Spring 5-15-2018

Discovering Rare Hematopoietic Clones Harboring Leukemia-Associated Mutations Using Error-Corrected Sequencing

Andrew Lee Young
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Molecular Genetics and Genomics

Dissertation Examination Committee:
Todd E. Druley, Chair
Donald F. Conrad
Timothy J. Ley
Daniel C. Link
Matthew J. Walter

Discovering Rare Hematopoietic Clones Harboring Leukemia-Associated Mutations Using Error-Corrected Sequencing

by
Andrew Lee Young

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

May 2018
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

First, I would like thank Todd Druley for fostering a creative and positive laboratory environment. His optimism for scientific discovery and willingness to embark on high risk ventures made all of this work possible. I could not imagine a better training environment to explore genomics, merging both experimental and computational techniques.

The members of the Druley Lab further promoted this positive and productive research environment. I thank Mark Valentine, Andrew Hughes, Wing Hing Wong, Spencer Tong and Sara Chasnoff for their enthusiasm, excitement, and rigor. I have learned a tremendous amount from them and enjoyed my time in the lab. I thank our funding sources including the American Society of Hematology, the National Institutes of Health, the Children's Discovery Institute of Washington University and St Louis Children's Hospital, the Washington University Medical Scientist Training Program training grant, and Hyundai Hope on Wheels. I also thank Jeffery Gordon and the Center for Genome Sciences and Systems Biology (CGSSB). The technology we developed would not have been possible without the infrastructure of the CGSSB. Specifically, I would like to thank Jessica Hoisington-Lopez, Eric Martin and Brian Koebbe.

I want to thank my thesis committee. Their expansive knowledge of genetics, stem cell biology, hematopoiesis, and leukemia is unparalleled. It was exciting and humbling to train under their guidance. I have learned a great deal from our long discussions, thoughtful feedback and productive collaborations. I thank them for the time and resources they have invested in my education.

I thank the Washington University Medical Scientist Training Program (MSTP) for creating an amazing and supportive environment. Specifically, I thank Wayne Yokoyama for his leadership and guidance throughout my time in the program. I would like to thank the program's

administrators Brian Sullivan, Christy Durbin, Liz Bayer, and Linda Perniciaro for their advice and kind support throughout my training.

I would also like to thank Elliott Margulies, my post-baccalaureate mentor at the National Institutes of Health/National Human Genome Research Institute. He provided my first positive independent research experience and spurred me towards a career as a physician scientist. The computational and bioinformatics skills that I brought to Washington University were all learned under his mentorship.

I would like to thank The Family of Love, my MSTP classmates who entered the program in the fall of 2010. I am continually astounded and inspired by their knowledge, creativity and dedication to science. I am proud to call them my friends and cannot wait to see how they will change the world.

I want to thank my family. I thank my parents for nurturing my inquisitiveness and creativity from a young age. It must have been exhausting. I would not be here without your love and inspiration. I thank my brother for his love, support and friendship over the years. I thank my wife for her love, encouragement and endless patience that made all of this possible. Finally, I would like to thank our son. He is a little bundle of joy and happiness that inspires me every day.

Andrew Lee Young

Washington University in St. Louis

May 2018

ABSTRACT OF THE DISSERTATION

Discovering Rare Hematopoietic Clones Harboring Leukemia-Associated Mutations Using Error-Corrected Sequencing

by

Andrew Lee Young

Doctor of Philosophy in Biology and Biomedical Sciences

Molecular Genetics and Genomics

Washington University in St. Louis, 2018

Associate Professor Todd E. Druley, Chair

Cancer is a heterogeneous group of diseases that currently takes over half a million lives per year in the United States alone. Our understanding of cancer has improved dramatically over the last forty years, beginning with the discovery that cancer is a disease of the genome. Currently, the set of somatic mutations found in malignancy are largely known. The specific somatic mutations driving an individual's disease can be readily assessed at clinical presentation. Additionally, the functional consequences for many of these mutations are known as well as their role in tumorigenesis. Despite this understanding, a cure for cancer remains elusive.

Acute myeloid leukemia (AML) is a particularly deadly example, which currently kills about 10,000 people per year and has a 5-year survival rate of only 25%. While the current outlook for these patients is grim, much is known about the disease, which will fuel future improvements in detection and therapy. Existing research has identified the spectrum of somatic mutations driving most cases of AML and has elucidated the oligoclonal nature of the disease. Following treatment, relapse often arises from a minor clone that was inconspicuous at

presentation, but resistant to treatment. The current gold standard for assessing response to treatment is multiparameter flow cytometry (MPFC), which identifies persistent leukemic cells marked by a patient-specific leukemia-associated immunophenotype. Unfortunately, MPFC is only useful in a subset of patients and not sensitive to the clonal diversity present in many tumors. Conversely, virtually every case of AML is marked by leukemia-specific somatic mutations that theoretically distinguish every leukemic cell from its normal counterparts.

These limitations of MPFC and the general need for improved residual disease detection were early motivations for this thesis work: to develop a sequencing-based modality for rare leukemic-clone detection. Previous efforts to develop a sequencing-based platform for residual disease detection had largely failed because of the intrinsic error rate of next-generation sequencing (NGS) technology, which precludes the detection of leukemic clones less common than 1:20 cells (0.025 variant allele fraction for heterozygous mutations). For comparison, MPFC is sensitive and prognostic to a detection limit of 1:10,000 cells. To address this limitation, we developed methods for targeted error-corrected sequencing that mitigated the effect of sequencing errors. After an extensive development and validation process, we applied this technology to study two fundamental questions in AML and hematopoiesis in general.

First, we applied our error-corrected sequencing methods to study leukemogenesis in therapy-related AML (t-AML). This aggressive form of leukemia arises months to years following treatment with chemotherapy or radiation for a primary malignancy. The prevailing notion was that antecedent therapy introduced somatic mutations in hematopoietic stem and progenitor cells (HSPCs) that directly caused the development of t-AML. We used error-corrected sequencing to demonstrate that leukemogenic *TP53* mutations were present at low frequency months to years before the diagnosis of t-AML and in some cases preceded the initial chemotherapy exposure. These findings redefined the etiology of t-AML. Instead of being introduced by chemotherapy, these *TP53* mutations likely arose stochastically in HSPCs throughout the patient's lifetime and were selected for by cytotoxic therapy, eventually spawning malignancy.

Second, we applied error-corrected sequencing to further our understanding of benign clonal hematopoiesis in healthy individuals over time. Recent work had identified benign hematopoietic clones harboring leukemia-specific somatic mutations in the blood of healthy individuals. The prevalence of this phenomenon increased as a function of age; while rare below 50, clones were detected in up to 10% of individuals by 70 years-old. These findings were made with conventional NGS and, likewise, did not detect rare clonal mutations in fewer than 1:20 cells. We sought to characterize the prevalence, stability and mutation spectrum of benign hematopoietic clones below this threshold. Using our error-corrected sequencing approach, we demonstrated that approximately 95% of disease-free individuals have hematopoietic clones harboring leukemia-associated mutations by 50-60 years of age. We also demonstrated that these clonal mutations were stable over time and originated in long-lived HSPCs.

These findings demonstrate the utility of our error-corrected sequencing platform to identify and characterize previously undetectable leukemia-associated somatic mutations. We applied these techniques to unveiled new insights into clonal HSPC biology and the development of t-AML. Future work will apply this technology as a sequencing-based modality for residual disease detection in pediatric AML. We believe this technology will improve the detection of residual leukemia, identify the step-by-step molecular perturbations driving relapse, inform therapeutic selection, and improve clinical outcomes and survival.

Chapter 1: Introduction

1.1 Cancer is a Genetic Disease

Cancer is the second most common cause of death in the United States, predicted to take over half a million lives in 2016[1]. History has demonstrated that effective treatment of this heterogeneous disease requires a thorough understanding of the molecular alterations that drive each individual's malignancy[2]. Forty years ago it was postulated by Peter Nowell that cancer is an evolutionary process by which normal cells sequentially acquire somatic mutations, experience drift and are selected for by the environment[3]. This theory built upon the groundbreaking discoveries by Janet Rowley, who first characterized the t(8;21)(q22;q22) *RUNX1*/*RUNX1T1* translocation found in 5% of acute myeloid leukemia (AML) cases, the universal Philadelphia chromosome t(9;22)(q34;q11) *BCR/ABL* translocation found in chronic myeloid leukemia (CML), and the canonical t(15;17)(q24.1;q21) *PML*/*RARA* translocation found in acute promyelocytic leukemia[4–7]. Additional contemporary work characterized X-inactivation skewing in tumor samples from female patients, which characterized the monoclonal (CML, Burkitt's lymphoma, polycythemia vera, myelofibrosis) or polyclonal (hereditary neurofibromas) origin of several neoplasms[8–13]. Interestingly, while neurofibromas had a polyclonal origin, malignant transformation into a neurofibrosarcoma arose from a single cell[14]. These findings definitively demonstrated that cancer is a genetic disease, likely originating from a single cell, and founded the field of cancer genetics[15]. These concepts were further bolstered by studies examining clonal evolution and heterogeneity at the chromosomal level via karyotype analysis[16,17]. This was followed by the seminal discovery of the first cancer-causing DNA sequence change—a guanine to thymine substitution in codon 12 of *HRAS*[18,19]. This discovery utilized Maxam-Gilbert sequencing—a newly developed technique for quickly reading the nucleotide sequence of short DNA fragments[20]. This technology and another concurrently developed sequencing technology—Sanger sequencing—were the primary workhorses for cancer sequencing studies over the next 20 years and launched a new era of cancer genomics[21].

1.2 The Human Genome Project, Sequencing Cancer Genomes, A Catalog of Cancer Associated Genes

Shortly after the turn of the 21st century, completing the first reference human genome sequence provided a framework to map the evolutionary process of tumorigenesis at the level of individual nucleotides[22,23]. Focusing on malignancy, early exome-sequencing studies leveraged this reference genome to provide the first characterization of mutations in breast and colorectal cancer[24,25]. These herculean efforts sequenced hundreds of thousands of PCR amplicons to discover thousands of germline and somatic mutations present in the tumor samples. These studies highlighted the necessity of comparing the sequencing results from individually matched tumor and normal samples in order to distinguish somatic mutations from constitutional variants that differed from the reference genome sequence. Through targeted sequencing of candidate gene panels, other large collaborative sequencing studies identified a spectrum of somatic mutations in multiple types of malignancies including glioblastoma, colorectal cancer, adenocarcinoma, renal carcinoma, myeloproliferative neoplasms and pancreatic cancer[26–32]. Together these observations elucidated the vast heterogeneity of somatic mutation burden and substitution types between different tumor types that resulted from the specific environmental exposures, such as UV light in melanoma, carcinogen exposure in lung cancer and DNA damage repair defects[33,34]. During this period, each new discovery expanded the list of specific genetic alterations that caused cancer, estimating that only a few hundred out of the 20,000-25,000 protein-coding genes in the human genome routinely contribute to malignant transformation[35,36]. Today, these lists of curated cancer-associated genes and detected somatic mutations are maintained at the Cancer Gene Census and Catalogue of Somatic Mutations in Cancer (COSMIC), respectively[35,37,38].

The subsequent development of massively parallel sequencing—also known as next-generation sequencing (NGS) or second-generation sequencing—reduced costs by two orders of magnitude and moved nucleic acid sequencing out of the a global network of large production facilities that generated the first human reference genome and into smaller research labs studying the panoply of biological processes and diseases[39–42]. This technology enabled the first matched tumor/normal whole genome sequencing (WGS) study, which characterized the spectrum of somatic single nucleotide substitutions and small insertion/deletion (indel) mutations within a single case of normal karyotype adult *de novo* AML[43]. Surprisingly, this study identified only 10 coding somatic mutations, again, highlighting the importance of a matched normal sample to filter out inherited constitutional variants[43]. A subsequent study of another normal karyotype AML case revealed only 12 coding somatic mutations including the hot spot *IDH1* R132C mutation, which was subsequently identified in several other AML cases[44]. Reanalysis of the first AML case to undergo WGS identified a frameshift mutation in *DNMT3A*—a DNA methyltransferase that catalyzes the *de novo* methylation of cytosine in CpG dinucleotides[45]. Abnormal DNA methylation leading to epigenetic dysregulation was hypothesized to contribute to the development of cancer[46]. This finding led to the characterization of recurrent mutations in *DNMT3A* (present in one-third of normal karyotype AML cases) and hotspot mutations affecting the arginine 882 amino acid (present in almost two thirds of *DNMT3A* mutations in AML)[45]. A similar German study characterized the spectrum and prevalence of *IDH1* and *IDH2* mutations (found in the second AML WGS study) identifying *IDH1* or *IDH2* mutations in 16% of adults with AML[47]. These studies highlight the power of unbiased WGS studies to uncover potent drivers of malignancy and subsequently quantify the prevalence of those mutations and outcomes in affected individuals.

These early studies opened the floodgates for the unbiased detection of somatic mutations in any and all forms of malignancy. Subsequent studies characterized the spectrum of somatic mutations in multiple malignancies including breast cancer, malignant melanoma, and small-cell lung cancer[48–52]. These studies provided the first broad characterization of "cancer genome landscapes," specifically identifying the few genes that harbor somatic mutations in a wide variety of malignancies and the many genes that are less frequently mutated[53]. A comprehensive review from Bert Vogelstein et al. summarized the following observations: 1) the number of non-synonymous mutations per cancer type was highly variable with up to 1000 in colorectal cancer with microsatellite instability, 100-200 in lung cancer and melanoma (due to environmental mutagen exposure), and approximately 10 in liquid tumors (e.g. AML and CML) and pediatric cancers (e.g. glioblastoma, neuroblastoma and medulloblastoma); 2) a typical tumor only contained 2-8 "driver" mutations and the rest were inconsequential passenger mutations that arose throughout the natural history of the cell that founded the malignancy; 3) tumors were almost universally heterogeneous, which would impact response to treatment[53]. Subsequently, the Cancer Genome Atlas' analysis of 3,281 tumors across 12 cancer types and The Broad Institute's analysis of 4,742 tumors across 21 cancer types cemented these concepts, producing a comprehensive catalog of cancer associated genes[54,55]. In AML the genomic landscape of somatic mutations was further refined by leukemia/normal WGS or whole exome sequencing (WES) for 200 cases of AML[56]. The overwhelming amount of data generated by these studies required the development of new tools to distinguish putative driver mutations from the bevy of passenger mutations that were also detected[57]. These discoveries established the foundation necessary to realize the goal of personalized medicine, where each patient's cancer will be genotyped accurately at diagnosis, the cancer-specific pathways and susceptibilities will be identified, and a personalized therapeutic plan will be initialized[58]. However, as our understanding of the cancer genome has progressed, so has our understanding of clonal heterogeneity and evolution, curtailing early hopes for an easily produced, genomically forged cure for cancer.

1.3 Clonal Heterogeneity and Clonal Evolution

With an accurate reference set of cancer associated genes and mutations, the next important step in understanding the genomic basis of malignancy was to describe the heterogeneity within a single individual's malignancy and the clonal evolution of that malignancy over time. Heterogeneity within an individual's tumor was first observed long ago using karyotype analysis[17,59]. However, NGS made it possible to characterize this heterogeneity throughout the genome. One elegant study sequenced multiple sections from a single pancreas inundated with carcinoma along with several distant metastases to clearly show that genomic heterogeneity increased geographically across the primary lesion. Additionally, peritoneal metastases were similar to the primary tumor, and the liver and lung metastases were drastically different from the primary tumor[60]. Another study described likely convergent evolution within a single case of renal cell carcinoma in which three different, geographically separated somatic mutations in *SETD2* were identified in a tumor already missing the other copy of *SETD2* because of a ubiquitous chromosome 3p deletion[61]. These initial studies demonstrated the geographic heterogeneity found within solid malignancies and provide some insights into the mechanisms for metastasizing, treatment resistance and recurrence.

Focusing on liquid tumors, while a rare group of diseases overall, there are several key advantages regarding the study of clonal evolution that are worth noting. For liquid tumors, samples are often serially banked over the disease course of a single individual, samples are relatively free of contaminating normal tissue, phenotypically identical cells are easily sorted for analysis of specific subpopulations of cells, and comparable healthy tissue is easily obtained. These features enable the in-depth study of clonal heterogeneity within an individual's disease. The clonal structure of AML was elegantly described by John Welch et al. in a study utilizing WGS to characterize the somatic mutations present in 12 cases of French-American-British (FAB) classified M3 AML (each containing the canonical PML-RARα translocation) and 12 cases of M1 AML with an unknown initiating lesion[62]. This study presented several interesting findings: 1) most somatic mutations found in AML were benign events that occur during the natural history of the initiating cell before leukemic transformation; 2) the AML samples were almost universally oligoclonal with multiple clones present at diagnosis; and 3) the founding clone in M1 AML frequently had mutations in *DNMT3A*, *IDH1*, *TET2* or *NPM1*[62]. Another important study described the clonal architecture of secondary AML (sAML)—AML that arises from antecedent myelodysplastic syndrome (MDS) in the setting of ineffective hematopoiesis[63]. This study observed that clonal hematopoiesis in the bone marrow was indistinguishable between MDS and sAML, all of the sAML samples were oligoclonal (2-5 clones detected per person), and sAML arose from persistent MDS clones that acquired additional functional mutations. While these were both groundbreaking studies, it is humbling to think that these observations were predicted by Peter Nowell, Janet Rowley and others nearly 40 years earlier. Interestingly, the Welch et al. study demonstrated that AML arose in cells with a somatic mutation burden similar to healthy age-matched hematopoietic stem and progenitor cells (HSPCs), suggesting that AML could arise without an elevated intrinsic mutation rate required for other cell types to undergo malignant transformation[58,62].

