this phenomenon, but one would expect that this is the logical next step to take once we could accurately account for clonal fitness in modeling disease trajectory.

7.3 Cell-tagging to investigate clonal evolution

The advances in single-cell sequencing technologies over the recent years have enabled the research community to interrogate complex biological processes with unparalleled resolution[225]. Earlier optimizations of the technology focused on eliminating batch effects, and increasing the number of cells that could be sequenced in a single run. With these, we are now able to examine the RNA profiles, DNA damages, and epigenetic changes within a single cell. However, the results from these single-cell sequencing were essentially still a ‘snapshot’ capture of the evolutionary trajectories of the cells, because it was hard to discern if the cells sequenced at a later time point came from the same clonal populations of cells sequenced prior[225]. Over the recent 2 years, an interesting development has emerged – tagging individual cells with some forms of barcode (not unlike UMI in ECS)[226-228] that would allow researchers to trace the cell fate of tagged cells *in vitro* or *in vivo* over a longitudinal timeline. In essence, cells sharing the same barcode is presumed to have arisen from the same ancestral cell. With this new development, it opens new doors to investigate clonal evolution of diseases. For example, one could tagged every cells in a heterogynous 3D tumor organoid model, and examine the clonal trajectories of cells with distinct genetic profiles. Does a minor clone accumulate more mutations in response to treatment as part of the biological processes in resistance development? If so, what genetic alterations were acquired? These are some of the questions that can be asked now, given the resolution afforded by single-cell sequencing in tandem with cell barcoding.

7.4 Conclusion

Upon reflection, it has been a humbling journey that we could study these questions that were previously thought to be either untenable or clinically insignificant, of course this is due to generous support from the patients and the agencies responsible for them. This work has enabled new knowledge pertaining to clonality in various disease states in the pediatric and AYA cohorts. In summary, we found that 1) clonal hematopoiesis was common among neonates and young adults; 2) clonal hematopoietic mutations were stable over time and could engraft recipients during hematopoietic stem cell transplantation; 3) Down syndrome children with or without leukemia had distinct clonal mutation profiles; 4) disease ontogeny in NF1-glioblastoma was not due to bi-allelic loss of the *NF1* gene in a patient whom we examined. We hope that others would continue to expand on our findings. The dissertation is the beginning of my future scientific adventure to study clonality with more depth.

References

1. Secker-Walker LM. The meaning of a clone. *Cancer Genet. Cytogenet.* 1985; 16:87-88
2. Wainscoat JS, Fey MF. Assessment of clonality in human tumors: A review. *Cancer Research* 1990; 50:1355-1360
3. McGranahan N, Swanton C. Clonal heterogeneity and tumor evolution: Past, present, and the future. *Cell* 2017; 168(4):613-628
4. Nowell PC. The clonal evolution of tumor cell populations. *Science* 1976; 194:23-28
5. Beutler E, Yeh M, Fairbanks VF. Normal human female as mosaic of X-chromosome activity: studies using the gene for G6PD deficiency as a marker. *PNAS* 1962; 48:9-16
6. Linder D, Gartler SM. Glucose-6-phosphate dehydrogenase mosaicism: utilization as a cell marker in the study of leiomyomas. *Science* 1965; 150:67-69
7. Linder D. Gene loss in human teratomas. *PNAS* 1969; 63:699-704
8. Fialkow PJ, Gartler SM, Yoshida A. Clonal origin of chronic myelocytic leukemia in man. *PNAS* 1967; 58:1468-1471
9. Fialkow PJ, Klein G, Clifford P. Second malignant clone underlying a Burkitt-tumor exacerbation. Lancet 1972; 2:629-631
10. Fialkow PJ, Sagebiel RW, Gartler SM, Rimoin DL. Multiple cell origin of hereditary neurofibromas. *NEJM* 1971; 284:298-300
11. Rowley JD. Chromosome abnormalities in cancer. *Cancer Genet. Cytogenet.* 1980; 2:175-198
12. Maxam M, Gilbert W. A new method for sequencing DNA. *PNAS* 1977; 74:560-564
13. Sanger F, Nicklen S, Coulson R. DNA sequencing with chain-terminating inhibitors. *PNAS* 1977; 74:5463-5467

14. Noguchi S, Motomura K, Inaji H, Imaoka S, Koyama H. Clonal analysis of fibroadenoma and phyllodes tumor of the breast. *Cancer Research* 1993; 53:4071-4073

15. Mao L, Lee DJ, Tockman MS, Erozan YS, Askin F, Sidransky D. Microsatellite alterations as clonal markers for the detection of human cancer. *PNAS* 1994; 91:9871-9875

16. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* 2001; 409:860-921

17. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* 2004; 431:931-945

18. Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* 2006; 314:268-274

19. Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* 2007; 314:1108-1113

20. Beerenwinkel N, Antal T, Dingli D, Traulsen A, Kinzler KW, Velculescu VE *et al.* Genetic progression and the waiting time to cancer. *PLoS Computational Biology* 2007; 3(11):e255

21. Shendure J, Ji H. Next-generation DNA sequencing. *Nature Biotechnology* 2008; 26:1135-1145

22. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* 2008; 456:66-72

23. Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS *et al.* Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* 2012; 481:506-510

24. Shain Ah, Yeh I, Kovalyshyn I, Sriharan A, Talevich E, Gagnon A *et al.* The genetic evolution of melanoma from precursor lesions. *NEJM* 2015; 373:1926-1936

25. Yates LR, Knappskog S, Wedge D, Farmery JHR, Gonzalez S, Martincorena I *et al.* Genomic evolution of breast cancer metastasis and relapse. *Cancer Cell* 2017; 32:169-184

26. Rogers ZN, McFarland CD, Winters IP, Seoane JA, Brady JJ, Curtis C *et al.* The fitness landscape of tumor suppression in lung adenocarcinoma in vivo. *Nature Genetics* 2018; 50:483-486

27. Welch JS, Ley TJ, Link DC, Miller CA, Larson DE, Koboldt D *et al.* The origin and evolution of mutations in acute myeloid leukemia. *Cell* 2012; 150:264-278

28. Desai P, Mencia-Trinchant N, Savenkov O, Simon MS, Cheang G, Lee S *et al.* Somatic mutations precede acute myeloid leukemia years before diagnosis. *Nature Medicine* 2018; 24:1015-1023

29. Abelson S, Collord G, Ng SWK, Weissbrod O, Cohen OM, Niemeyer E *et al.* Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* 2018; 559:400-404

30. Young AL, Tong RS, Birmann BM, Druley TE. Clonal haematopoiesis and risk of acute myeloid leukemia. *Haematologica* 2019; doi:10.3324/haematol.2018.215269

31. Wong TN, Ramsingh G, Young AL, Miller CA, Touma W, Welch JS *et al.* Role of *TP53* mutations in the origin and evolution of therapy-related acute myeloid leukaemia. *Nature* 2015; 518:552-555

32. Gupta RG, Somer RA. Intratumor heterogeneity: Novel approaches for resolving genomic architecture and clonal evolution. *Molecular Cancer Research* 2017; 15:1127-1137

33. Sun R, Hu Z, Sottoriva A, Graham TA, Harpak A, Ma Z *et al.* Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nature Genetics* 2017; 49:1015-1024

34. Sottoriva A, Kang H, Ma Z, Graham TA, Salomon MP, Zhao J *et al.* A Big Bang model of human colorectal tumor growth. *Nature Genetics* 2015; 47:209-216

35. Pogrebniak KL, Curtis C. Harnessing tumor evolution to circumvent resistance. *Trends in Genetics* 2018; 34:639-651

36. Gerlinger M, Rowan AJ, Horswell S, Math M, Larkin J, Endesfelder D *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *NEJM* 2012; 366:883-891

37. Gerlinger M, Horswell S, Larkin J, Rowan AJ, Salm MP, Varela I *et al.* Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nature Genetics* 2014; 46:225-233

38. Ma P, Fu Y, Cai M, Yan Y, Jing Y, Zhang S *et al.* Simultaneous evolutionary expansion and constraint of genomic heterogeneity in multifocal lung cancer. *Nature Communications* 2017; 8:823

39. Siravegna G, Mussolin B, Buscarino M, Corti G, Cassingena A, Crisafulli G *et al.* Clonal evolution and resistance to EGFR blockade in the blood of colorectal cancer patients. *Nature Medicine* 2015; 21:795-801

40. Kohsaka S, Petronczki M, Solca F, Maemondo M. Tumor clonality and resistance mechanisms in EGFR mutation-positive non-small-cell lung cancer: implications for therapeutic sequencing. *Future Oncology* 2018; 15:637-652

41. Garg M, Nagata Y, Kanojia D, Mayakonda A, Yoshida K, Haridas KS *et al.* Profiling of somatic mutations in acute myeloid leukemia with FLT3-ITD at diagnosis and relapse. *Blood* 2015: 126:2491-2501

42. Kronke J, Bullinger L, Teleanu V, Tschurtz F, Gaidzik VI, Kuhn MW *et al.* Clonal evolution in relapsed NPM1-mutated acute myeloid leukemia. *Blood* 2013; 122:100-108

43. Shlush LI, Mitchell A, Heisler L, Abelson S, Ng SWK, Trotman-Grant A *et al.* Tracing the origins of relapse in acute myeloid leukaemia to stem cells. *Nature* 2017; 547:104-108

44. Blakely CM, Watkins TBK, Gini B, Chabon JJ, McCoach CE, McGranahan N *et al.* Evolution and clinical impact of co-occurring genetic alterations in advanced-stage EGFR-mutant lung cancers. *Nature Genetics* 2017; 49:1693-1704

45. Li X, Puri S, Negrao MV, Nilsson MB, Robichaux J, Boyle T *et al.* Landscape of EGFR-dependent and –independent resistance mechanisms to osimertinib and continuation therapy beyond progression in EGFR-mutant NSCLC. *Clinical Cancer Research* 2018; 24:6195-6203

46. Greaves M, Maley CC. Clonal evolution in cancer. *Nature* 2012; 481:306-313

47. Lacina L, Coma M, Dvorankova B, Kodet O, Melegova N, Gal P *et al.* Evolution of cancer progression in the context of Darwinism. *Anticancer Research* 2019; 39:1-16

48. Wood BL. Principles of minimal residual disease detection for hematopoietic neoplasms by flow cytometry. *Cytom. Part B Clin. Cytom.* 2016; 90:47-53

49. Spencer DH, Abel HJ, Lockwood CM, Payton JE, Szankasi P, Kelley TW *et al.* Detection of FLT3 internal tandem duplication in targeted, short-read-length, next-generation sequencing data. *J. Mol. Diagn*. 2013; 15:81-93

50. Levine RL, Valk PJM. Next-generation sequencing in the diagnosis and minimal residual disease assessment of acute myeloid leukemia. *Haematologica* 2019; 104:868-871

51. Borowitz MJ, Devidas M, Hunger SP, Bowman WP, Carroll AJ, Carroll WL *et al.* Clinical significance of minimal residual disease in childhood acute lymphoblastic leukemia and its

relationship to other prognostic factors: a Children's Oncology Group Study. *Blood* 2008; 111:5477-5485