The effect of therapy on the clonal characteristics of AML were demonstrated by studies of relapsed AML. The clonal evolution of relapsed AML was described in a WGS study of eight cases of AML at diagnosis and relapse, which found that relapse clones could either arise directly from the primary leukemia (three cases) or arise from a subclone that survived the initial treatment (five cases)[64]. The vast majority of somatic mutations detected were shared between the primary sample and the relapsed sample, again demonstrating that most somatic mutations in AML arose prior to leukemic transformation. Additionally, these cases frequently acquired additional mutations at relapse (even if recurring directly from the primary leukemia). This reiterates the challenge of effectively treating this disease, which necessitates the eradication of the primary tumor and all subclones that could seed relapse. A subsequent study of clonal evolution in *NPM1*-mutated AML observed that at relapse *DNMT3A* mutations co-occurring at diagnosis were almost universally retained, but NPM1 mutations were occasionally lost, suggesting that *DNMT3A* mutations might have occurred earlier in the founding clone than the *NPM1* mutations[65]. Another interesting case report described the clonal evolution of a single *IDH1* R132L-mutated AML, which acquired a canonical spliceosome *SF3B1* K700E mutation at relapse 19 years later[66]. Similar findings were also described in acute lymphoblastic leukemia (ALL), chronic lymphocytic leukemia (CLL), CML and multiple myeloma[67–74]. These are just a handful of the studies that demonstrated the dynamics of clonal complexity in liquid tumors that have relapsed following treatment.

The recent development of single-cell sequencing technology has enabled further study of clonal heterogeneity in malignancy. The earliest forms of this technology were applied to solid malignancies to enable the separation of malignant cells from adjacent non-malignant stroma. An early single-cell sequencing approach utilized whole-genome amplification of flow sorted nuclei and low coverage sequencing to assess clonal diversity at the level of copy-number changes, demonstrating a large phylogenetic separation between a primary breast tumor and hepatic metastases[75]. Another early study used single-cell whole exome sequencing to describe the clonal architecture of clear cell renal carcinoma[76]. Surprisingly, while the normal cells clustered tightly together, the malignant cells did not, suggesting that clear cell renal carcinoma may be more clonally diverse than expected[76]. A concurrent study reported the monoclonal origin of a single case of *JAK2*-negative essential thrombocythemia[77]. While this technology has matured in recent years there are still many associated technical challenges such as single cell isolation, unbiased amplification of genomic DNA and sequence data analysis[78]. The single-cell analysis of liquid tumors is somewhat easier, but still developing. One of the largest single cell studies to date analyzed approximately 800 cells from 6 pediatric ALL patients using targeted resequencing of mutations identified in their bulk tumors[79]. Despite a high level of allelic dropout, the investigators were able to accurately describe the clonal architecture of each individual's disease and the hierarchy of clonal mutations. Surprisingly, they demonstrated that mutations present at similar variant allele fractions (VAFs) often occurred in separate clones of similar size. In a separate study, the single-cell analysis of three individuals with sAML refined the clonal hierarchy predicted by prior bulk sequencing, which enabled the accurate clustering of variants that were outliers in the bulk sequencing analysis[63,80]. Another contemporary study examined the functional consequences of clonal heterogeneity in AML and discovered that most clones circulate in the peripheral blood, subclonal mutations appeared in non-leukemic normal hematopoietic compartments, and engraftment potential in xenograft models varied drastically and unpredictably between subclones[81]. This study also used single-cell sequencing to verify the clonal hierarchy predicted by bulk sequencing. Even when a single participant of this study, AML31, was re-sequenced at 10-fold greater coverage (yielding 11-fold more "platinum" variant calls) in a follow-up study, the clonal hierarchy was still informed by the previous single-cell sequencing results[81,82]. Additionally, this ultra-deep sequencing study demonstrated that clonal diversity was much more complicated than previously thought; standard WGS identifying 3 clones at diagnosis and one at relapse compared to deep sequencing which identified 1 founding clone, 3 subclones in the primary tumor, 1 separate clone enriched from diagnosis to relapse, and at least 1 clone gained in relapse[64,82]. These studies have expanded our understanding of clonal dynamics and heterogeneity in AML. As the technology improves, the vast complexity of AML, and malignancy in general, is slowly being unveiled.

1.4 Pre-Leukemic Hematopoietic Stem Cells

The studies mentioned previously also supported the growing understanding of pre-leukemic HSPCs, which carry several of the activating mutations necessary for leukemic transformation, but still maintain ostensibly normal hematopoiesis. Early evidence from this theory came from the observation of RUNX1-RUNX1T1 (AML-ETO) t(8;21)(q22;q22) translocation in normal-appearing long-term remission samples from individuals treated for AML[83]. They observed expression of the RUNX1-RUNX1T1 translocation in healthy hematopoietic stem cells (HSCs), monocytes and B lymphocytes, suggesting that HSCs with the translocation were capable of self-renewal and differentiation into mature blood cells. In another case, pre-leukemic HSCs were detected in two year-old twins discordant for ALL containing the

ETV6-RUNX1 t(12;21)(p13;q22) translocation[84]. An early application of genomics to this field utilized exome sequencing to identify somatic mutations in six cases of AML (3-19 per person) and then examined the prevalence of these mutations in non-leukemic HSCs[85]. Non-leukemic HSCs were isolated by fluorescence activated cell sorting (FACS) and functionally validated by xenograft transplant into immunodeficient mice. In five out of six cases they identified some, but not all, of the leukemia-associated mutations in these HSCs that preceded the AML clone. Using single-cell colony formation and genotyping, the researchers temporally ordered the sequence of mutation acquisition from non-leukemic HSCs containing different subsets of the leukemia-associated mutations. Follow-up work from the same group examined the clonal evolution of these pre-leukemic HSCs and how they respond to induction chemotherapy[86]. Interestingly, they observed that mutations affecting epigenetic regulation (DNA methylation, histone modification and chromatin looping) occurred early in the development of disease and mutations conferring a proliferative advantage occurred late. Using single-cell genotyping, they elegantly and convincingly demonstrated the multistep process of mutation acquisition within individual HSCs that contain increasing subsets of the somatic mutations present in the leukemic sample. Furthermore, this work demonstrated that pre-leukemic HSCs harboring early driver mutations survive induction chemotherapy and may be an important cause of relapse. A concurrent study made similar observations in cases of *DNMT3A* and *NPM1* mutated AML, where pre-leukemic HSCs with only the *DNMT3A* mutation survived chemotherapy treatment and were capable of multilineage engraftment in mice[87]. More recent studies reported persistent *DNMT3A* R882 and *IDH2* R140Q mutations in long-term remission following treatment for AML[88,89]. These findings supported the theory that the somatic mutations in AML arise through the sequential acquisition of (largely non-functional) mutations in self-renewing HSCs and that pre-leukemic HSCs exist at diagnosis that harbor some, but not all, of the mutations present in the founding AML clone.

1.5 Residual Disease Detection

At the beginning of this thesis work, virtually all of the driver mutations in AML had been uncovered, the clonal structure and temporal evolution of several AML cases had been described, and there was growing evidence that pre-leukemic HSCs harboring a subset of the AML-associated somatic mutations survived chemotherapy and could spawn relapse. The genomic tools and understanding were in place to develop a sequencing-based modality for residual disease detection in AML. Specifically, following induction therapy, could persistent leukemia-associated mutations in peripheral blood or bone marrow samples predict relapse and overall survival? While initially unsuccessful, we made several advancements in the detection of rare hematopoietic clonal mutations that will hopefully be useful in the future realization of this goal.

The current gold standard for assessing residual disease following treatment for AML is multiparameter flow cytometry (MPFC)[90]. Leukemic cells are identified by a leukemia-associated immunophenotype (LAIP) that is not present on normal hematopoietic cells. These differences can manifest as different expression levels of normal cell surface markers, abnormalities in timing of marker expression given the normal differentiation program or the co-occurrence of markers not normally present on the same cell[90,91]. In pediatric AML, detecting any residual disease by MPFC (using a different-from-normal approach) at the end of induction (1 or 2) or end of therapy was associated with an increased risk of relapse and shorter relapse free survival[92]. In a separate study of pediatric AML, detecting residual disease above 1 leukemic cell in 1,000 mononuclear bone-marrow cells by MPFC (using a LAIP approach) was also associated with an increased risk of relapse and shorter relapse free survival[93]. These results extended to young adults (<60 years old) and older adults (>=60 years old) treated for AML[94,95]. While most studies simply reported presence or absence of an LAIP during assessment, one study suggested that a detection cutoff of 0.15% maximized the receiver operator characteristics for predicting relapse[96]. However, other groups advocated for a more sensitive cutoff for positive residual disease of 0.035% residual leukemic cells in the bone marrow[97]. This variability between study designs reflects a broader lack of standardization in the field that hinders large-scale application and prevents comparison of results from different studies[98,99]. Another alternative method for residual disease detection is quantitative PCR (qPCR), which targets leukemia-associated translocations. This method is sensitive to detect residual leukemic cells present at 1:10,000-1:1,000,000 cells, about two orders of magnitude lower that MPFC[100]. However, it is only suitable in the 15-60% of AML patients with a canonical translocation or a suitable NPM1 frameshift mutation for which PCR primers have already been designed[98,101]. Despite these limitations, both methods have been used successfully to stratify patient outcomes based on detecting residual leukemic cells following treatment.

Even with these findings, there was mounting evidence that the LAIP identified at diagnosis frequently changed by relapse (up to 91% of AML cases)[102]. Additionally, heterogeneity in the primary AML tumor made the detection of residual or relapsed disease originating from previously uncharacterized subclones more challenging[103]. In B-ALL this was an especially dire problem as treatment often directly targeted CD19, the antigen primarily used for detection, such that relapse clones often lacked the CD19 antigen[104]. These are phenotypic manifestations of the underlying genomic clonal evolution and selection richly characterized by the aforementioned sequencing studies that nonetheless limits the efficacy of MPFC for detecting recurrent disease. Conversely, virtually all cases of AML have somatic mutations that uniquely mark the primary tumor, subclones or the pre-leukemic clone that will seed relapse. In broad strokes, sequencing could determine the spectrum of somatic mutations in the primary tumor and predominant subclones at diagnosis to inform targeted treatment selection and sequencing-based residual disease detection. While useful for tracking, knowing the spectrum of somatic mutations at diagnosis would not be essential for residual disease detection. By querying all of the recurrently mutated genes in AML, the assay would detect relapsing clones with or without the somatic mutations identified at diagnosis. Early applications of NGS for residual disease detection focused on simply detecting indels in recurrently mutated loci that were easy to capture by PCR amplification for sequencing. One report demonstrated that FLT3 internal tandem duplications (15-300 bp long) could be reliably captured from genomic DNA, sequenced with short paired-end NGS (101 bp paired-end reads), and aligned/assembled using a combination of bioinformatics tools to identify the samples with the mutation[105]. Subsequent studies could also assess somatic single nucleotide variants for residual disease detection. One early example in T-lineage acute lymphoblastic leukemia/lymphoma demonstrated that sequencing the T-cell receptor—which is a unique clonal marker for disease—at post-treatment day 29 was much more sensitive than MPFC for detecting residual disease[106]. Similar results were observed by the same group when sequencing the immunoglobulin heavy chain locus in cases of B lymphoblastic leukemia[107]. Interestingly, in both studies, MPFC failed to detect residual disease in several cases where sequencing identified residual disease in greater than 1:1,000 cells, well above the usual limit of detection for MPFC[106,107]. In AML, residual disease was reliably detected in studies specifically sequencing NPM1, RUNX1, and FLT3-ITD and often detected in cases originally deemed negative by MPFC[108–110]. In the RUNX1 study, the frequency of detected residual

RUNX1 mutations was also prognostic of event-free survival and overall survival[109]. These proof of principle studies clearly demonstrated the feasibility of sequencing-based residual disease detection. However, each of these genes were only mutated in a subset of AML cases and broad application of NGS as a residual disease detection modality required expanding sequencing to encompass the spectrum of somatic mutations in AML.

Subsequent studies have expanded somatic mutation detection at remission using WGS or WES, attempting to predict relapse risk by assessing clonal expansion after therapy. A study of 50 patients sequenced at remission with enhanced exome sequencing (exome sequencing supplemented with capture reagents for AML-specific genes) or targeted capture of diagnosis-specific variants determined that mutation persistence was associated with an increased risk of relapse, shorter relapse-free survival and reduced overall survival[111]. Interestingly, this study characterized a mutation as cleared if it was not detected above 0.025 VAF, the limit of detection for NGS, even when MPFC has demonstrated prognostic value in detecting residual AML at 1:1,000 cells or less. Even in individuals who cleared all of their mutations at remission, median event free survival was still only 17.9 months. This suggested that a lower limit of detection would perhaps improve risk stratification by identifying rare persistent clones in individuals destined to relapse. In follow-up study, deeper analysis of 15 of cases identified five cases where clonal expansion occurred following therapy, but harbored different somatic mutations than the diagnostic AML sample[112]. These "rising clones" appeared to expand following therapy, were non-leukemic and did not appear related to the founding AML clone. These findings highlighted two additional challenges facing NGS as a platform for residual disease detection, namely, residual disease with prognostic value was likely present below the 0.025 VAF threshold of detection for NGS and clonal expansion occurred in relapse-free individuals.

Had these findings been known at the outset of this thesis work, the goals and design may have differed. However, at the genesis of this project the limit of detection for NGS appeared to be the primary constraint on sequencing-based residual disease detection. Specifically, the error-rate of NGS precluded the detection of somatic mutations below approximately 0.02 VAF[113]. Based on the literature from residual disease detection with MPFC and qPCR we knew that there was prognostic information in detecting as few as 1:1,000-1:10,000 residual leukemic cells[114–116]. Additionally, as described previously, individuals frequently relapsed when the detection cutoff was 0.025 VAF or 1:20 cells for heterozygous mutations[111]. Fortunately, several tools had been recently developed to mitigate the effect of sequencing errors enabling the reliable detection of variants as rare as 0.0001 VAF[117–125]. In general, these methods capitalized on the same experimental trick, tag each individual DNA molecule with a unique molecular identifier (UMI), sequence each tagged molecule multiple times, use the UMI to identify sequence reads originating from the same molecule and correct the sequencing errors. These tools have been applied to study a variety of biological processes including HIV virus diversity[124,126], early detection of ovarian and endometrial cancers[127], age-associated somatic mutations in mitochondria[128], transcriptome analysis[125,129], and resequencing of tumor samples for hotspot mutation identification[130,131]. Our goal was to adapt these techniques to determine whether a more sensitive sequencing approach that could identify rare leukemia-associated somatic mutations would improve residual disease detection and outcome prognostication.

We created a novel platform for amplicon-based error-corrected sequencing (ECS) that enable the reliable detection of rare clonal somatic mutations. Our first application of this technique was through a collaboration with Daniel C. Link and Terrence N. Wong in the Department of Medicine at Washington University School of Medicine in St. Louis, who were studying the role of TP53 mutations in therapy-related AML (t-AML) and therapy-related myelodysplastic syndrome (t-MDS). These diseases arise following treatment with chemotherapy, radiation therapy and/or immunotherapy; are marked by poor outcomes; and often do not respond to treatment[132]. At the time, the prevailing theory was that DNA damage introduced by antecedent cytotoxic therapy introduced the somatic mutations that drove later disease[133]. Surprisingly, tumor/normal sequencing of several t-AML/t-MDS cases did not identify an increased number of somatic mutations compared to *de novo* AML. However, they did observe a different spectrum of mutations with lower rates of *DNMT3A*, *NPM1* and *FLT3* mutations and increased rates of *TP53* and ABC transporter mutations. They were interested in determining when the leukemogenic *TP53* mutations arose in the intervening years between the cytotoxic therapy exposure and development of t-AML/t-MDS. Shockingly, using our targeted ECS approach, we demonstrated that leukemogenic *TP53* mutations were present at very low frequencies years before the development of t-AML/t-MDS and in two cases before chemotherapy exposure (Chapter 3)[134]. We subsequently used this approach to demonstrate that other non-*TP53* mutations were present months to years before the development of t-AML/t-MDS[135]. These paradigm-shifting findings suggested that the leukemia-associated mutations were not introduced by the chemotherapy, but were already present in the individuals and were selected for by the treatment. Additionally, we demonstrated that ECS could reliably detect rare clonal mutations below the error-rate of NGS. These findings also suggested that the detection of rare clonal leukemia-associated mutations in healthy individuals could predict who would later develop hematological malignancy.

1.6 Clonal Hematopoiesis in Disease-Free Individuals

The concept of aberrant clonal expansion preceding and predicting leukemic transformation has grown in recent decades. Early evidence of benign clonal expansion in the hematopoietic compartment was demonstrated using X-inactivation studies in healthy women[136,137]. They observed that X-inactivation skewing increased as a function of age and presciently suggested stem cell exhaustion or clonal hematopoiesis as likely culprits. Later it was shown that X-inactivation skewing predominantly occurred in the myeloid compartment[138]. Subsequently, somatic mutations in *TET2* were identified in individuals with hematopoietic X-inactivation skewing, establishing a genetic cause for this phenomenon[139]. These findings were quickly bolstered by large scale microarray and sequencing studies of healthy blood samples. Somatic mosaicism of large chromosomal anomalies (duplications, deletions and uniparental disomy) was detected using microarray data from blood samples in <0.5% of individuals <=50 years old and 2-3% of individuals >50 years old, which was associated with a 10-fold increased risk of developing hematological malignancy[140]. Similarly, a concurrent study observed clonal mosaicism in 2% of individuals over 50 years-old and 20% of individuals who later developed myeloid or lymphocytic leukemia[141]. While these were examples of clinically silent clonal expansion in healthy individuals, there are also well described benign conditions that precede malignancy. Monoclonal gammopathy of undetermined significance (MGUS) is a benign condition occurring in up to 2% of individuals 50 years-old or older and progresses to multiple myeloma or a related disease at a rate of 1% per year[142]. Similarly, monoclonal B-cell lymphocytosis, another benign condition, progresses to CLL at a rate of 1.1% per year[143]. Recent genome sequencing efforts have sought to determine the spectrum of somatic mutations driving clonal hematopoiesis and understand the additional steps required for transformation into fulminant malignancy. One study examined the normal blood samples analyzed as control tissue from 2,728 participants in The Cancer Genome Atlas project and identified 77 clonal blood-specific somatic mutations in 58 individuals[144]. These clones frequently harbored mutations in leukemia-associated genes such as *DNMT3A*, *TET2* and *JAK2* and were more prevalent in older individuals. Two subsequent reports characterized clonal hematopoiesis in two large (>12,000 person) cohorts unselected for hematological malignancy and observed that clonal hematopoiesis increased with age (rare at 50 years-old and present in approximately 10% of 70 year-olds) and frequently harbored mutations in *DNMT3A*[145,146]. Interestingly, they only observed the canonical leukemia-associated *DNMT3A* R882H mutation in a sixth of clones with a *DNMT3A* mutation compared to two thirds of AML cases with a *DNMT3A* mutation[45,145]. An additional report studying hot spots recurrently mutated in AML demonstrated that clonal hematopoiesis harboring *DNMT3A* R882H/R882C and *JAK2* V617F mutations arose in younger-aged participants and mutations affecting spliceosome genes arose in older participants[147]. In an interesting case report, WGS of a single 115 year-old woman's blood suggested that the majority of her hematopoietic compartment originated from two clonally related HSCs, as they identified two clusters of somatic mutations at 0.22 and 0.32 VAF[148]. Even the Welch et al. report describing clonal evolution of AML observed leukemia associated mutations in the blood of health individuals, which became more prevalent with age[62]. This type of clinically silent benign clonal expansion has been termed clonal hematopoiesis of indeterminate potential (CHIP)[149].

Despite these provocative findings, all of these studies could only detect common hematopoietic clones due to the error-rate of NGS. We sought to characterize the prevalence and spectrum of mutations below the detection limit of NGS and establish whether these mutations arose in long-lived HSPCs or more terminally differentiated cells. Fortunately, in our ill-fated efforts to create a platform for sequencing-based residual disease detection, we had created the perfect tools to answer this question. Through collaboration with the Nurses' Health Study at Brigham and Women's Hospital and Harvard University, we were able to study serially banked blood samples from healthy women. Characterizing these samples with targeted error-corrected sequencing, we discovered that clonal hematopoiesis is a ubiquitous finding by middle age and revealed new insights into the biology or normal and leukemogenic clonal hematopoiesis (Chapter 4)[150].

Together this body of work demonstrates the inherent complexity of hematopoiesis and leukemogenesis. Outlined in these chapters are the steps we took to develop our platform for error-corrected sequencing (Chapter 2) and the discoveries made with this technology (Chapters 3&4). This process has been an exciting and humbling experience. It is fascinating and terrifying to think that everybody has a hematopoietic compartment chock-full of expanding clones harboring leukemogenic mutations by middle age. Yet, AML is such a rare disease that virtually all of these clones must benignly co-exist with their host. Witnessing this phenomenon makes me optimistic for the future, when we will have to tools and knowledge to accurately predict which clonal mutations, groups of mutations or epigenetic signatures are harbingers of disease and which mark stable benign clonal hematopoiesis.

Chapter 2: Creating a Platform for Targeted Error-Corrected Sequencing

2.1 Introduction

Our initial goal was to develop a sequencing-based modality for residual disease detection in cases of AML. Specifically, we sought to replace MPFC—the gold standard for residual disease detection—because it was only useful in a subset of individuals and insensitive to clonal diversity. In contrast, a sequencing-based approach would be applicable in virtually every case of AML and sensitive to the clonal diversity present within an individual's tumor. This platform would target some or all of the leukemia-specific somatic mutations present in persistent leukemic and pre-leukemic clones that survived therapy and could initiate relapse. Additionally, founder mutations—the initiating lesions acquired early during leukemogenesis—would tag all leukemia-specific clones and a subset of pre-leukemic clones that could initiate relapse.

We initially viewed the sequencing error rate of NGS, which precluded detection of SNVs rarer than 0.02 VAF, as the predominant limitation of sequencing-based residual disease detection[113]. Conversely, MPFC, when applicable, provided prognostic information to a detection limit of 1 leukemic cell in 1,000 total cells[93]. At the outset of this project, one NGS-based residual disease detection study had achieved a limit of detection similar to MPFC by targeting indel events in FLT3 and NPM1[110]. This was possible because indel errors were rarely made by the NGS platform. Unfortunately, SNVs occur approximately 10-times more frequently than indels in AML[56]. Likewise, the sensitive detection of SNVs was essential to capture the spectrum of somatic mutations in leukemic clones. To address this limitation, we focused on improving the limit of detection for accurate SNV calling.

Fortunately, when we began this project, two papers had been recently published that circumvented the error-rate of NGS using unique-molecular identifiers (UMIs)[117,120]. One method termed Safe-SeqS tagged each strand of individual DNA fragments with a unique 12- or 14-base random oligonucleotide index (UMI) during library preparation[120]. When sequenced, multiple sequence reads containing the same UMI originated from the same original single-stranded DNA molecule. Computationally, these reads were compared to each other and sequencing errors present in a single read were corrected by comparison to the other reads from the same original tagged molecule. This mitigated the effect of sequencing errors and enabled the detection of mutations as rare as 0.001 VAF for most classes of substitutions. Specifically, the limit of detection for G to T and C to T mutations was still closer to 0.01 VAF due. The source of these errors are described below. Despite this limitation, we developed our targeted error-corrected sequencing (ECS) approach based on these techniques.

Another method termed Duplex Sequencing enabled a lower limit of detection than Safe-SeqS by tagging both strands of each DNA fragment with complementary UMIs[117]. While every double-stranded DNA (dsDNA) molecule in the library was tagged with a different UMI, each complementary strand comprising a single dsDNA molecule was tagged with a complementary UMI. This enabled linking of sequenced reads from each strand of the original dsDNA molecule, enabling the correction of strand-specific artifacts and PCR errors introduced early during library preparation. Interestingly, by tagging both complementary DNA strands, they were able to demonstrate that most errors present in libraries only UMI-tagging single-strands of DNA (e.g. Safe-SeqS) were due to DNA damage. Specifically, they observed guanine oxidation to 8-oxoguanine (G to T mutations) and cytosine deamination to uracil (C to T mutations), which were both well characterized mechanisms of DNA degradation[151,152]. When DNA was treated with an oxidizing agent, they observed an increase in G to T mutations in single-strand consensus sequences, but not in duplex consensus sequences[117]. Unfortunately, this improvement in error-correction required four times more sequencing than the Safe-SeqS methods. While useful in in small model systems (e.g. the mitochondrial genome), Duplex Sequencing was not suited to target large regions of the human genome. Conversely, the Safe-SeqS approach could be adapted to enable the identification of clonal somatic mutations in all of the recurrently mutated genes in AML.

Having settled on a framework for generating ECS libraries, we next developed methods to target specific loci in the genome. Previous applications of ECS were in small systems (mitochondrial DNA, plasmids) that could be sequenced entirely. To apply these techniques for residual disease detection in patients with AML, we needed to target specific loci in the genome. Initially, we attempted this using liquid-phase hybridization capture with biotinylated oligonucleotide baits, which was the primary capture strategy for exome sequencing[153]. This method enabled sampling from diverse, randomly sheared genomic DNA libraries and avoided many of the issues found with PCR amplification (jackpotting, allelic skewing). However, we were totally unsuccessful. The capture yield for individual targets was exceptionally low and the off-target rate was unacceptably high. Granted we were trying to capture a handful of loci (hundreds of bp) from the entirety of the genome (3 billion bp), so even a 1000-fold enrichment by capture still resulted in an unusable library.

We next moved on to PCR-based capture. This method could reliably capture individual exons from the genome, variant calls were quantitative, and PCR artifacts were minimal. This technique enabled targeted-ECS for our collaboration with Dan Link and Terrence Wong (Chapter 3), studying *TP53* mutations in therapy-related AML (t-AML). We were able to identify variants identified at diagnosis for t-AML at low frequency in samples banked years prior to diagnosis. However, with this method, we still noticed a high rate of G to T substitutions, indicative of guanine oxidation to 8-oxoguanine in the primary sample. Initially, this precluded the detection rare clonal G to T substitutions. However, we later developed a binomial statistical framework to model position specific error profiles. This enabled us to identify likely clonal G to T (and C to T) mutations above the background error rate due to DNA damage artifacts. For other substitutions, this model us to reliably identify variants as rare as 0.0001 VAF.

The effect of DNA damage was also observed in artefactual false positives observed in our validation experiments using droplet digital PCR (ddPCR)[154]. With this technique, DNA was partitioned into microfluidic droplets that were genotyped individually. Genotyping was accomplished by amplification with primers spanning the variant of interest and querying with a variant-specific TaqMan probe. Droplets receiving a copy of the wild-type allele would fluoresce with the wild-type probe and droplets receiving a copy of the mutant allele would fluoresce with the mutant probe. Occasionally, we observed droplets that received two different alleles from separate genomic DNA molecules. These double-positive droplets occurred at predicable, low rates. However, when the wild-type allele was guanine and the mutant allele was thymine (guanine oxidation), we observed many more double-positive droplets than expected by chance. The same was true for cytosine to thymine substitutions (cytosine deamination). Due to these known sources of artifacts, we modified Bio-Rad's approach to calculating VAF to improve the accuracy of rare variant quantification.

The final set of methods that we developed were for multiplex capture with ECS. PCR-based capture, described previously, enabled targeting at a handful of loci within a single sample. This was useful when resequencing samples with known mutations. However, a different approach was required to create a single reagent that could detect rare leukemia-associated clonal mutations in nearly all individual with AML. A broad panel was necessary as every case of AML contains a different spectrum of somatic mutations and the clonal mutations that drive relapse are often undetectable at diagnosis[56,64]. We combined the Illumina TruSight Myeloid panel with our ECS library preparation to enable rare variant detection at multiple loci recurrently mutated in AML. With this reagent, we targeted a tractable subset of the genome (141 kb covering 54 genes) that covered recurrently mutated loci in AML. By combining these two protocol, we were able to reliably detect leukemia-associated variants as rare as 0.0003 VAF. Unfortunately, our initial application of this technology as a modality for residual disease detection ended in failure. We received a bad lot of reagents from Illumina that introduced an unacceptably number of PCR artifacts during library preparation. While initially unsuccessful, future planned studies in the lab will compare targeted-ECS to MPFC for residual disease detection in AML.

While applied here to the study of AML and clonal hematopoiesis, these tools are broadly applicable for rare variant detect with any set of genes and in any tissue type. Presented here are some of the key experimental successes and failures conducted to develop our targeted error-corrected sequencing protocol.

2.2 Protocol Development

2.2.1 General principles for ECS

Our adaptation of error-corrected sequencing built upon the established Safe-SeqS design[120]. In general, DNA molecules were tagged with unique molecular identifiers (UMIs), amplified by PCR and sequenced to yield multiple sequenced reads per original tagged molecule. Sequencing errors present at one position in one of the sequenced reads would be identified by observing the correct nucleotide call in the other reads originating from the same tagged molecule (Figure 2.1). Practically, random 16-bp UMIs were introduced into the standard Illumina Y-shaped adapters (Figure 2.2). These adapters were ligated to genomic DNA fragments during library preparation such that each molecule was tagged with a different UMI. Multiple identical copies of each tagged molecule were created by PCR amplification. This amplified library was then submitted for sequencing. After sequencing, the reads originating from the same original molecule were grouped together based on their UMI sequence. Initially, we allowed up to two mismatches per UMI to allow for sequencing errors in the 16-bp UMI sequence. However, further analysis demonstrated that the UMI sequences were not as random as advertised by Integrated DNA Technologies, our oligonucleotide vendor. Allowing a UMI mismatch correction frequently grouped together reads from two separate uniquely tagged molecules that differed in UMI sequence by only a single nucleotide (Figure 2.3). Based on these observations, we did not allow mismatches in the UMI sequence when generating read families—groups of reads originating from the same tagged molecule.

Several parameters of the ECS library preparation protocol were determined experimentally. One critical parameter was library concentration after ligation with the UMI-tagged adapters. In the t-AML study, library concentration was quantified using quantitative PCR (qPCR). Later, experiments used ddPCR to more accurately quantify library concentration. Regardless, successful rare variant detection with ECS absolutely required that multiple copies of each original UMI-tagged molecule were sequenced concurrently. Likewise, the number of UMI-tagged molecules had to be accurately restricted before a final amplification step, so amplicons would be sequenced for each tagged molecule. If the library was too dilute, sequencing bandwidth would be wasted with too many reads covering a handful of UMIs. Conversely, if the library was too concentrated then each UMI would only be covered by a single read at most, preventing error-correction by multiple sequence alignment. This calculation was based on the sequencing bandwidth of the machine run planned, the number of samples multiplexed on the run, the number of targets per samples, the number of sequenced reads per UMI, and the desired limit of detection. After several experiments, we decided that 10x coverage per UMI resulted in the optimal balance between sequencing coverage per UMI and number of UMI-tagged molecules per library. The remaining parameters were decided for each unique experiment. For example, to target a single-loci in a single sample with an Illumina MiSeq Nano (1M reads) run, the library must contain 100,000 UMI-tagged molecules that would be amplified for sequencing. If the calculation was incorrect and 500,000 molecules were selected instead, the average coverage per UMI would only be 2x and the results would be uninterpretable. Given the precision required at the steps, extra care was taken to accurately quantify library concentration before dilution.

The second set of parameters that factored into experiment planning were the number of genome equivalents entering the reaction, the expected yield of capture and the desired limit of detection. In the previous example, to query 100,000 genome equivalents, the reaction needed to start with 330 ng of human genomic DNA (3.3 pg per haploid human genome). Capture efficiencies varied dramatically depending on the targeting protocol. For the clonal hematopoiesis studies (Chapter 4), the capture reagent had a capture efficiency of approximately 5%. Based on that estimate, we used 250-500 ng of input DNA (75,000-150,000 genome equivalents) to ensure a limit of detection of at least 1:1,000 variant allele fraction (VAF). Reducing the amount of input DNA would worsen the limit of detection irrespective of sequencing depth.

Several other bioinformatics parameters we empirically derived. We established a minimum read family size of five reads. Other groups had generated read families from two or three reads sharing the same UMI[120]. However, we found a lower false positive rate when using a higher minimum read family size. Additionally, we required that 90% of the reads present at a given position within a read family call the same nucleotide to call a consensus nucleotide, otherwise we reported an N at that position. Finally, consensus sequences were not reported if more that 10% of the positions were called as an N. We found that modifying these last two parameters did not significantly affect the specificity of the variant calls. While the capture methods differed significantly over the projects presented here, the fundamental parameters for read family generation and variant calling remained constant.

2.2.1 Co-opting liquid-phase hybridization capture for ECS

The first attempt at selecting genomic loci for targeted ECS was with liquid-phase hybridization capture (Figure 2.4). With this method, randomly sheared genomic DNA was hybridized to biotinylated oligonucleotide baits (complementary to a region of interest). Genomic DNA fragments hybridizing to the biotinylated baits were captured with streptavidin coated magnetic beads. These steps were identical to standard exome capture protocols. The adapter sequences were designed to be compatible with the Illumina sequencing chemistry (Figure 2.2). While standard exome sequencing often targeted 30-70 Mb or approximately 1% of the human genome, we sought to target individual regions that were approximately 300 bp long or 0.00001% of the human genome. Initially, this didn’t seem like such a foolish idea, but hindsight tells a different story of youthful optimism. Using 90-bp long oligonucleotide baits (Table 2.1), we attempted to recreate the hybridization conditions used with standard exome capture. The primary experimental condition that we varied was the molar ratio of bait molecules to target loci in the genomic DNA. We initially started with a bait:target molar ratio of 100:1, but did not pull down any DNA. Subsequently, we tested molar ratios of bait:target from 10^5:1 to 10^9:1 using 8 baits simultaneously for a single locus (Table 2.2). The optimal molar ratio was 10^7:1, where we observed approximately 10% of the reads were on target. However, these reads were generated from only a couple hundred molecules and reported coverage was inflated by sequencing PCR duplicates of these few molecules. Alone, the low on target rate would have been acceptable. However, few of the genome equivalents present in the hybridization reaction were successfully captured for sequencing. We started these experiments with 500 ng of genomic DNA, which was approximately 150,000 genome equivalents. After accounting for PCR duplicates, we recovered at most 500 distinct molecules from the targeted reaction or 0.3% of target genome equivalents present. As outlined previously, this low capture efficiency precluded the detection of rare clonal mutations.