52. Orfao A, Ortuno F, de Santiago M, Lopez A, San Miguel J. Immunophenotyping of acute leukemias and myelodysplatic syndromes. *Cytometry A.* 2004; 58:62-71
53. Loken MR, Alonzo TA, Pardo L, Gerbing RB, Raimondi SC, Hirsch BA *et al.* Residual disease detected by multidimensional flow cytometry signifies high relapse risk in patients with de novo acute myeloid leukemia: a report from Children's Oncology Group. *Blood* 2012; 120:1581-1588
54. Rubnitz JE, Inaba H, Dahl G, Ribeiro RC, Bowman WP, Taub J *et al.* Minimal residual disease-directed therapy for childhood acute myeloid leukaemia: results of the AML02 multicentre trial. *Lancet Oncology* 2010; 11:543-552
55. Al-Mawali A, Gillis D, Lewis I. The role of multiparameter flow cytometry for detection of minimal residual disease in acute myeloid leukemia. *American Journal of Clinical Pathology* 2009; 131:16-26
56. Patel JP, Gonen M, Figueroa ME, Fernandez H, Sun Z, Racevskis J *et al.* Prognostic relevance of integrated genetic profiling in acute myeloid leukemia. *NEJM* 2012; 366:1079-1089
57. Bashford-Rogers RJM, Nicolaou KA, Bartram J, Goulden NJ, Loizou L, Koumas L *et al.* Eye on the B-ALL: B-cell receptor repertoires reveal persistence of numerous B-lymphoblastic leukemia subclones from diagnosis to relapse. *Leukemia* 2016; 30:2312-2321
58. Morita K, Kantarjian HM, Wang F, Yan Y, Bueso-Ramos C, Sasaki K *et al.* Clearance of somatic mutations at remission and the risk of relapse in acute myeloid leukemia. *Journal of Clinical Oncology* 2018; 36:1788-1797

59. Klco JM, Miller CA, Griffith M, Petti A, Spencer DH, Ketkar-Kulkarni S *et al.* Association between mutation clearance after induction therapy and outcomes in acute myeloid leukemia. *JAMA* 2015; 314:811-822

60. Genovese G, Kahler AK, Handsaker RE, Lindberg J, Rose RA, Bakhoum SF *et al.* Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *NEJM* 2014; 371:2477-2487

61. Jaiswal S, Fontanillas P, Flannick J, Manning A, Grauman PV, Mar BG *et al.* Age-related clonal hematopoiesis associated with adverse outcomes. *NEJM* 2014; 371:2488-2498

62. Xie M, Lu C, Wang J, McLellan MD, Johnson KJ, Wendl MC *et al.* Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nature Genetics* 2014; 20:1472-1478

63. Steensma DP, Bejar R, Jaiswal S, Lindsley RC, Sekeres MA, Hasserjian RP *et al.* Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. *Blood* 2015; 126:9-16

64. Shlush LI. Age-related clonal hematopoiesis. Blood 2018; 131:496-504

65. Spencer DH, Russler-Gemain DA, Ketkar S, Helton NM, Lamprecht TL, Fulton RS *et al.* CpG island hypermethylation mediated by DNMT3A is a consequence od AML progression. *Cell* 2017; 168:801-816

66. Chin R, Chen K, Usmani A, Chua C, Harris PK, Binkley M *et al.* Detection of solid tumor molecular residual disease using ctDNA. *Molecular Diagnosis & Therapy* 2019; 23:311-331

67. Garcia-Murillas I, Schiavon G, Weigelt B, Ng C, Hrebien S, Cutts RJ *et al.* Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. *Science Translational Medicine* 2015; 7:302ra133

68. Chaudhuri AA, Chabon JJ, Lovejoy AF, Newman AM, Stehr H, Azad RD *et al.* Early detection of molecular residual disease in localized lung cancer by circulating tumor DNA profiling. *Cancer Discovery* 2017; 7:1394-1403

69. Tie J, Wang Y, Tomasetti C, Li L, Springer S, Kinde I *et al.* Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer. *Science Translational Medicine* 2016; 8:346ra92

70. Kamila Naxerova, Reiter JG, Brachtel E, Lennerz JK, van de Wetering M, Rowan A *et al.* Origins of lymphatic and distant metastases in human colorectal cancer. *Science* 2017; 357:55-60

71. Hu Z, Ding J, Ma Z, Sun R, Seoane JA, Shaffer JS *et al.* Quantitative evidence for early metastatic seeding in colorectal cancer. *Nature Genetics* 2019; 51:1113-1122

72. Cady B. Lymph node metastases: Indicators, but not governors of survival. *Arch. Surg.* 1984; 119:1067-1072

73. Casadaban L, Rauscher G, Aklilu M, Villenes D, Freels D, Maker AV *et al.* Adjuvant chemotherapy is associated with improved survival in patients with stage II colon cancer. *Cancer* 2016; doi:10.1002/cncr.30181

74. Dudley JC, Schroers-Martin J, Lazzareschi DV, Shi WY, Chen SB, Esfahani MS *et al.* Detection and surveillance of bladder cancer using urine tumor DNA. *Cancer Discovery* 2019; 9:500-509

75. Abbosh C, Birkbak NJ, Wilson GA, Jamal-Hanjani M, Constantin T, Salari R *et al*. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature* 2017; 545:446-451

76. Risitano AM, Selleri Carmine. Clonal non-malignant hematological disorders: Unraveling molecular pathogenic mechanisms to develop novel targeted therapeutics. *Transl. Med. UniSa.* 2014; 8:1-3

77. Gilliland DG, Blanchard KL, Bunn HF. Clonality in acquired hematologic disorders. *Annual Review of Medicine* 1991; 42:491-506

78. Nash R, Storb R, Neiman P. Polyclonal reconstitution of human marrow after allogeneic bone marrow transplantation. *Blood* 1988; 72:2031-2037

79. Shepherd BE, Guttorp P, Lansdorp PM, Abkowitz JL. Estimating human hematopoietic stem cell kinetics using granulocyte telomere lengths. *Exp. Hematol.* 2004; 32:1040-1050

80. Dingli D, Pacheco JM. Ontogenic growth of the haematopoietic stem cell pool in humans. *Proc. Biol. Sci.* 2007; 274:2497-2501

81. Mahe E, Pugh T, Kamel-Reid S. T cell clonality assessment: Past, present and future. *Journal of Clinical Pathology* 2018; 71:195-200

82. Groenen PJTA, van Raaij A, van Altena MC, Rombout PM, van Krieken JMH. A practical approach to diagnostic Ig/TCR clonality evaluation in clinical pathology. *Journal of Hematopathology* 2012; 5:16-25

83. van Dongen JJ, Wolvers-Tettero IL. Analysis of immunoglobulin and T cell receptor genes. Part 1: Basic and technical aspects. Clin. Chim. Acta 1991; 198:1-92