We attempted to improve capture efficiency by reversing the adapter ligation and capture steps. We believed the adapter sequences, especially the long UMI index, would interfere with the hybridization stoichiometry or lead to daisy-chaining. Daisy-chaining occurred when one molecule was hybridized correctly to a biotinylated bait, but the adapter sequence for that molecule was hybridized to another molecule that would be unintentionally pulled down in the capture step. Theoretically, by reversing the steps, only fragmented genomic DNA would be captured by the baits without the risk of daisy-chaining. However, this was unsuccessful because too little DNA was captured for the subsequent library preparation steps. To troubleshoot this problem, we added synthesized amplicons during the ligation step to improve the ligation stoichiometry. These amplicons all contained uracil instead of thymine and we degraded after ligation with uracil DNA glycosylase (Figure 2.5). Unfortunately, this was also unsuccessful as the synthetic amplicons that escaped cleavage by UDG were more than enough to overwhelm the few captured genomic DNA molecules.

Given this limitation of liquid-phase hybridization capture, we were forced to abandon this as a method for capturing genomic loci. For sequencing-based residual disease detection, we required reliable identification of genomic variants from one leukemic cell out of 1,000 normal cells. Given the limitations observed, the amount of input genomic DNA necessary to achieve this limit of detection was unreasonably high. Likewise, we redirected our efforts to different methods for capturing genomic loci.

2.2.3 PCR amplification-based targeting of genomic loci

We next sought to capture individual loci from genomic DNA using PCR primers spanning the region of interest. This was a much easier method than the original liquid phase-hybridization capture, but there were several potential drawbacks. First, errors introduced during the early steps of PCR amplification, before UMI-tagging, would be indistinguishable from true rare variants. Second, PCR jackpotting could skew allelic ratios, especially for the low frequency variants we intended to detect[155]. Jackpotting occurred when a low number of template molecules were not uniformly amplified during PCR. The most extreme example is the loss of a constitutional heterozygous polymorphism due to selective amplification of one allele, which frequently occurs during whole genome amplification. To circumvent this limitation, we split template samples into eight separate reactions for amplification, mixed the samples back together after purification and then repeated the process. This allowed us to dilute out the effect of jackpotting. Additionally, we used the Q5 High-Fidelity DNA Polymerase (NEB) to minimize the number of error introduced during PCR amplification.

The general workflow for amplicon-targeted ECS was straightforward (Figure 2.6). Primers were designed to target regions of interest (e.g. exons 4-7 of *TP53*). The resulting amplicons were individually tagged with UMI-containing adapter sequences by ligation. The libraries were diluted to restrict the number of molecules seeding the sequencing run, amplified by PCR and then submitted for sequencing. The resulting sequenced reads were grouped into read families based on their UMI, sequencing errors were identified, and an error-corrected consensus sequence was generated.

2.2.4 Multiplex hybridization-extension-ligation for leukemia-associated target capture

PCR amplification-based capture enabled the analysis of up to about a dozen loci simultaneously. The primary limitation to targeting more loci with PCR was the amount of starting template material required. To detect rare clonal mutation (at 0.001 VAF) approximately 500 ng of genomic DNA was required per experiment. Other methods could target up to 50 loci simultaneously by introducing a pre-amplification step with all of the primers followed by individual amplification with each primer pair[156]. However, for our study of clonal hematopoiesis in healthy individuals (Chapter 4), we needed to increase the number of target loci by yet another order of magnitude to query all of the recurrently mutated genes in AML. Our ultimate goal was to develop a broadly applicable platform for residual disease assessment in AML. This broad utility hinged on being able to detect leukemia-associated mutations in virtual every unique case of AML.

To accomplish this goal, we sought to adapt our error-corrected sequencing indexing strategy to a pre-existing capture reagent that had been already balanced and benchmarked. Ultimately, we selected the Illumina TruSight Myeloid Panel, which targeted 141 kb of the human genome with 568 amplicons in 54 genes. The method for capture was similar to molecular inversion probes, where primers were designed to span a given locus of interest and anneal to the same strand of DNA (Figure 2.7a-c)[130]. Beginning from the upstream primer, a single step of extension would fill in the gap between the primers (Figure 2.7b). Next a ligation step would connect the filled in strand with the annealed downstream primer (Figure 2.7c).

Similar to the Y-shaped adapters described previously, these capture primers each had asymmetrical, non-complementary tails that were compatible with the Illumina sequencing chemistry. While the sequences of the Y-shaped adapters were reported by Illumina, they did not disclose the sequences flanking each primer pair in the TruSight Myeloid panel. When we tried to use our existing UMI adapter chemistry with the TruSight Myeloid panel, the library preparation failed. We suspected that the adapter sequences had changed, but Illumina could not confirm our suspicions. Since we could not use the standard TruSight Myeloid adapters for ECS, we needed design our own compatible UMI-adapters. We determined the Illumina adapter design by Sanger sequencing a library generated with the standard TruSight Myeloid protocol and subsequently designed our own UMI-adapters (Figure 2.8). With this information, we continued adapting ECS to the TruSight panel (Figure 2.7d-g). We used PCR to add in our custom designed adapter sequences that contained a fixed-index for sample multiplexing and a UMI-index to enable ECS. Once these molecules were generated, the same process of accurate quantification, dilution, amplification, sequencing and analysis were conducted to call rare variants. This platform was used for novel rare variant detection in our study of clonal hematopoiesis in healthy individuals (Chapter 4).

We conducted several experiments to assess library quality and technical reproducibility. We determined that the coverage per target was highly concordant between the standard Illumina

TruSight Myeloid protocol and our modified protocol to introduce adapters with UMIs (Figure 2.9). We also observed that coverage per target was highly correlated between replicate libraries produced from the same samples (Figure 2.10). We also observed that generating the error-corrected consensus sequence did not bias coverage per amplicon. While some amplicons were covered poorly in these experiments, they were covered poorly with the standard protocol as well. These findings demonstrated that the ECS protocol integrated into the TruSight Myeloid protocol without disrupting the capture efficiency.

2.2.5 Binomial error modelling

The predominant limitation of our method for error-corrected sequencing was that only single strands of DNA were UMI-tagged. Likewise, we could not correct substitutions originating from DNA degradation of the original template or early PCR errors. The errors that we observed were usually G to T substitutions, due to guanine oxidation to 8-oxo guanine, and C to T substitutions, due to cytosine deamination to uracil. We modeled the prevalence of these artifacts to observe their position-specific distribution to potentially distinguish rare variants from artifacts introduced by DNA damage.

Surprisingly, we observed a strong position specific effect on the substitution error rate. While the G to T substitution rate varied widely across a specific region, the error rate at a single position was consistent between samples. Based on this observation we modeled the position specific error profile as a binomial process, analogous to a coin-flipping experiment. The variant calls made by the error-corrected consensus sequences (ECCSs) were treated as a series of coin flips. If we observed 1000x coverage at a given position, that corresponded to 1000 coin flips. At each independent position, we counted ECCSs identifying the wild-type or variant nucleotide (heads vs tails). By sequencing multiple individuals and replicates, we could build a robust model of the substitution error rate at every position captured by the panel. Thus, for each variant observed in a sample, we could estimate the probability that variant was artefactual given the error profile at that position.

We also accounted for multiple hypothesis testing to reduce the rate of false positive calls. Since the capture panel targeted 141 kb of sequence and there were three substitution types per position, there were almost half a million hypothesis tests per sample. We employed a stringent Bonferroni correction on our variant calls that corrected for the number of samples, replicates, bases covered and substitution classes in the entire cohort. As a result, our variant calls were highly specific. While the sensitive of the assay was never directly assessed, this statistical framework likely grossly underreported the number of true rare variants present. Conversely, every substitution called by ECS, that we subsequently validated with ddPCR, was observed at nearly identical VAFs (Chapters 3 and 4).

2.2.6 Validation with ddPCR

Given the extremely low VAFs identified by ECS, we needed an equally sensitive method for validation. While expensive to implement, we use the Bio-Rad QX200 ddPCR platform to validate our findings. This technology combined allele-specific TaqMan probes with microfluidic partitioning to permit extremely sensitive and specific variant quantification. Similar to ECS, the primary constraint on the limit of detection was the number of genome equivalents used in the assay. Experimentally, primers and probes were designed to target a specific single nucleotide variant or small indel. Genomic DNA was partitioned into microfluidic droplets such that on average one or fewer genome equivalents of the region of interest was present in a single droplet. The region of interest was amplified by PCR and subsequently assayed with the allele-specific TaqMan probes. The droplets were then analyzed to assess fluorescence intensity for the wild-type or variant probe. Variant calls as rare as 0.0001 VAF were easily validated on this platform. For negative controls, we selected samples from the same cohort where the variant of interest was not observed. Surprisingly, we frequently observed zero positive mutant droplets in our negative control experiments, which often had up to 500,000 total droplets.

Interestingly, the platform was so sensitive that we observed the effect of DNA degradation on our genomic DNA samples. These were predominantly C to T (cytosine deamination) and G to T (guanine oxidation) mutations, as described previously. These artifacts manifested as a higher number of "double positive" droplets than expected in experiments with a G to T or C to T mutation. These double positive droplets arose normally when a droplet was formed containing a wild-type genomic DNA fragment and a fragment harboring the variant. We could estimate the number of expected double positive droplets from the frequency of wild-type only and variant only droplets because they were independent processes that followed a Poisson distribution. To prevent these artefactual double-positive droplets from inflating the VAF estimated by ddPCR, we ignored the double positive droplets during analysis. Instead, we estimated the VAF from the number of mutant only positive droplets and the Poisson estimated number of singleton droplets. Since the variants were rare we assumed that variant positive droplets only contained a single genome equivalent of the variant allele. Likewise, we calculated the VAF by dividing the number of variant only droplets by the estimated number of droplets containing one genome equivalent of genomic DNA, which harbored either the wild-type or variant allele. For low frequency variants, where the rate of cytosine deamination could double or triple the estimated VAF, this method provided a more accurate approximation of the VAF.

This method was used for validation experiments in both the t-AML study (Chapter 3) and the study of clonal hematopoiesis in healthy individuals (Chapter 4).

2.3 Discussion

These are some of the key experiments undertaken to develop our platform for targeted ECS-based rare variant detection. These experiments underlie the decisions made regarding choice of reagents, protocol design and bioinformatics analysis. These parameters were not selected randomly. We strove to enable leukemia-associated rare variant detection in the most efficacious manner possible. Every decision was a compromise that optimized many factors including cost of sequencing, efficient biospecimen use, ease of application, breadth of utility, limit of detection and false positive rate. For example, we opted not to implement Duplex Sequencing because we did not believe the lower limit of detection was worth the dramatically higher cost of sequencing and amount of input material required. To compensate, the statistical framework for variant calling was very stringent to reduce the rate of false positive variant calls. Conversely, many true positives were likely missed.

There are many ways to improve this technology. Future development of targeted-ECS could proceed down several avenues. First, the efficiency of capture could be dramatically improved. We used the Illumina TruSight Myeloid panel to target genomic loci that were recurrently mutated in AML. While almost all of the captured molecules were on target, only approximately 5% of the genome equivalents present were captured. Liquid-phase hybridization capture had many off target reads and poor capture efficiency. As capture technology improves, it will directly benefit the detection of rare clonal mutations.

Second, improvements to the sequencing technology may make capture unnecessary. Instead of targeting regions of interest and accepting the limitations of capture, whole genome

ECS would enable the unbiased identification of rare clonal mutations. Already, one study sequenced UMI-tagged molecules from a whole genome library preparation to demonstrate that somatic mutation rates differ based on age and tissue of origin[157]. As throughput increases and cost decreases these types of studies may become more feasible.

Third, single-cell sequencing technology has the potential to revolutionize this field. With our approach, we could not determine which rare variants co-occurred in the same cells. Robust single cell sequencing could cleanly describe how clonal mutations are partitioned within a biological sample, such as a tumor, peripheral blood or sorted stem cells. Already, these techniques have been applied crudely to describe clonal architecture in pediatric ALL and pre-leukemic clonal hematopoiesis[79,85,86]. However, these studies relied on bulk sequencing to identify the somatic mutations that were re-sequenced in single cells. Improvements in single-cell genomic DNA isolation and capture would enable a new world of discovery into the biology of clonal hematopoiesis.

In conclusion, the technology developed here enabled the characterization of previously undetectable rare hematopoietic clonal mutations. We have further refined our understanding of the intricate and complicated biological processes that underlie stem cell homeostasis, clonal evolution and leukemogenesis. Fortunately, the technology is advancing rapidly and studies that are now strictly theoretical will soon become feasible.

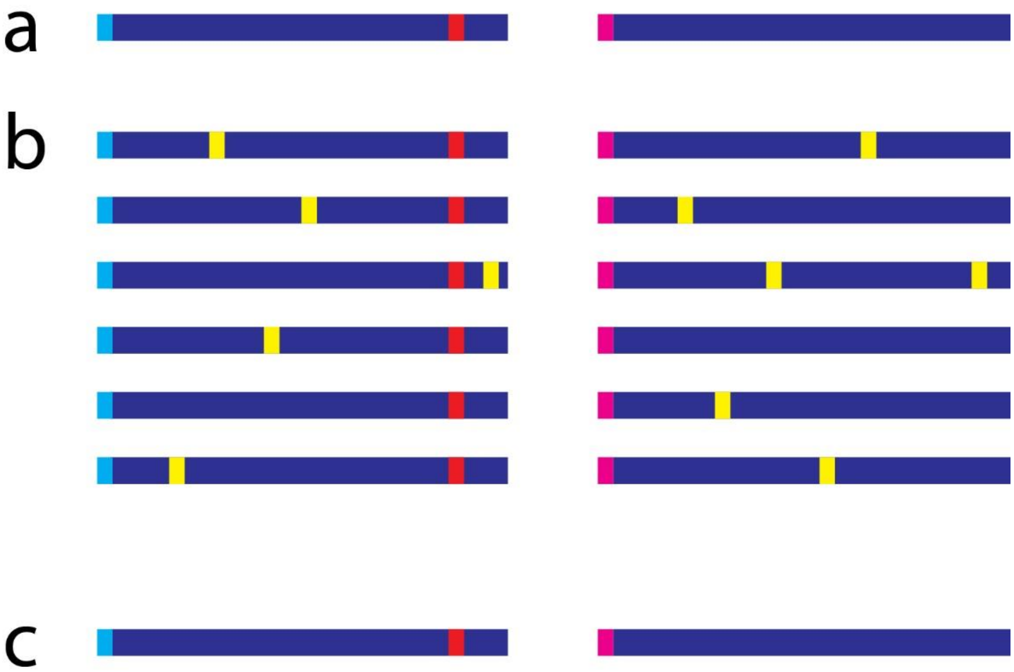

Figure 2.1 General schematic for error-corrected sequencing. a) Individual molecules of genomic DNA were tagged with a unique molecular identifier (blue and magenta). b) PCR amplification and sequencing produced multiple sequenced reads from each originally tagged molecule. Errors introduced by PCR or by sequencing (yellow) were randomly distributed across the sequenced reads. c) By comparing the sequenced reads from the same original molecule (marked by the same UMI) the sequencing errors were identified and corrected revealing true variant (red).

```
Adapter 1: 5′ AGACGGCATACGAGATNNNNNNNNNNNNNNNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
Adapter 2: 5′ pGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT
PCR Primer 1: 5′ CAAGCAGAAGACGGCATACGAGAT
PCR Primer 2: 5′ AATGATACGGCGACCACCGAGATC
Illumina Primer 1: 5′ ATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
Illumina Primer 2: 5′ TGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
Illumina Index Primer: 5′ TCGGAAGAGCACACGTCTGAACTCCAGTCAC
```

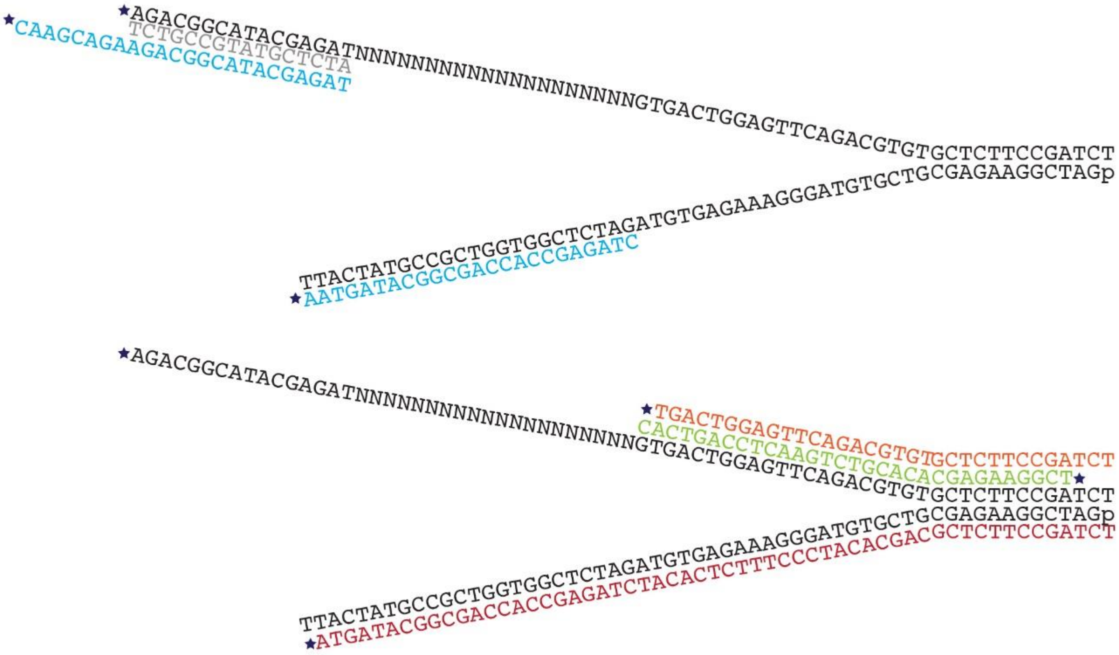
★AGACGGCATACGAGATNNNNNNNNNNNNNNNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
TCTGCCGTATGCTCTA
★CAAGCAGAAGACGGCATACGAGAT
TTACTATGCCGCTGGTGGCTCTAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGp
★AATGATACGGCGACCACCGAGATC
★AGACGGCATACGAGATNNNNNNNNNNNNNNNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
★TGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
CACTGACCTCAAGTCTGCACACGAGAAGGCT★
TTACTATGCCGCTGGTGGCTCTAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGp
★ATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

Figure 2.2 Adapter sequences that enabled error-corrected sequencing. The two sequences in black were the Y-shaped Illumina adapter sequences. The homopolymer run of N’s in the top half of the Y-shaped adapter encoded the random unique molecular identifier for each adapter sequence. The bottom half of the Y-shaped adapter was a fixed sequence. In blue were the PCR primers that amplify these molecules during library preparation after the ligation step. The other colored-coded sequences were the Illumina-specific sequencing primers. Sequences were placed to demonstrate annealing orientation. The star denoted the 5-prime end of the oligonucleotides and the lower-case “p” denoted the phosphorylated 5-prime end of the adapter, which enabled ligation.

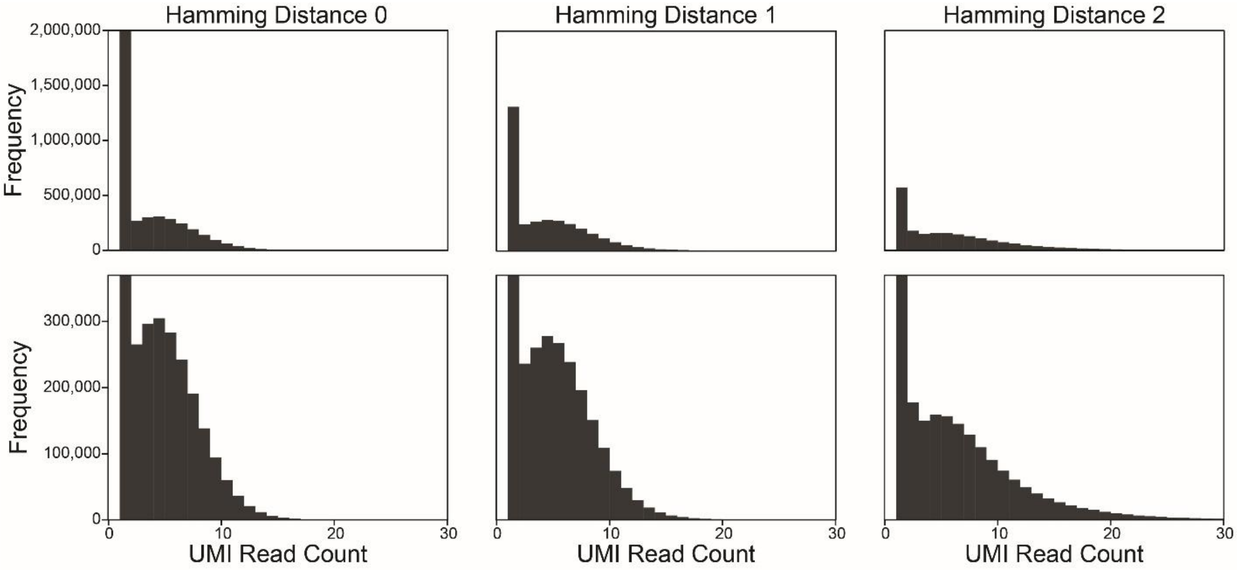

Figure 2.3 Hamming distance correction in UMI sequence. Each plot depicted the histogram of read family sizes (sequenced reads that shared the same UMI). The top and bottom rows depicted the same data with different scales on the y-axis. As the hamming distance increased, the number of read families of size one decreased, which indicated that indexes with a single nucleotide error in the UMI were correctly grouped. However, these histograms also demonstrated that many average-sized read families were collapsed with a hamming distance of 1 or 2. This suggested that the UMI sequences were not as random as originally thought.

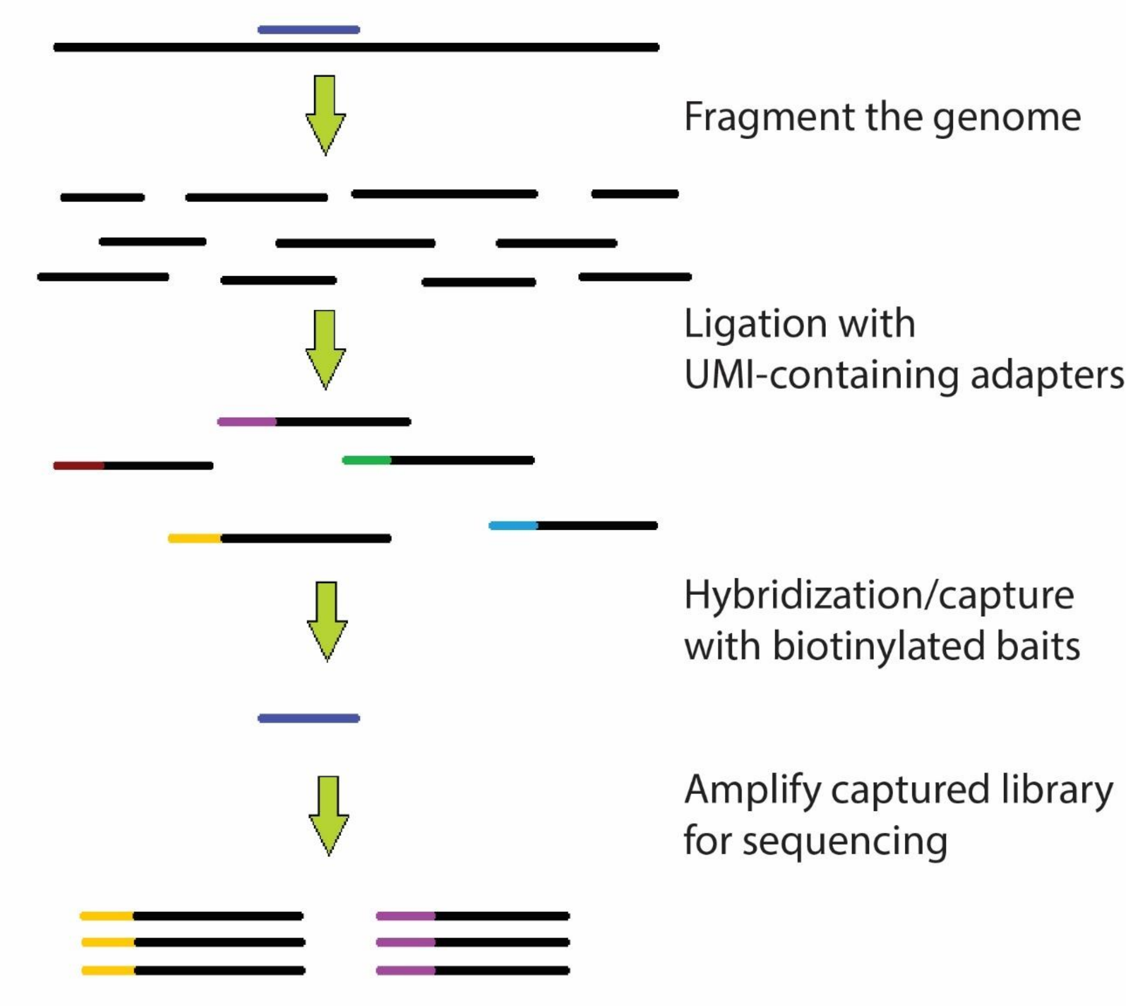
Fragment the genome
Ligation with UMI-containing adapters
Hybridization/capture with biotinylated baits
Amplify captured library for sequencing

Figure 2.4 Hybridization capture with biotinylated baits. Genomic DNA was sheared into 300 bp fragments. Adapters were ligated to the fragmented DNA molecules that contained random unique molecular identifiers for error-corrected sequencing. Hybridization-capture pulled down the fraction of the library in the region of interest based on the biotinylated oligonucleotide baits. These captured fragments were amplified and sequenced.

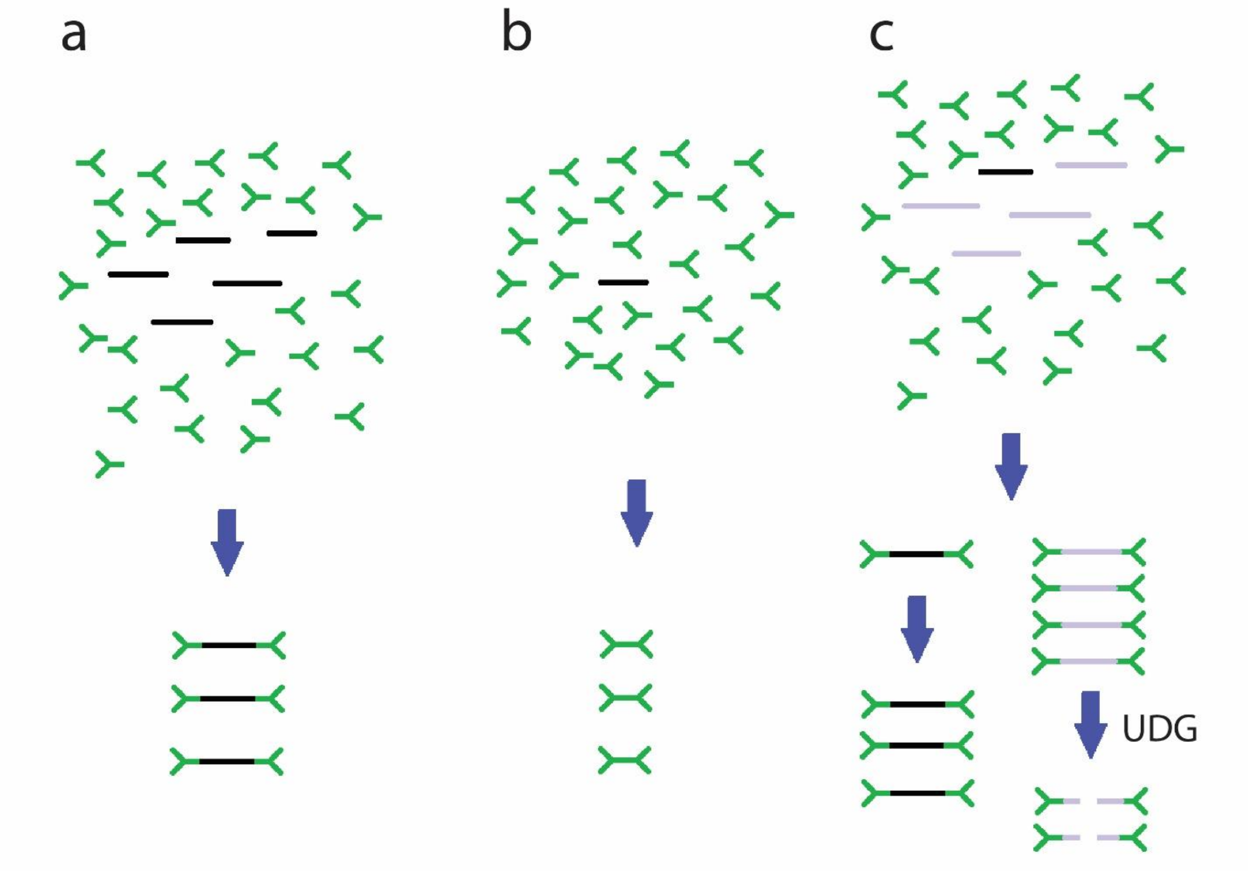

Figure 2.5 Troubleshooting Illumina Y-shaped adapter ligation. a) Under normal conditions, DNA fragments (black) and Y-shaped adapters (green) were in favorable molar ratios resulting in the ligation of one adapter to each end of the DNA fragments. b) When too few DNA molecules were present, adapter dimers formed instead. c) Adding synthesized amplicons (cyan) containing uracil to the few captured DNA fragments improved the ligation stoichiometry. The synthetic amplicons were degraded by uracil DNA glycosylase (UDG) after the ligation step.

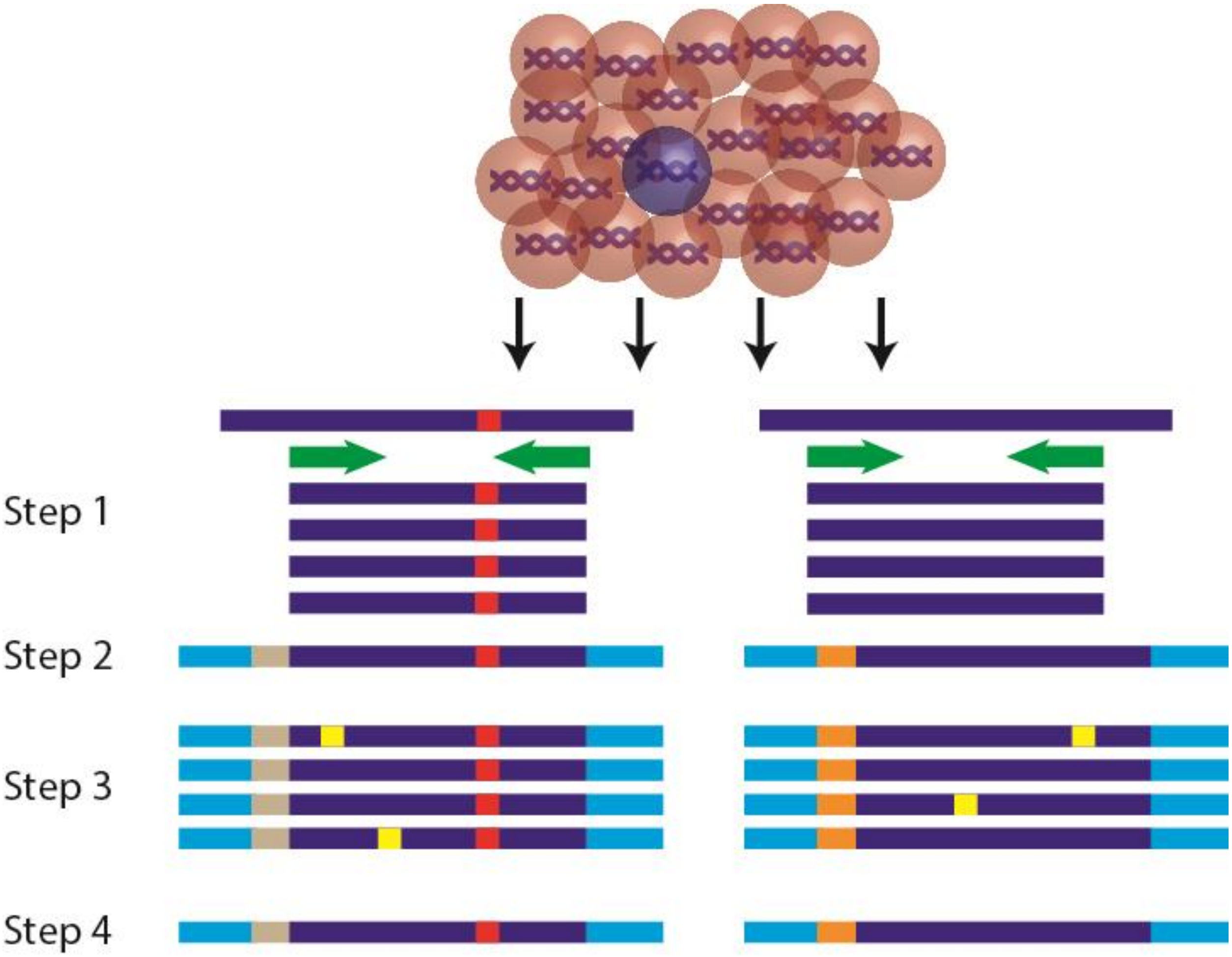

Figure 2.6 Schematic for amplicon-targeted error-corrected sequencing. Step 1: Amplification of genomic DNA with target-specific primers (green arrows) yielded a subset of amplicons containing a rare clonal mutation (red). Step 2: Randomly indexed adapters (tan and orange) were ligated to each amplicon. Step 3: Read families containing the same index sequence originated from a single UMI-tagged molecule. Sequencing errors (yellow) were randomly distributed across the sequenced reads within a read family. Step 4: Multi-sequence alignment of reads within a single read family enabled the identification of sequencing errors and the subsequent generation of an error-corrected consensus sequence.