84. Hershberg U, Prak ETL. The analysis of clonal expansions in normal and autoimmune B cell repertoires. *Phil. Trans. R. Soc. B* 2015; 370:20140239

85. Anderson SM, Khalil A, Uduman M, Hershberg U, Louzoun Y, Haberman AM *et al.* Taking advantage: high affinity B cells in the germinal center have lower death rates, but similar rates of division, compared to low affinity cells. *J. Immunol.* 2012; 183:7314-7325

86. Burkovitz A, Sela-Culang I, Ofran Y. Large-scale analysis of somatic hypermutations in antibodies reveals which structural regions, positions and amino acids are modified to improve affinity. *FEBS J.* 2014; 281:306-319

87. Gibson CJ, Steensma DP. New insights from studies of clonal hematopoiesis. *Clinical Cancer Research* 2018; 19:4633-4642

88. Jaiswal S, Natarajan P, Silver AJ, Gibson CJ, Bick AG, Shvartz E *et al.* Clonal hematopoiesis and risk of atherosclerotic cardiovascular disease. *NEJM* 2017; 377:111-121

89. Fuster JJ, MacLauchlan S, Zuriaga MA, Polackal MY, Ostriker AC, Chakraborty R *et al.* Clonal hematopoiesis associated with TET2 deficiency accelerates atherosclerosis development in mice. *Science* 2017; 355:842-847

90. Sano S, Oshima K, Wang Y, MacLauchlan S, Katanasaka Y, Sano M *et al.* Tet2-mediated clonal hematopoiesis accelerates heart failure through a mechanism involving the IL-1B/NLRP3 inflammasome. *J. Am. Coll. Cardiol.* 2018; 71:875-886

91. Gibson CJ, Kennedy JA, Nikiforow S, Kou FC, Alyea EP, Ho V *et al.* Donor-engrafted CHIP is common among stem cell transplant recipients with unexplained cytopenias. *Blood* 2017; 130:91-94

92. Frick M, Chan W, Arends CM, Hablesreiter R, Halik A, Heuser M *et al.* Role of donor clonal hematopoiesis in allogeneic hematopoietic stem-cell transplantation. *Journal of Clinical Oncology* 2019; 37:375-385

93. Young AL, Challen GA, Birmann BM, Druley TE. Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. *Nature Communications* 2016; 7:12484

94. Yasuda T, Ueno T, Fukumura K, Yamato A, Ando M, Yamaguchi H *et al.* Leukemic evolution of donor-derived cells harboring IDH2 and DNMT3A mutations after allogeneic stem cell transplantation. *Leukemia* 2014; 28:426-428

95. Spencer DH, Tyagi M, Vallania F, Bredemeyer AJ, Pfeifer JD, Mitra RD *et al.* Performance of common analysis methods for detecting low-frequency single nucleotide variants in targeted next-generation sequence data*. J. Mol. Diagn*. 2014; 16:75-88

96. Schmitt MW, Fox EJ, Prindle MJ, Reid-Bayliss KS, True LD, Radich JP *et al.* Sequencing small genomic targets with high efficiency and extreme accuracy. *Nature Methods* 2015; 12:423-425

97. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. *PNAS* 2011; 108:9530-9535

98. Young AL, Wong TN, Hughes AEO, Heath SE, Ley TJ, Link DC, Druley TE. Quantifying ultra-rare pre-leukemic clons via targeted error-corrected sequencing. *Leukemia* 2015; 29:1608-1611

99. Duncacage EJ, Jacoby MA, Chang GS, Miller CA, Edwin N, Shao J *et al.* Mutation clearance after transplantation for myelodysplastic syndrome. *NEJM* 2018; 379:1028-1041

100. Salk JJ, Schmitt MW, Loeb LA. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. *Nature Review Genetics* 2018; 19:269-285

101. Petrackova A, Vasinek M, Sedlarikova L, Dyskova T, Schneiderova P, Novosad T *et al.* Standardization of sequencing coverage depth in NGS: Recommendation for detection of clonal and subclonal mutations in cancer diagnostics. *Frontier in Oncology* 2019; doi:10.3389/fonc.2019.00851

102. Wong WH, Tong RS, Young AL, Druley TE. Rare event detection using error-corrected DNA and RNA sequencing. *Journal of Visualized Experiment* 2018; 3;138

103. Hoang ML, Kinde I, Tomasetti C, McMahon KW, Rosenquist TA, Grollman AP *et al.* Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. *PNAS* 2016; 113:9846-9851

104. O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm E, Coe BP *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* 2012; 485:246-250

105. Patel JP, Gonen M, Figueroa ME, Fernandez ME, Sun Z, Racevskis J *et al.* Prognostic relevance of integrated genetic profiling in acute myeloid leukemia. *NEJM* 2012; 366:1079-1089

106. Gazzola A, Mannu C, Rossi M, Laginestra MA, Sapienza MR Fuligni F *et al.* The evolution of clonality testing in the diagnosis and monitoring of hematological malignancies. *Ther. Adv. Hematol*. 2014; 5:35-47

107. Lin E, Cao T, Nagrath S, King MR. Circulating tumor cells: Diagnostic and therapeutic applications. *Annual Review of Biomedical Engineering* 2018; 20:329-352

108. Huang K, Mashl RJ, Wu Y, Ritter DI, Wang J, Oh C *et al.* Pathogenic germline variants in 10,389 adult cancers. *Cell* 2018; 173:355-370

109. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A *et al.* Comprehensive characterization of cancer driver genes and mutations. *Cell* 2018; 173:371-385

110. Farrar JE, Schuback HL, Ries RE, Wai D, Hampton OA, Trevino LR *et al.* Genomic profiling of pediatric acute myeloid leukemia reveals a changing mutation landscape from disease diagnosis to relapse. *Cancer Research* 2016; 76:2197-2205

111. Shugay M, Zaretsky AR, Shagin DA, Shagina IA, Volchenkov IA, Shelenkov AA *et al.* MAGERI: Computational pipeline for molecular-barcoded targeted resequencing. *PLoS Computational Biology* 2017; 13:e1005480

112. Krimmel JD, Schmitt MW, Harrell MI, Agnew KJ, Kennedy SR, Emond MJ *et al.* Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic TP53 mutations in noncancerous tissues. *PNAS* 2016; 113:6005-6010

113. Phallen J, Sausen M, Adleff V, Leal A, Hruban C, Whilte J *et al.* Direct detection of early-stage cancers using circulating tumor DNA. *Science Translational Medicine* 2017; 9:eaan2415

114. Egorov ES, Merzlyak EM, Shelenkov AA, Britanova OV, Sharonov GV, Staroverov DB *et al.* Quantitative profiling of immune repertoires for minor lymphocyte counts using unique molecular identifiers. *Journal of Immunology* 2015; 194:6155-6163

115. Milholland Brandon, Auton Adam, Suh Yousin, Vijg Jan. Age-related somatic mutations in the cancer genome. *Oncotarget* 2015; 6:24627-24635

116. Blokzijl, de Ligt J, Jager M, Sasselli V, Roerink S, Sasaki N *et al.* Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* 2016; 538:260-264

117. Pietras EM, Warr MR, Passegue E. Cell cycle regulation in hematopoetic stem cells. *Journal of Cell Biology* 2011; 195:709

118. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* 2015; 526:68-74

119. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G *et al.* Integrative Genomics Viewer. *Nature Biotechnology* 2011; 29:24-26

120. Koboldt D, Zhang Q, Larson D, Shen D, McLellan M, Miller LL *et al.* VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research* 2012; 22:568-576

121. Osorio FG, Huber AR, Oka R, Verheul M, Patel SH Hasaart K *et al.* Somatic mutations reveal lineage relationships and age-related mutagenesis in human hematopoiesis. *Cell Reports* 2018; 25:2308-2316

122. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV *et al.* Signatures of mutational processes in human cancer. *Nature* 2013; 500:415-421

123. Takizawa H, Regoes RR, Boddupalli CS, Bonhoeffer S, Manz MG. Dynamic variation in cycling of hematopoietic stem cells in steady state and inflammation. *Journal of Experimental Medicine* 2011; 208:273

124. Wong TN, Miller CA, Jotte MRM, Bagegni N, Baty JD, Schmidt AP *et al.* Cellular stressors contribute to the expansion of hematopoietic clones of varying leukemic potential. *Nature Communications* 2018; 9:455

125. Watson CJ, Papula A, Poon YPG, Wong WH, Young AL, Druley TE *et al.* The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *BioRxiv* 2019; doi:10.1101/569566

126. Bejar R, Sekeres MA. Attack of the clones: CHIP in the clinic. *The Hematologist* 2019; 16(1).

127. Paashuis-Lew YR, Heddle JA. Spontaneous mutation during fetal development and post-natal growth. *Mutagenesis* 1998; 13:613-617

128. Be The Match. Why a donor's age matters. Available from https://bethematch.org/transplant-basics/matching-patients-with-donors/why-donor-age-matters/

129. Wong WH, Bhatt S, Trinkaus K, Pusic I, Elliott K, Mahajan N *et al.* Engraftment of rare, pathogenic donor hematopoietic clones in unrelated hematopoietic stem cell transplantation. *Science Translational Medicine* 2019; in press

130. Shenoy S, Boelens JJ. Advances in unrelated and alternative donor hematopoietic cell transplantation for non-malignant disorders. *Curr. Opin. Pediatr* 2015; 27:9-17

131. Gyurkocza B, Rezvani A, Storb RF. Allogeneic hematopoietic cell transplantation: the state of the art. *Expert Rev. Hematol*. 2010; 3:285-299

132. Afessa B, Peters SG. Major complications following hematopoietic stem cell transplantation. *Semin. Respir. Crit. Care Med.* 2006; 27:297-309

133. Scott JM, Armenian S, Giralt S, Moslehi J, Wang T, Jones LW. Cardiovascular disease following hematopoietic stem cell transplantation: Pathogenesis, detection, and the cardioprotective role of aerobic training. *Crit. Rev. Oncol. Hematol.* 2016; 98:222-234

134. Kato M, Yamashita T, Suzuki R, Matsumoto K, Nishimori H, Takahashi S *et al.* Donor cell-derived hematological malignancy: a survey by the Japan Society for Hematopoietic Cell Transplantation. *Leukemia* 2016; 30:1742-1745

135. Jian J, Hao H, Yuan C. Donor-cell-derived myelodysplastic syndrome involving U2AF1 mutation developing 8 years after matched unrelated bone marrow transplantation for acute leukemia and literature review. *European Journal of Biology and Medical Science Research* 2017; 5:1-7

136. Bowman RL, Busque L, Levine RL. Clonal hematopoiesis and evolution to hematopoietic malignancies. *Cell Stem Cell* 2018; 22:157-170

137. Cancer Genome Atlas Research Network, Ley TJ, Miller C, Ding L, Raphael BJ, Mungall AJ *et al.* Genomic and epigenomic landscapes od adult de novo acute myeloid leukemia. *NEJM* 2013; 368:2059-2074

138. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research* 2018; 47:886-894

139. Arends CM, Galan-Sousa J, Hoyer K, Chan W, Jager M, Yoshida K *et al.* Hematopoietic lineage distribution and evolutionary dynamics of clonal hematopoiesis. *Leukemia* 2018; 32:1908-1919

140. Charlesworth B. The effects of deleterious mutations on evolution at linked sites. *Genetics* 2012; 190:5-22

141. Dorscheimer L, Assmus B, Rasper T, Ortmann CA, Ecke A, Abou-El-Ardat K *et al.* Association of mutations contributing to clonal hematopoiesis with prognosis in chronic ischemic heart failure. *JAMA Cardiology* 2019; 4:25-33

142. Horton SJ, Giotopoulos G, Yun H, Vohra S, Sheppard O, Bashford-Rogers R *et al.* Early loss of Crebbp confers malignant stem cell properties on lymphoid progenitors. *Nature Cell Biology* 2017; 19:1093-1104

143. Rosche WA, Foster PL. The role of transient hypermutators in adaptive mutation in Escherichia coli. *PNAS* 1999; 96:6862-6867