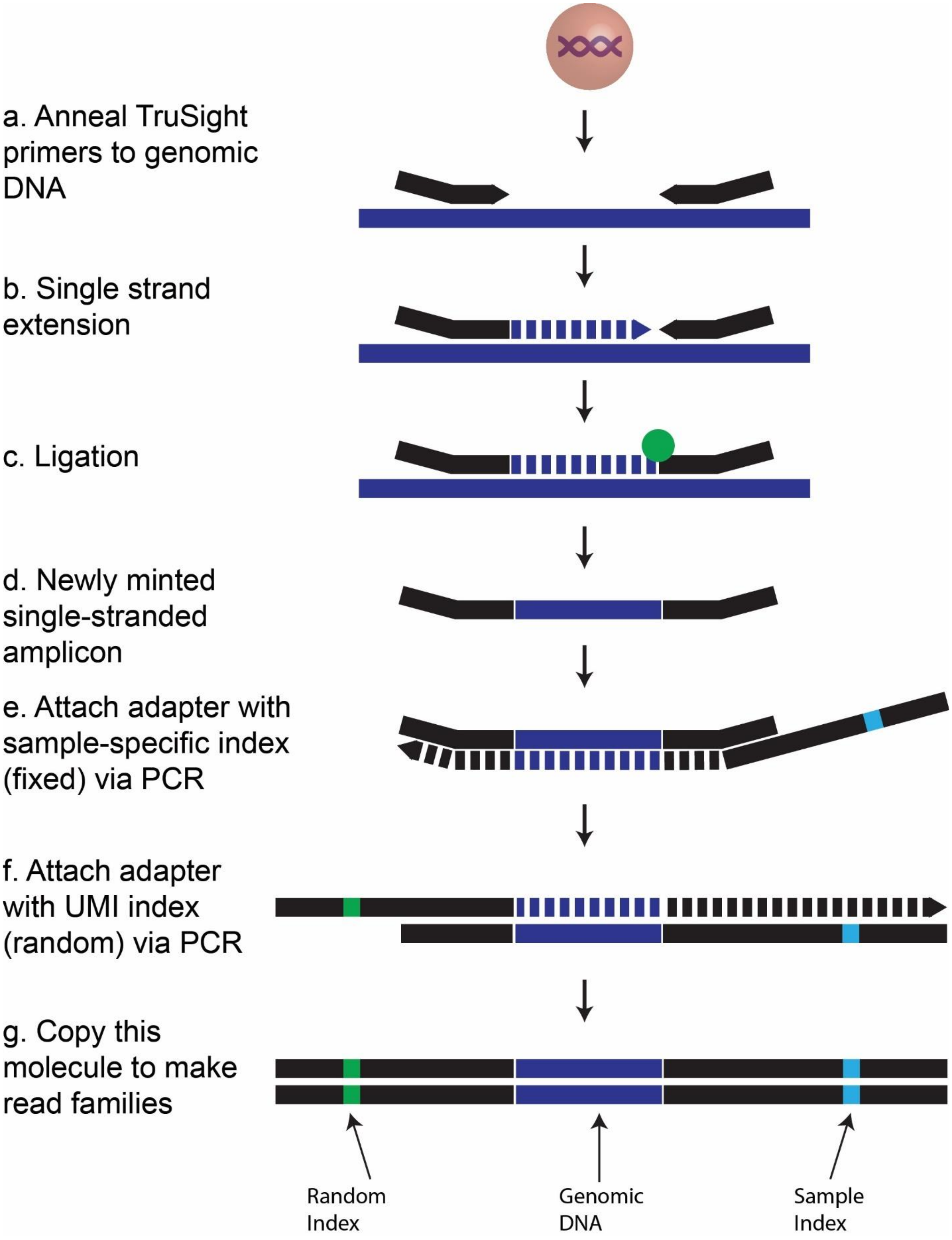
a. Anneal TruSight primers to genomic DNA
b. Single strand extension
c. Ligation
d. Newly minted single-stranded amplicon
e. Attach adapter with sample-specific index (fixed) via PCR
f. Attach adapter with UMI index (random) via PCR
g. Copy this molecule to make read families
Random Index
Genomic DNA
Sample Index

Figure 2.7 Illumina TruSight Myeloid capture and error-corrected sequencing. a) Primers (black) designed to flank a region of interest were annealed to genomic DNA (blue). b) Single strand extension filled in the nucleotides between the two primers, recording the genomic information from the template molecule. c) Ligation connected the extended strand to the downstream primer. d) The single hybridization-extension-ligation step enabled the simultaneous capture of 568 amplicons from the genomic DNA sample. e) The ends of each primer were compatible with the Illumina sequencing chemistry. We introduced the sequencing adapter and a fixed sample-specific index (cyan) using PCR directed to one end of the captured molecule. f) We introduced another sequencing adapter containing the random UMI index (green) to the other end of the molecule with PCR. g) The resulting molecules were then further prepared for sequencing.

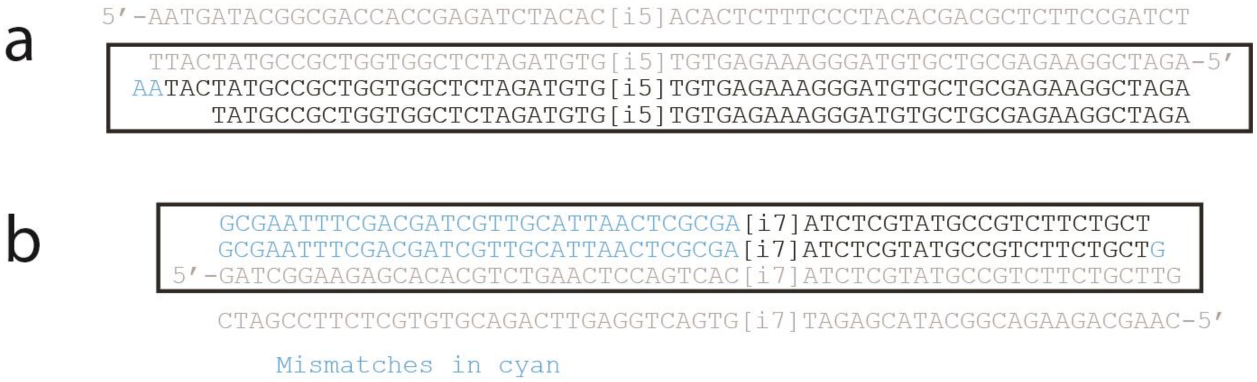
a
5′-AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCT
TTACTATGCCGCTGGTGGCTCTAGATGTG[i5]TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGA-5′
AATACTATGCCGCTGGTGGCTCTAGATGTG[i5]TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGA
TATGCCGCTGGTGGCTCTAGATGTG[i5]TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGA
b
GCGAATTTCGACGATCGTTGCATTAACTCGCGA[i7]ATCTCGTATGCCGTCTTCTGCT
GCGAATTTCGACGATCGTTGCATTAACTCGCGA[i7]ATCTCGTATGCCGTCTTCTGCTG
5′-GATCGGAAGAGCACACGTCTGAACTCCAGTCAC[i7]ATCTCGTATGCCGTCTTCTGCTTG
CTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG[i7]TAGAGCATACGGCAGAAGACGAAC-5′
Mismatches in cyan

Figure 2.8 Illumina TruSight Myeloid adapter sequences revealed by Sanger sequencing. By designing Sanger sequencing primers that annealed within one of the target regions, we were able to sequence out into the adapter sequence and determine the differences between the TruSight Myeloid adapter and the original Y-shaped Illumina adapter. The known Y-shaped adapter sequences (grey) were used to map the Sanger reads from both sequencing experiments. a) The upstream (i5) adapter sequence contained the index sequence we wished to replace with our random UMI index. Sanger sequencing of the adapter (alignments in black in the box, mismatches in cyan) revealed that the TruSight Myeloid kit used the same i5 adapter. b) The downstream (i7) adapter sequence contained the fixed index sequence we used for sample multiplexing. Sanger sequencing of the adapter (alignments in black in the box, mismatches in cyan) revealed that the i7 adapter sequence was different than the original Y-shaped adapter.

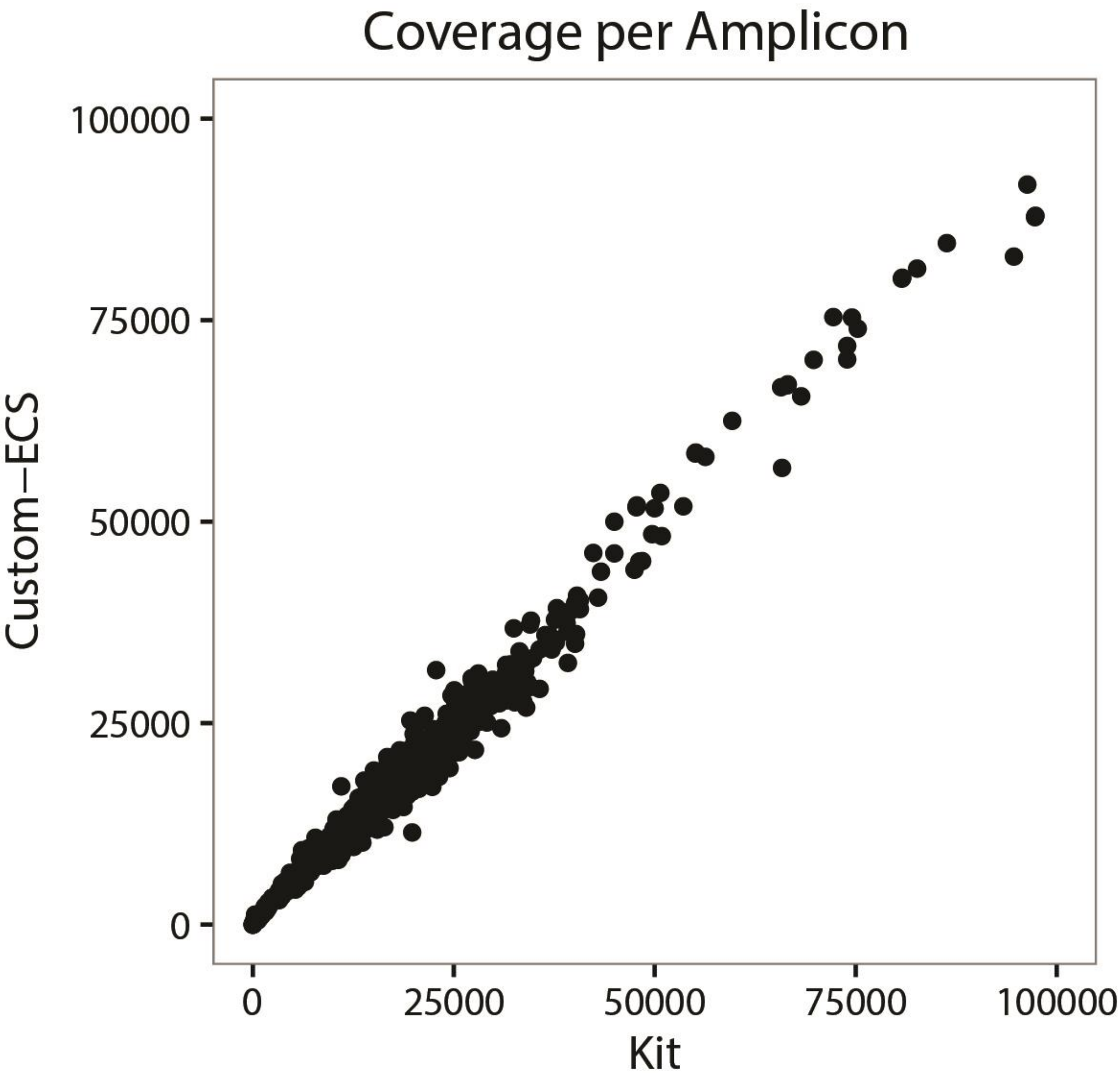

Figure 2.9 Coverage per target amplicon for standard vs ECS protocol. Sequencing of a library prepared with the Illumina TruSight Myeloid kit protocol (x-axis) was compared to the modified protocol to incorporate UMIs and ECS (y-axis). Coverage per amplicon was highly concordant between the two protocols.

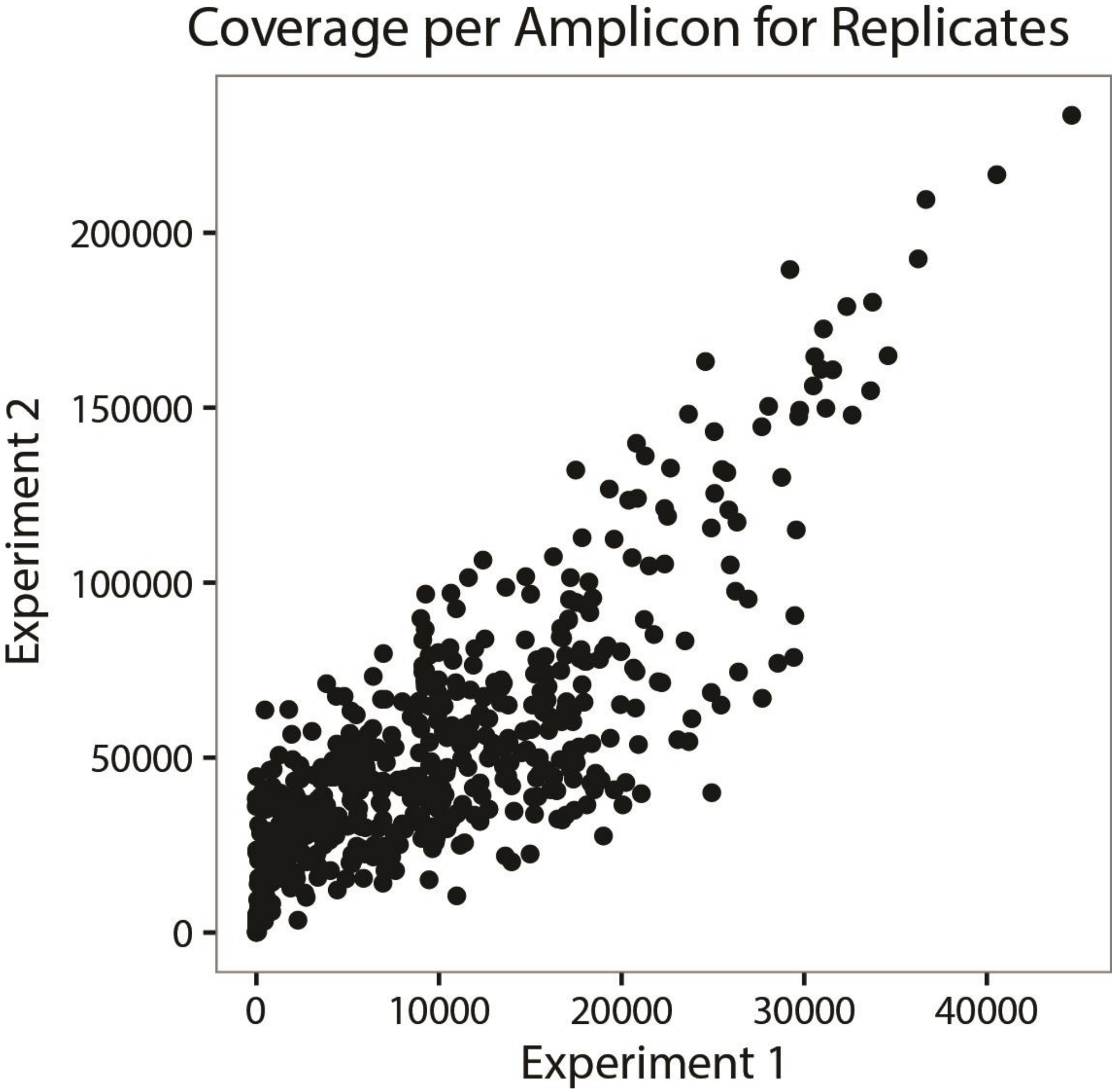

Figure 2.10 Coverage per target amplicon for two replicate ECS libraries. While the second experiment had more total coverage (y-axis), coverage per target was highly correlated with the first experiment (x-axis).

Table 2.1 Biotinylated baits example. These baits were designed for an early experiment targeting *TP53* exon 7. The 5'-biotin label was denoted as /5Biotin/.

Target	Sequence
P53exon7 Bait1	/5Biotin/ATGGGTAGTAGTATGGAAGAAATCGGTAAGAGGTGGGCCCAGGGGTCAGAGGCAAGCAGAG GCTGGGGCACAGCAGGCCAGTGTGCAGGG
P53exon7 Bait2	/5Biotin/TCAGAGGCAAGCAGAGGCTGGGGCACAGCAGGCCAGTGTGCAGGGTGGCAAGTGGCTCCTG ACCTGGAGTCTTCCAGTGTGATGATGGTG
P53exon7 Bait3	/5Biotin/TGGCAAGTGGCTCCTGACCTGGAGTCTTCCAGTGTGATGATGGTGAGGATGGGCCTCCGGTTC ATGCCGCCCATGCAGGAACTGTTACAC
P53exon7 Bait4	/5Biotin/AGGATGGGCCTCCGGTTCATGCCGCCCATGCAGGAACTGTTACACATGTAGTTGTAGTGGAT GGTGGTACAGTCAGAGCCAACCTAGGAG
P53exon7 Bait5	/5Biotin/ATGTAGTTGTAGTGGATGGTGGTACAGTCAGAGCCAACCTAGGAGATAACACAGGCCCAAGA TGAGGCCAGTGCGCCTTGGGGAGACCTG
P53exon7 Bait6	/5Biotin/ATAACACAGGCCCAAGATGAGGCCAGTGCGCCTTGGGGAGACCTGTGGCAAGCAGGGGAGG CCTTTTTTTTTTTTTTTTGAGATGGAATC

Table 2.2 Biotinylated bait capture efficiency. The capture efficiency for hybridization capture for 8 baits designed against *TP53* exon 7. Capture efficiency was poor throughout the experiment. At a bait:target molar ratio of 10^7:1 the on target fraction was close to 10%, however these reads were PCR duplicates from only a couple hundred captured molecules.