144. Galhardo RS, Hastings PJ, Rosenberg SM. Mutation as a stress response and regulation of evolvability. *Crit. Rev. Biochem. Mol. Biol.* 2007; 42:399-435

145. Feinberg AP, Koldobskiy MA, Gondor A. Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nature Review Genetics* 2016; 17:284-299

146. Xiao H, Wang LM, Luo Y, Lai X, Li C, Shi J *et al.* Mutations in epigenetic regulators are involved in acute lymphoblastic leukemia relapse following allogeneic hematopoietic stem cell transplantation. *Oncotarget* 2016; 7:2696-2708

147. Jeong M, Park HJ, Celik H, Ostrander EL, Reyes JM, Guzman A *et al.* Loss of Dnmt3a immortalizes hematopoietic stem cells in vivo. *Cell Reports* 2018; 23:1-10

148. Bittles AH, Glasson EJ. Clinical, social, and ethical implications of changing life expectancy in Down syndrome. *Dev. Med. Child. Neurol.* 2004; 46:282-286

149. Roizen NJ, Patterson D. Down’s syndrome. *Lancet* 2003; 361:1281-1289

150. Spahis JK, Wilson GN. Down syndrome: perinatal complications and counseling experiences in 216 patients. *Am. J. Med. Genet.* 1999; 89:96-99

151. Ali FE, Al-Bustan MA, Al-Busairi WA, Al-Mulla FA, Esbaita EY. Cervical spine abnormalities associated with Down syndrome. *Int. Orthop*. 2006; 30:284-289

152. Graber E, Chacko E, Regelmann MO, Costin G, Rapaport R. Down syndrome and thyroid function. *Endocrinol. Metab. Clin. North Am*. 2012; 41:735-745

153. Bergstrom S, Carr H, Petersson G, Stephansson O, Bonamy AE, Dahlstrom A *et al.* Trends in congenital heart defects in infants with Down syndrome. *Pediatrics* 2016; 138:e20160123

154. Wechsler J, Greene M, McDevitt MA, Anastasi J, Karp JE, Le Beau MM *et al*. Acquired mutations in *GATA1* in the megakaryoblastic leukemia of Down syndrome. *Nature Genetics* 2002; 32(1):148-52.

155. Hasle H, Clemmensen IH, Mikkelsen M. Risk of leukaemia and solid tumours in individuals with Down’s syndrome. *Lancet* 2000; 355:165-169

156. Ceballos I, Nicole A, Briant P, Grimber G, Delacourt A, Flament S *et al.* Expression of human CuZn superoxide dismutase gene in transgenic mice: model for gene dosage effect in Down’s syndrome. *Free Radical Res. Commun.* 1991; 13:581-589

157. Maluf SW, Erdtmann B. Genomic instability in Down syndrome and Fanconi anemia assessed by micronucleus analysis and single-cell gel electrophoresis. *Cancer Genetics and Cytogenetics* 2001; 124:71-75

158. Morawiec Z, Janik K, Kowalski M, Stetkiewicz T, Szaflik J, Morawiec-Bajda A *et al.* DNA damage and repair in children with Down’s syndrome. *Mutat. Res*. 2008; 635:118-123

159. Ahmed M, Sternberg A, Hall G, Thomas A, Smith O, O’Marcaigh A *et al*. Natural history of *GATA1* mutations in Down syndrome. *Blood* 2004; 103:2480-2489.

160. Alford KA, Reinhardt K, Garnett C, Norton A, Böhmer K, von Neuhoff C *et al*. Analysis of *GATA1* mutations in Down syndrome transient myeloproliferative disorder and myeloid leukemia. *Blood* 2011; 118(8):2222-38.

161. Takahashi S, Komeno T, Suwabe N, Yoh K, Nakajima O, Nishimura S *et al.* Role of GATA1 in proliferation and differentiation of definitive erythroid and megakaryocytic cells in vivo. *Blood* 1998; 92:434-442

162. Bhatnagar N, Nizery L, Tunstall O, Vyas P, Roberts I. Transient Abnormal Myelopoiesis and AML in Down Syndrome: an Update. *Current Hematologic Malignancy Reports* 2016; 11:333-341

163. Gamis AS, Alonzo TA, Gerbing RB, Hilden JM, Sorrell AD, Sharma M *et al.* Natural history of transient myeloproliferative disorder clinically diagnosed in Down syndrome

neonates: a report from the Children's Oncology Group Study A2971. *Blood* 2011; 118:6752-6759

164. Kanezaki R, Toki T, Terui K, Xu G, Wang R, Shimada A *et al.* Down syndrome and GATA1 mutations in transient abnormal myeloproliferative disorder: mutation classes correlate with progression to myeloid leukemia. *Blood* 2010; 116:4631-4638

165. Malinge S, Izraeli S, Crispino JD. Insights into the manifestations, outcomes, and mechanisms of leukemogenesis in Down syndrome. *Blood* 2009; 113:2619-2628

166. Roberts I, Alford K, Hall G, Juban G, Richmond H, Norton A *et al.* GATA1-mutated clones are frequent and often unsuspected in babies with Down syndrome: identification of a population at risk of leukemia. *Blood* 2013; 112:3908-3917

167. Yoshida K, Toki T, Okuno Y, Kanezaki R, Shiraishi Y, Sato-Otsubo A *et al*. The landscape of somatic mutations in Down syndrome-related myeloid disorders. *Nature Genetics* 2013; 45:1293-1299.

168. Obenauer JC, Kavelaars FG, Sanders MA, de Vries ACH, de Haas V, Beverloo HB et al. Recurrently affected genes in juvenile myelomonocytic leukaemia. British Journal of Haematology 2018; 182:135-138

169. Steensma DP. Clinical consequences of clonal hematopoiesis of indeterminate potential. *Blood Advances* 2018; 2(22):3404-3410

170. Fisher JB, McNutty M, Burke MJ, Crispino JD, Rao S. Cohesin mutations in myeloid malignancies. *Trends in Cancer* 2017; 3:282-293

171. Li Z, Godinho FJ, Klusmann JH, Garriga-Canut M, Yu C, Orkin SH. Developmental stage-selective effect of somatically mutated leukemogenic transcription factor *GATA1*. *Nature Genetics* 2005; 37(6):613-619

172. Labuhn M, Perkins K, Matzk S, Varghese L, Garnett C, Papaemmanuil E *et al*. Mechanisms of progression of myeloid preleukemia to transformed myeloid leukemia in children with Down syndrome. *Cancer Cell* 2019; 36:123-138.

173. Pan X, Ohneda O, Ohneda K, Lindeboom F, Iwata F, Shimizu R *et al*. Graded levels of *GATA1* expression modulate survival, proliferation, and differentiation of erythroid progenitors. *Journal of Biological Chemistry* 2005; 280:22385-22394

174. Nei Y, Obata-Ninomiya K, Tsutsui H, Ishiwata K, Miyasaka M, Matsumoto K *et al*. *GATA-1* regulates the generation and function of basophils. *PNAS* 2013; 110(46):18620-18625.

175. Janiszewska M, Tabassum DP, Castaño Z, Cristea S, Yamamoto KN, Kingston NL *et al*. Subclonal cooperation drives metastasis by modulating local and systemic immune microenvironment. *Nature Cell Biology* 2019; 21:878-888.

176. Vinci M, Burford A, Molinari V, Kessler K, Popov S, Clarke M *et al*. Functional diversity and cooperativity between subclonal populations of pediatric glioblastoma and diffuse intrinsic pontine glioma cells. *Nat Med* 2018; 24(8):1204-1215.

177. Chen MJ, Yokomizo T, Zeigler BM, Dzierzak E, Speck NA. *Runx1* is required for the endothelial to haematopoietic cell transition but not thereafter. *Nature* 2009; 457:887-891.

178. Horvath S, Garagnani P, Bacalini MG, Pirazzini C, Salvioli S, Gentilini D *et al*. Accerelated epigenetic aging in Down syndrome. *Aging Cell* 2015; 14(3):491-495

179. Wong WH, Junck L, Druley TE, Gutmann DH. NF1-glioblastoma clonal profiling reveals KMT2B mutations as potential somatic oncogenic events. *Neurology* 2019; doi:10.1212/WNL.0000000000008623

180. Gutmann DH, Rasmussen SA, Wolkenstein P, MacCollin MM, Guha A, Inskip PD *et al*. Gliomas presenting after age 10 in invidivuals with neurofibromatosis type 1. *Neurology* 2002; 59:759-761

181. Friedman JM, Birch PH. Type 1 neurofibromatosis: A descriptive analysis of the disorder in 1728 patients. *American Journal of Human Genetics* 1997; 70:138-143

182. Karaconji T, Whist E, Jamieson RV, Flaherty MP, Grigg JRB. Neurofibromatosis Type 1: Review and update on emerging therapies. *Asia-Pacific Journal of Ophthalmology* 2019; 8:62-72

183. Hirbe AC, Gutmann DH. Neurofibromatosis Type 1: A multidisciplinary approach to care. *The Lancer Neurology* 2017; 13:834-843

184. Gutmann DH, Parada LF, Silva AH, Ratner N. Neurofibromatosis type 1: Modeling CNS dysfunction. *Journal of Neuroscience* 2012; 32:14087-14093

185. Brems H, Beert E, de Ravel T, Legius E. Mechanisms in the pathogenesis of malignant tumours in neurofibromatosis type 1. *The Lancet Oncology* 2009; 10:508-515

186. Abdolrahimzadeh B, Piraino DC, Albanese G, Cruciani F, Rahimi S. Neurofibromatosis: An update of ophthalmic characteristics and applications of optical coherence tomography. *Clin. Opthalmol*. 2016; 10:851-860

187. de Blank PMK, Fisher MJ, Liu GT, Gutmann DH, Listermick R, Ferner RE, Avery RA. Optic pathway gliomas in neurofibromatosis type 1: An update: Surveillance, treatment indications, and biomarkers of vision. *Journal of Neuro-Ophthalmology* 2017; 37:23-32

188. Ferner RE, Huson SM, Thomas N, Moss C, Willshaw H, Evans DG *et al.* Guidelines for the diagnosis and management of individuals with neurofibromatosis 1. *J. Med. Genet.* 2007; 44:81-88

189. Stern JMD, DiGiacinto GVMD, Housepian EMMD. Neurofibromatosis and optic glioma: Clinical and morphological correlations. *Neurosurgery* 1979; 4:524-528

190. Graf N. Glioblastoma in children with NF1: the need for basic research. *Pediatr. Blood Cancer* 2010; 54:870-871

191. Rasmussen SA, Yang Q, Friedman JM. Mortality in neurofibromatosis 1: An analysis using U.S. death certificates. *American Journal of Human Genetics* 2001; 68:1110-1118

192. D'Angelo F, Ceccarelli M, Tala, Garofano L, Zhang J, Frattini V *et al.* The molecular landscape of glioma in patients with Neurofibromatosis 1. *Nature Medicine* 2019; 25:176-187

193. Gutmann DH, McLellan MD, Hussain I, Wallis JW, Fulton LL, Magrini V *et al.* Somatic neurofibromatosis type 1 (NF1) inactivation characterizes NF1-associated pilocytic astrocytoma. *Genome Research* 2013; 23:431-439

194. Jeong T, Yee G. Glioblastoma in a patient with Neurofibromatosis Type 1: A case report and review of the literature. *Brain Tumor Research and Treatment* 2014; 2:36-38

195. Hakan T, Aker FV. Case report on a patient with neurofibromatosis type 1 and a frontal cystic glioblastoma. *Neurol. Neurochir. Pol.* 2008; 42:362-365

196. Fortunato JT, Reys B, Singh P, Pan E. Brainstem glioblastoma multiforme in a patient with NF1. *Anticancer Research* 2018; 38:4897-4900