Molar Ratio Bait:Target	On Target Reads	Total Reads	Fraction
10^9:1	366	136,991	0.0027
10^8:1	9,522	145,036	0.0657
10^7:1	12,267	126,691	0.0968
10^6:1	418	166,583	0.0025
10^5:1	1,615	117,851	0.0137
NC 0:1	2	201,205	0

Chapter 3: Rare Pre-Leukemic Clone Detection in Therapy-Related AML

3.1 Introduction

The quantification of rare clonal and subclonal populations from a heterogeneous DNA sample has multiple clinical and research applications for the study and treatment of leukemia. Specifically, in the hematopoietic compartment, recent reports demonstrate the presence of subclonal variation in normal and malignant hematopoiesis[63,148], and leukemia is now recognized as an oligoclonal disease[62]. Currently, clonal heterogeneity in leukemia is studied using next-generation sequencing (NGS) targeting subclone-specific mutations. With this method, detecting mutations at 0.02-0.05 variant allele fraction (VAF) requires costly and time-intensive deep re-sequencing and identifying lower-frequency variants is impractical regardless of sequencing depth. Recently, various methods have been developed to circumvent the error rate of NGS[117,120]. These methods tag individual DNA molecules with unique oligonucleotide indexes or unique molecular identifiers (UMIs), which enable error-correction after sequencing.

Expanding upon these techniques, we developed methods for error-corrected sequencing (ECS) that enabled the study of clonal heterogeneity during leukemogenesis. We benchmarked these methods with dilution series experiments, which demonstrated quantitative SNV detection. Separately, the ECS error-profile revealed highly specific variant detection down to 1-2:10,000 molecules for all substitutions, except for G to T transversions. These substitutions were only detectable down to 1:500 due to oxidative DNA damage in the original samples.

As a pilot study for potential clinical utility, we applied ECS to identify leukemia-specific mutations in banked pre-leukemic blood and bone marrow samples from patients who later developed therapy-related acute myeloid leukemia (t-AML) or therapy-related myelodysplastic syndrome (t-MDS). These diseases occur in 1-10% of individuals who receive alkylator- or epipodophyllotoxin-based chemotherapy or radiation to treat a primary malignancy[132]. For the nine individuals surveyed in this study, matched leukemia/normal whole genome sequencing identified the t-AML/t-MDS-specific somatic mutations present at diagnosis. We applied our method for ECS to identify leukemia-specific mutations in six out of nine individuals using DNA extracted from blood and bone marrow samples collected years prior to diagnosis.

Results from two of these individuals (UPN530447, UPN341666) were published in a study specifically describing the role of *TP53* mutations in the development of t-AML/t-MDS[134]. Surprisingly, in two separate individuals in that study (not reported here), clonal TP53 mutations were detected before chemotherapy exposure, changing the established theory of leukemogenesis in t-AML/t-MDS. Previously, the chemotherapy or radiation exposure was thought to directly introduce the somatic mutations necessary for the development of t-AML/t-MDS[133]. Instead, these mutations were likely acquired in the hematopoietic compartment stochastically over the patient's lifetime and not introduced by therapy. This presented a testable hypothesis: were clonal *TP53* mutations detectable in the blood of healthy elderly individuals? Using our error-corrected sequencing approach for novel variant discovery (instead of resequencing known mutations), we demonstrated that 9/19 healthy elderly individuals harbored clonal *TP53* mutations in their peripheral blood.

Results from the remaining seven individuals were published in a subsequent study that expanded upon the methodological advancements made to enable rare pre-leukemic clone detection. In two of these seven individuals, clonal mutations were identified below the 1% threshold of detection governed by conventional NGS. These results highlighted the ability of targeted-ECS to identify clinically silent single nucleotide variations (SNV).

3.2 Methods

3.2.1 Study Design

Blood and bone marrow samples from patients treated for t-AML/t-MDS at Washington University were banked or accessed following informed consent under Human Research Protection Protocol #201011766. Patients included in this study underwent matched leukemia and non-cancer (skin) whole genome sequencing on the Illumina HiSeq 2500 platform, which identified tumor-specific somatic coding mutations in leukemia samples. Our study focused on identifying these known mutations from matched blood or bone marrow samples banked 1-13 years prior to the initial diagnosis of t-AML/t-MDS.

3.2.2 Sample Preparation

Genomic DNA was isolated from either FFPE or cryopreserved peripheral blood or bone marrow samples using the QIAamp DNA FFPE Tissue or DNA Mini Kit (Qiagen). PCR primers were designed using primer3[158] to amplify regions harboring individual leukemia-specific mutations from the banked biological samples (Table 3.1). The concentration of each purified DNA sample was determined using the Qubit dsDNA HS Assay Kit (Life Technologies). Genomic DNA (400-800 ng) was amplified using the Q5 High-Fidelity 2X Master Mix (New England Biolabs) in a 25 uL reaction with 0.5 uM primers (Figure 3.1a). The following conditions were used: 98C for 30s; 16-30 cycles of 98C for 10s, 62-72C (based on a separate optimization) for 30s and 72C for 30s; 72C for 2m; hold 10C. The PCR reactions were purified using the Agencourt AMPure XP (Beckman Coulter) bead-based protocol without modification.

For a few of the patient samples, the amount of input genomic DNA was limited. In these cases, modifications were made to the protocol to amplify multiple leukemia-specific mutations from the same biological sample (multiplex PCR). Patient-specific primers were pooled during a first round of PCR and amplified for roughly 16 cycles, similar to pre-amplification described in TAm-Seq[156]. After purification the DNA was split into a single PCR reaction per patient-specific SNVs and amplified using only that specific primer pair, again for roughly 16 cycles. This allowed us to generate diverse amplicon pools for multiple loci using only 400-800 ng of starting DNA.

3.2.3 ECS Library Preparation

The concentration of the purified PCR products was measured using the Qubit dsDNA HS Assay Kit (Life Technologies). NGS libraries were prepared from 800 ng of amplicons for each sample/mutation using the Illumina TruSeq DNA Sample Preparation Kit (Illumina). We replaced the Illumina-provided Y-shaped adapters with custom adapters containing a random 16 base pair oligonucleotide index sequence (Table 3.2). Adapters were diluted to 40 uM in Tris-EDTA with 5 nM NaCl and annealed using the following conditions: 95C for 5m then decreased by 1C every 30s to 4C. Aside from the custom adapters used for ligation, the library preparation protocol from Illumina was mostly unchanged (Figure 3.1b). Enrichment for correctly ligated products was completed using a 50 uL Q5 PCR amplification with 2 uL of ligation product and 0.5 uM Illumina specific primers under the following conditions: 98C for 30s; 6 cycles of 98C for 10s, 57C for 30s and 72C for 30s; 72C for 2m; hold 10C The PCR reaction was purified using a modified Ampure bead cleanup, which increased the size range of purification to remove adapter dimers. 100 uL of beads were washed twice with ddH_2O to remove the stock poly-ethylene glycol (PEG) solution. The solution was replaced with 25.5 uL 50% wt/vol PEG (Sigma), 37.5 uL 5M NaCl and 37 uL ddH_2O. The PCR reaction was added to this solution and purified per the standard Ampure protocol.

3.2.4 Quantification by qPCR

We sought to generate read families from a single randomly-indexed molecule with roughly seven-fold coverage. Given the bandwidth of a single Illumina MiSeq run was roughly 15-18 million read pairs, we sought to generate sequencing libraries from roughly 2.5 million molecules. To achieve this, we quantified the concentration of each library using the qPCR NGS Library Quantification Kit, Illumina GA (Agilent Technologies). Based on the measured concentration, each library was diluted to 0.4 pM such that a 10 uL volume of the diluted library would contain ~2.5 million molecules. The 10 uL aliquot of diluted sequencing library was then amplified for 16-20 cycles and purified with the same Q5 and modified Ampure bead protocol used for the previous enrichment PCR step. The final library was visualized on a 2% SYBR Safe gel (Life Technologies) and quantified using Qubit dsDNA HS Assay Kit. When multiplexing samples on a single lane of sequencing, individual sequencing libraries were combined in equimolar amounts after enrichment PCR and the pooled sample was diluted and quantified using qPCR as stated previously. However, we also found it possible to pool amplicons in equimolar amounts after the initial genomic DNA amplification and make a single sequencing library. Up to 7 different amplicons were multiplexed on a single MiSeq run. Multiplexing was only possible with mutations in different genes or within different exons of the same gene because the samples were demultiplexed by alignment.

3.2.5 Sequencing

Each library was sequenced on the Illumina MiSeq instrument as specified by the manufacturer (Figure 3.1c). Approximately, 5-10% of PhiX control DNA was spiked into each sequencing experiment. Each completed sequencing run contained roughly 15-18M paired-end 150 bp reads. Raw sequence reads were aligned to the PhiX genome using Bowtie 2[159]. Sequence reads aligning to PhiX were removed from further analysis. The remaining sequence reads were aligned to UCSC hg19/GRCh37 using Bowtie 2 for comparison against error-corrected consensus sequences (ECCS) derived from read families (below).

3.2.6 Error Corrected Consensus Sequences

Sequence reads containing the same index sequence (originated from the same randomly-indexed molecule) were aligned to each other to generate read families in a fashion similar to previously published methods[117,120] (Figure 3.1d). Previous studies used a minimum read family size of three[117]. We found using a more stringent cutoff of five reduced the error rate in the read families (Figure 3.2). The median read family size was seven reads per index (Figure 3.3). Paired-end reads within a read family were error corrected in a stepwise fashion (Figure 3.1e). First, at every position, the nucleotides called by each sequence read were compared and a consensus nucleotide was called if there was at least 90% agreement between the reads. If there was less than 90% agreement, an N was called in the consensus sequence at that position. Errors that occurred during library preparation and sequencing were removed because they were not shared between different reads within a read family. Second, an ECCS was thrown out if less than 90% of the 300 nucleotides comprising the paired-end read were assigned a non-N nucleotide. These ECCSs were locally aligned to UCSC hg19/GRCh30 using Bowtie2[159] (Figure 3.1f). The aligned ECCSs were processed with Mpileup[160] using the parameters *–BQ0 –d 10000000000000*. This removed the coverage thresholds to ensure that all of the pileup output was returned regardless of variant allele fraction (VAF) or coverage. Variant allele factions comprised of both the expected mutations and the background errors for each sample were visualized using IGV[161] and graphically represented using ggplot2[162]. Each known variant was plotted relative to the error-profile of that specific substitution class (e.g. an expected C to T transition was compared against the C to T error profile). Variants distinguishable from the noise for that specific error class and located at the expected position within the amplicon were called true positives. The threshold for calling true variants varied based on the error profile of that substitution class. Based on our benchmarking studies we were 99% specific to detect variants above 0.0034 VAF for G to T (C to A) substitutions, 0.00020 VAF for C to T (G to A) substitutions and 0.000079 VAF for the other eight possible substitutions.

3.2.7 Healthy control methods

Amplicons were prepared from healthy control genomic DNA samples using primers designed to amplify exons four through eight of TP53 (Table 3.3). Patient specific barcodes, six nucleotides in length, were appended to the 5-prime end of each primer to enable pooling of multiple samples for sequencing. Amplicons generated from each TP53 exon/patient sample combination were generated as previously described and purified products were pooled in equimolar amounts. The pooled barcoded amplicons were prepared for error-corrected sequencing as previously described. Sequencing was completed on the Illumina Hi-Seq 2500 platform. Sequenced reads were demultiplexed based on the known patient-specific barcode sequences using a two nucleotide hamming distance. Demultiplexed sequence reads were organized into read families based on their random oligonucleotide index sequence and error-corrected as outlined previously. Read families comprised of three reads or more were used for analysis. A binomial distribution of the substitution rate at each covered base in TP53 was used to identify individuals with TP53 mutations. A variant was called if the binomial p-value was less than 10^{-6}, the VAF was greater than 1:10,000, the individual read family coverage was greater than 10,000x, at least 10 read families called the variant and the VAF in the individual was greater than five times the mean VAF for all individuals with greater than 10,000x coverage at that specific nucleotide. Read families from one patient sample (barcode GTACGGC) were removed from analysis due to a high error rate.

3.3 Results

3.3.1 Design and Validation

We employed ECS by tagging individual DNA molecules with adapters containing 16 bp random oligonucleotide molecular indexes in a manner similar to other reports[117,120,163]. Our implementation of ECS easily targeted loci of interest through single or multiplex PCR and inserted seamlessly into the standard NGS library preparation (Figure 3.1). Our only deviations from the standard protocol were ligation of customized adapters containing random indexes instead of the manufacturer's supplied adapters and a qPCR quantification step prior to sequencing (Table 3.2). Following sequencing, sequence reads containing the same index and originating from the same molecule were grouped into read families. Sequencing errors were identified by comparing reads within a read family and removed to create an error corrected consensus sequence (ECCS).

We performed two dilution series experiments to assess bias during library preparation and determine the limit of detection for ECS. For the first experiment, we spiked DNA from a t-AML sample into control human DNA, which was serially diluted over five orders of magnitude. The experiment was comprised of two technical replicates targeting two separate mutations (20 total independent libraries). The results demonstrate that ECS is quantitative to a VAF of 1:10,000 molecules and provides a highly reproducible digital readout of tumor DNA prevalence in a heterogeneous DNA sample (r^2 of 0.9999 and 0.9991, Figure 3.4). A second dilution series experiment using a leukemia sample with a somatic *TP53* H179L mutation highlighted the background rate of G to T (C to A) substitutions that likely arose from DNA damage in the original samples (Figure 3.5). The limit of detection for ECS was approximately 0.002 VAF and the limit of detection for substitutions other than G to T (C to A) was approximately 0.0002 VAF (Figure 3.6).

3.3.2 Error Profile of Error-Corrected Consensus Sequences

We next characterized the error profile based on the wild-type nucleotides included in the first dilution series experiment. Variant identification using the ECCSs was 99% specific at a VAF of 0.0016 versus 0.0140 for deep sequencing alone (Figure 3.7a). We noticed that ECCS errors were heavily biased towards G to T transversions and to a lesser degree C to T transitions (Figure 3.7b, Figure 3.8), as previously observed[117,122]. When separated by substitution type, variants identified from the ECCSs were 99% specific at a VAF of 0.0034 for G to T (C to A) mutations, 0.00020 for C to T (G to A) mutations and 0.000079 for the other eight possible substitutions. While excess G to T mutations were a known consequence of DNA oxidation leading to 8-oxo-guanine conversion[117], the pre-treatment of samples with formamidopyrimidine-DNA glycosylase (Fpg) prior to PCR amplification did not appreciably improve the error profile of G to T mutations (Figure 3.9).

3.3.3 Rare Clonal Mutation Detection in Pre-Leukemic Samples

As proof of principle, we applied ECS to study rare pre-leukemic clonal hematopoiesis in nine individuals who later developed t-AML/t-MDS. Leukemia/normal whole genome sequencing at diagnosis was used to identify the leukemia-specific somatic mutations in each patient’s malignancy (Table 3.4). We applied targeted ECS to query these 26 different loci in 12 cryopreserved or formalin-fixed paraffin-embedded (FFPE) blood and bone marrow samples that were 9-22 years-old and banked up to 12 years prior to diagnosis (Table 3.5).

We generated approximately 50 Gb of 150 bp paired-end reads from 11 Illumina MiSeq runs. We targeted 1-7 somatic mutations per individual (26 mutations spanning 6.5 kb from 18 genes in total) and identified leukemia-specific clonal populations in six individuals up to 12 years prior to diagnosis (Table 3.6). For each sequencing library, we tagged approximately 2.5 million locus-specific amplicons generated from genomic DNA using high-fidelity PCR with UMI-indexed custom adapters. Sequencing errors were removed to create ECCSs as described above. Each ECCS was then aligned to the reference genome for variant calling (Figure 2.6).

Using conventional deep sequencing, we detected t-AML/t-MDS-specific mutations in prior banked samples at variant allele fractions between 0.03 and 0.87 (data not shown). In one individual (UPN 684949), deep sequencing alone was insufficient to distinguish known *ASXL1* and *U2AF1* mutations from the sequencing errors in samples banked five and three years prior to t-MDS diagnosis, respectively (Figure 3.10a,b). However, ECS identified the L866* nonsense mutation in *ASXL1* at a VAF of 0.004 (Figure 3.10c) and the S34Y missense mutation in *U2AF1* at a VAF of 0.009 (Figure 3.10d). In addition, ECS was able to temporally quantify these mutations from three pre-t-MDS samples banked yearly from 3 to 5 years prior to diagnosis (Figure 3.11, Figure 3.12). In two cases (UPN643006 and UPN942008), only a subset of the variants identified at diagnosis were present in the prior banked sample (Table 3.6). Specifically, in the UPN643006 sample, banked twelve years prior to diagnosis, a single nucleotide deletion in *ASXL1* was present at VAF 0.03. But, the G to T substitution in *ASXL1*, CTT deletion in *GATA2* and G to T substitution in *U2AF1* were not detectable in this prior banked sample. In two additional cases *TP53* mutations were detected prior to the developed of t-AML/t-MDS. In UPN530447, somatic TP53 K139N and TP53 R248Q mutations were detected six years prior to t-AML diagnosis at 0.007 VAF and 0.005 VAF, respectively (Figure 3.13). A co-occurring

CSMD1 mutation was also detected in the prior-banked sample at 0.004 VAF. But, the NUP98 Q1532H and TET2 K1299M mutations that were detected at t-AML diagnosis were not detected in the prior-banked sample. In UPN341666, a *TP53* R196* mutation was detected at approximately 0.001 VAF three years prior to the development of t-MDS (Figure 3.14). Interestingly, a RUNX1 W279* mutation that co-occurred at t-MDS diagnosis was not detected in the prior-banked sample, suggesting that this mutation was acquired later during the development of disease.