197. Hu Z, Curtis C. Inferring tumor phylogenies from multi-region sequencing. *Cell Systems* 2016; 3:12-14

198. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* 2012; 28:1811-1817

199. Ha G, Roth A, Khattra J, Ho J, Yap D, Prentice LM *et al.* TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Research* 2014; 24:1881-1893

200. Deshwar AG, Vembu S, Yung CK, Jang GH, Stein J, Morris Q. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology* 2015;16:35

201. Nikbakht H, Panditharatna E, Mikael LG, Li R, Gayden T, Osmond M *et al.* Spatial and temporal homogeneity of driver mutations in diffuse intrinsic pontine glioma. *Nature Communications* 2016; 7:11185

202. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A *et al.* Comprehensive characterization of cancer driver genes and mutations. *Cell* 2018; 17:371-385

203. Zhu Y, Guignard F, Zhao D, Liu L, Burns DK, Mason RP *et al.* Early inactivation of p53 tumor suppressor gene cooperating with NF1 loss induces malignant astrocytoma. *Cancer Cell* 2005; 8:119-130

204. Suzuki H, Aoki K, Chiba K, Sato Y, Shiozawa Y, Shiraishi Y *et al.* Mutational landscape and clonal architecture in grade II and III gliomas. *Nature Genetics* 2015; 47:458-468

205. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA *et al.* The cancer genome atlas pan-cancer analysis project. *Nature Genetics* 2013; 45:1113-1120

206. Fontebasso AM, Liu XY, Sturm D, Jabado N. Chromatin remodeling defects in pediatric and young adult glioblastoma: a tale of a variant histone 3 tail. *Brain Pathology* 2013; 23:210-216

207. Liau BB, Sievers C, Donohue K, Gillespie SM, Flavahan WA, Miller TE *et al.* Adaptive chromatin remodeling drives glioblastoma stem cell plasticitiy and drug tolerance. *Cell Stem Cell* 2017; 20:223-246

208. Gagne LM, Boulay K, Topisirovic I, Huot M, Mallete F. Oncogenic activities of IDH1/2 mutations: from epigenetics to cellular signaling. *Trends in Cell Biology* 2018; 27:738-752

209. Janke R, Iavarone AT, Rine J. Oncometabolite D-2-Hydroxyglutarate enhances gene silencing through inhibition of specific H3K36 histone demethylases. *eLife* 2017; 6:e22451

210. Love S. Post mortem sampling of the brain and other tissues in neurodegenerative disease. *Histopathology* 2004; 44:309-317

211. Paulukonis ST, Eckman JR, Snyder AB, Hagar W, Feuchtbaum LB, Zhou M *et al.* Defining sickle cell disease mortality using a population-based surveillance system, 2004 through 2008. *Public Health Rep*. 2016; 131:367-375

212. Brunson A, Keegan THM, Bang H, Mahajan A, Paulukonis S, Wun T. Increased risk of leukemia among sickle cell disease patients in California. *Blood* 2017; 130:1597-1599

213. Ho S, Johnson A, Tarapore P, Janakiram V, Zhang X, Leung Y. Environmental epigenetics and its implication on disease risk and health outcomes. *ILAR Journal* 2012; 53:289-305

214. Bateson P, Barker D, Clutton-Brock T, Deb D, D'Udine B, Foley RA *et al.* Developmental plasticity and human health. *Nature* 2004; 430:419-421

215. Pellegrino M, Sciambi A, Treusch S, Durruthy-Durruthy R, Gokhale K, Jacob J *et al.* High-throughput single-cell DNA sequencing of acute myeloid leukemia tumors with droplet microfluidics. *Genome Research* 2018; 28:1345-1352

216. Lo P, Zhou Q. Emerging techniques in single-cell epigenomics and their applications to cancer research. *Journal of Clinical Genomics* 2018; doi:10.4172/JCG.1000103

217. Rohan TE, Miller CA, Li T, Wang Y, Loudig O, Ginsberg M *et al.* Somatic mutations in benign breast disease tissue and risk of subsequent invasive breast cancer. *British Journal of Cancer* 2018; 118:1662-1664

218. Suda K, Nakaoka H, Yoshihara K, Ishiguro T, Tamura R, Mori Y *et al.* Clonal expansion and diversification of cancer-associated mutations in endometriosis and normal endometrium. *Cell Reports* 2018; 24:1777-1789

219. Altrock PM, Liu LL, Michor F. The mathematics of cancer: Integrating quantitative methods. *Nature Review Cancer* 2015; 15:730-745

220. Heyde A, Reiter JG, Naxerova K, Nowak MA. Consecutive seeding and transfer of genetic diversity in metastasis. *PNAS* 2019; 116:14129-14137

221. Reiter JG, Makohon-Moore A, Gerold JM, Heyde A, Attiyeh MA, Kohutek ZA *et al.* Minimal functional driver gene heterogeneity among untreated metastases. *Science* 2018; 361:1033-1037

222. Bozic I, Reiter JG, Allen B, Antal T, Chatterjee K, Shah P *et al.* Evolutionary dynamics of cancer in response to targeted combination therapy. *Elife* 2013; 2:e00747

223. Martens EA, Kostadinov R, Maley CC, Hallatschek O. Spatial structure increases the waiting time for cancer. *New Journal of Physics* 2011; 13:115014

224. McDougall SR, Anderson AR, Chaplain MA. Mathematical modeling of dynamic adaptive tumour-induced angiogenesis: Clinical implications and therapeutic targeting strategies. *Journal of Theoretical Biology* 2006; 241:564-589

225. Guo C, Kong W, Kamimoto K, Rivera-Gonzalez G, Yang X, Kirita Y *et al.* CellTag Indexing: genetic barcode-based sample multiplexing for single cell genomics. *Genome Biology* 2019; 20:90

226. Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* 2017; 357:89-94

227. Stoeckius M, Zheng S, Houck-Loomis B, Hao S, Yeung BZ, Mauck WM *et al.* Cell hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biology* 2018; 19:224

228. Biddy BA, Kong W, Kamimoto K, Guo C, Waye SE, Sun T *et al.* Single-cell mapping of lineage and identity in direct reprogramming. *Nature* 2018; 564:218-224