3.3.4 Rare Clonal *TP53* Mutations in Healthy Individuals

The frequency and profile of somatic single nucleotide mutations in the hematopoietic stem cells (HSCs) of normal individuals have been previously measured[62]. The number of somatic mutations increased with age and was estimated to occur at a rate of 3.2 x 10^{-9} mutations/nucleotide/year (95% CI 2.4-4.0 x 10^{-9}) for the average nucleotide in the exome. Thus, we predicted that an average 50-year-old person would have 1.6 x 10^{-7} mutations/position. These mutations would not be randomly distributed but biased towards C to T (G to A) transitions[62]. Previous studies have proposed that an individual possesses approximately 10,000 distinct HSCs[164]. We used a randomized Monte Carlo simulation to model the prevalence of somatic single nucleotide mutations in healthy 50-year-old individuals with 10,000 HSCs given a normal somatic mutational profile and mutation rate. Repeated simulation (n=100,000) allowed us to predict the distribution of aging-induced *TP53* somatic mutations. As expected, this simulation modeled a Poisson process. Mutations were deemed detrimental if they had a SIFT score less than 0.05 and on a list of putative driver TP53 mutations[37,38,165]. Using this simulation, we predicted that 44% of 50-year-old individuals harbored one or more HSCs with a detrimental *TP53* mutation (Figure 3.15). We likely underestimated that number of functional *TP53* mutations per individuals, as we are accounting only for single nucleotide somatic missense mutations in our model. Insertions, deletions, nonsense mutations, and splicing-altering mutations are not accounted for. Thus, 44% is likely a lower estimate for the number of healthy individuals with a detrimental TP53 somatic mutation in at least one HSC.

We also sought to experimentally determine the number of hematopoietic stem and progenitor cells (HSPCs) harboring *TP53* mutations in healthy individuals. We analyzed peripheral blood leukocytes from 20 elderly (68–89 years old) cancer-free donors, who had not received prior cytotoxic therapy. We targeted the DNA binding domain of *TP53* (exons 4-8) as most leukemogenic mutations occurred in this region. Using ECS, we identified *TP53* mutations in 9 of 19 evaluable cases, at 0.0001 VAF to 0.0037 VAF (Table 3.7). We used droplet digital PCR to validate these findings in all three cases tested. Most mutations detected had been observed previously in malignancy based on the COSMIC dataset[37,38]. Interestingly, we likely underreported the number of functional *TP53* mutations in healthy individuals because we only targeted a subset of the coding sequence and did not detect indel or splicing mutations. These findings suggested that somatically acquired functional *TP53* mutations in HSPCs may confer a subtle competitive advantage over time even without cytotoxic exposure.

3.4 Discussion

Here, we present a practical and clinically oriented application for targeted error-corrected NGS utilizing unique molecular identifiers (UMIs). This method easily integrated into existing NGS library preparation protocols and enabled the quantification of previously undetectable mutations in heterogeneous DNA samples. The only modification to the standard NGS library preparation was the replacement of the stock adapters with our randomly indexed adapters and the addition of a qPCR step before sequencing. The qPCR step limited the number of molecules sequenced, ensuring adequate coverage for each read family. With these two modifications, we achieve highly specific detection for rare mutations. The bioinformatics analysis was straightforward and did not require proprietary algorithms or tools (Supplementary Methods). Our results highlight the ability of this method to identify rare subclonal populations in a heterogeneous biological sample. As applied to t-AML/t-MDS, we demonstrated these previously undetectable mutations were present years prior to diagnosis and fluctuated in prevalence over time.

A clinical application of ECS is to quantify minimal residual disease (MRD). As the genomic characterization of leukemia becomes more readily available, identifying causative genetic lesions and rare therapy-resistant subclones will become increasingly useful for risk stratification, therapeutic selection and disease monitoring. Already, whole genome sequencing of AML has demonstrated that nearly every case of AML harbors one or more somatic single nucleotide variations (SNV)[56]. These SNVs are more reliable clonal markers of malignancy than cell surface markers, which can change over time. Leveraging this information, conventional NGS was implemented retrospectively to detect residual disease harboring leukemia-specific insertions/deletions (indels) as rare as 0.00001 VAF in *NPM1*[108] and 0.0001 VAF in *RUNX1*[109]. This was possible because indels were only rarely generated erroneously by NGS. Unfortunately, measuring rare leukemia-associated substitutions is limited due to the relatively high error profile of conventional NGS[166]. However, ECS can achieve the 1:10,000 limit of detection featured by conventional MRD platforms[116]. For patients whose leukemia lacks suitable markers for conventional MRD, ECS could offer an alternative with comparable sensitivity and specificity that is easy to implement in a clinical sequencing lab. Furthermore, the ability to multiplex targets for ECS enables the surveillance of known mutations and the simultaneous discovery of new somatic mutations. Ongoing work will directly compare gold-standard MRD methods to targeted ECS in patients with and without relapsed leukemia.

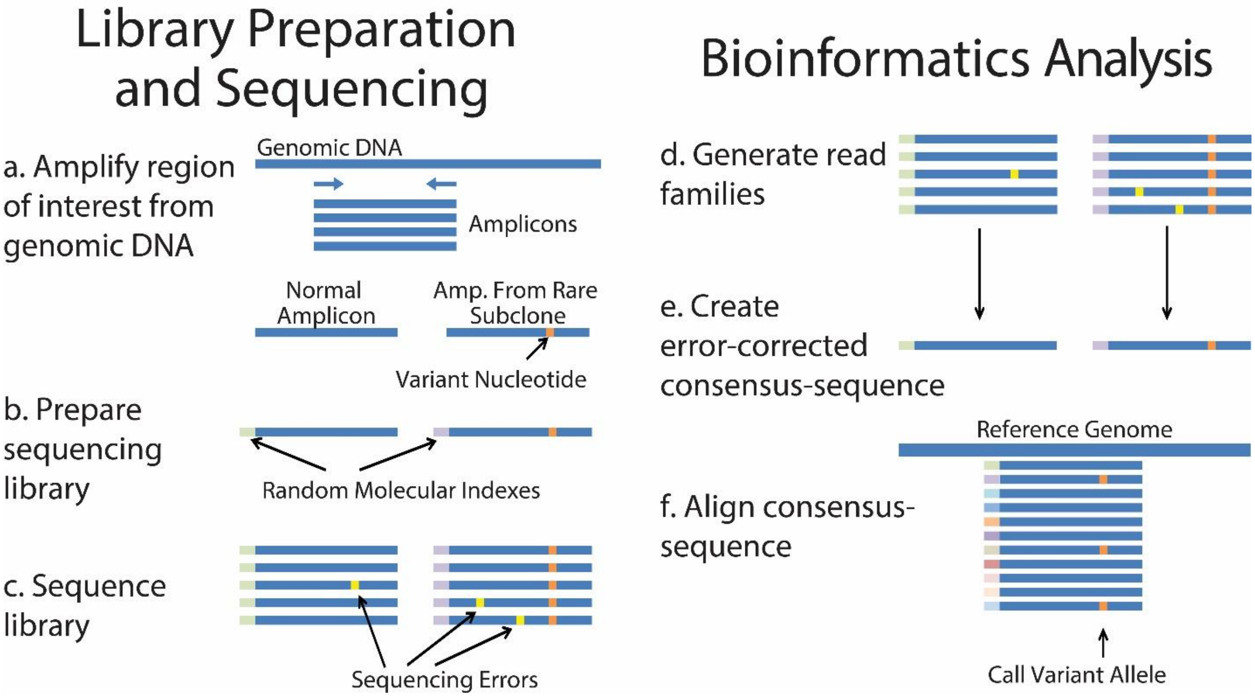
Library Preparation and Sequencing
Bioinformatics Analysis
a. Amplify region of interest from genomic DNA
Genomic DNA
Amplicons
Normal Amplicon
Amp. From Rare Subclone
Variant Nucleotide
b. Prepare sequencing library
Random Molecular Indexes
c. Sequence library
Sequencing Errors
d. Generate read families
e. Create error-corrected consensus-sequence
Reference Genome
f. Align consensus-sequence
Call Variant Allele

Figure 3.1 Schematic for amplicon-targeted error-corrected sequencing. a) Primers designed to span a locus of interest enabled the recovery of variants (orange) in that region. b) Adapter sequences containing a unique molecular identifier (UMI) were ligated to each captured amplicon. c) Amplification and sequencing for a restricted subset of these UMI-tagged molecules produced multiple sequenced reads per UMI. Sequencing errors (yellow) were randomly distributed across the sequenced reads. d) These errors were distinguishable from the correctly sequenced nucleotide in other reads from the same read family. e) Correcting the sequencing errors produced an error-corrected consensus sequence. f) The comparison of multiple error-corrected consensus sequences from different UMI-tagged read families enabled the detection of rare variants present below the error rate of conventional sequencing.

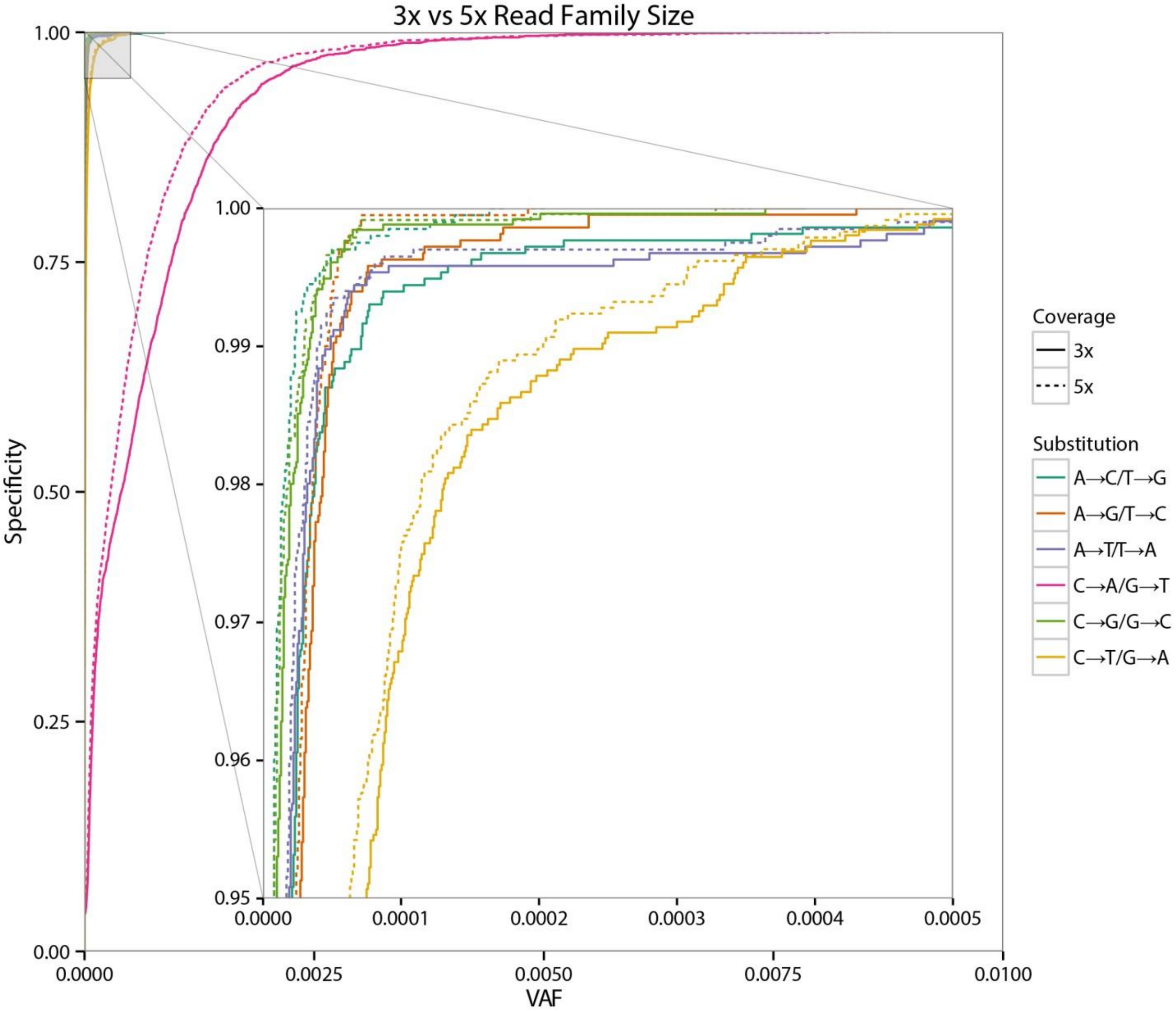

Figure 3.2 Error profile observed with increased read family size. Read families generated with 3x or greater coverage (solid line) had a higher cumulative distribution of erroneous substitutions called compared to read families with 5x or greater coverage (dotted line).

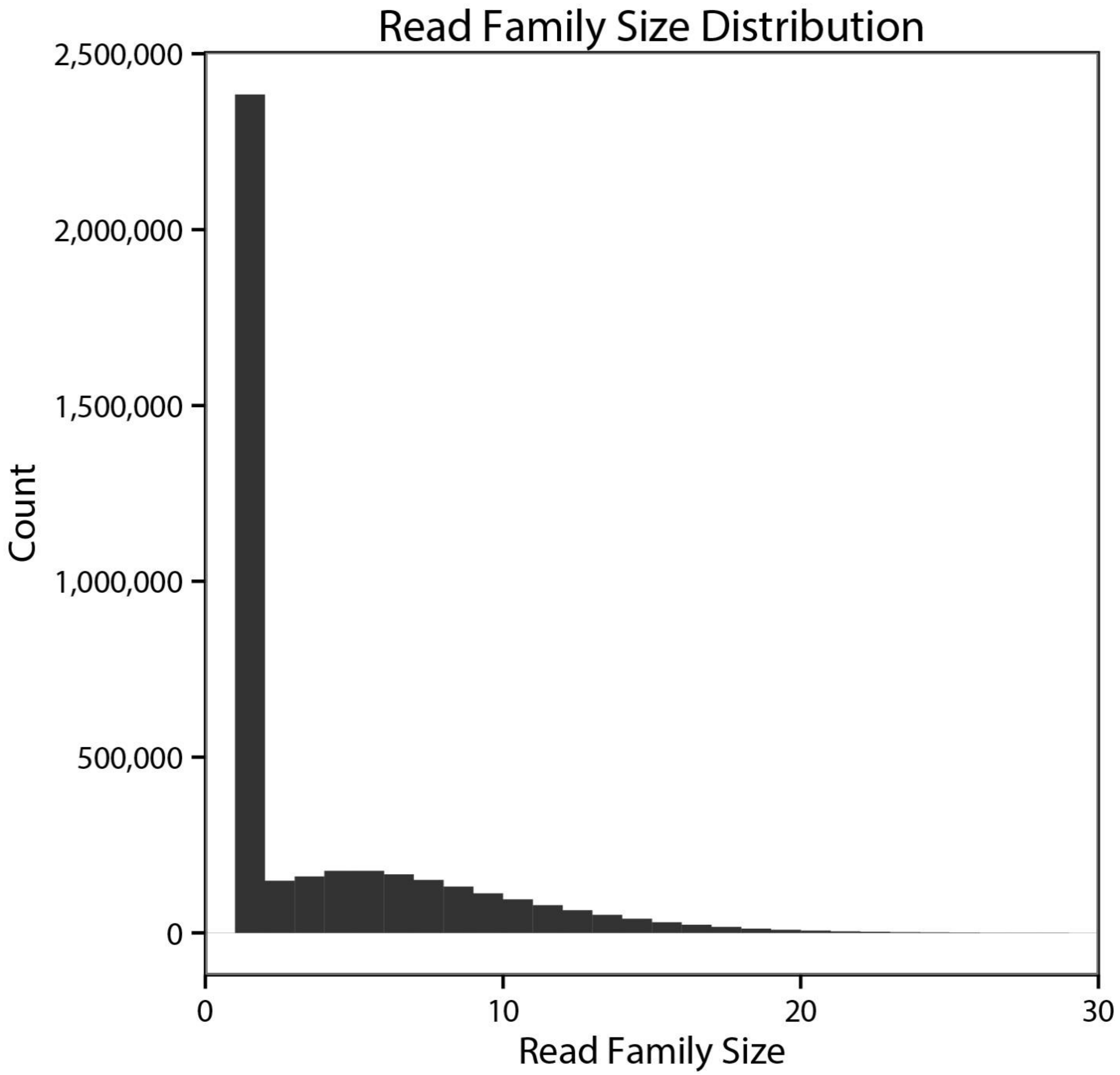

Figure 3.3 Representative distribution of read family size. Singletons predominantly represented index sequences containing a sequencing error. Excluding singletons, the median read family size was 7x (mean 7.4x). Only read families with 5-20 reads were included in ECS analysis.

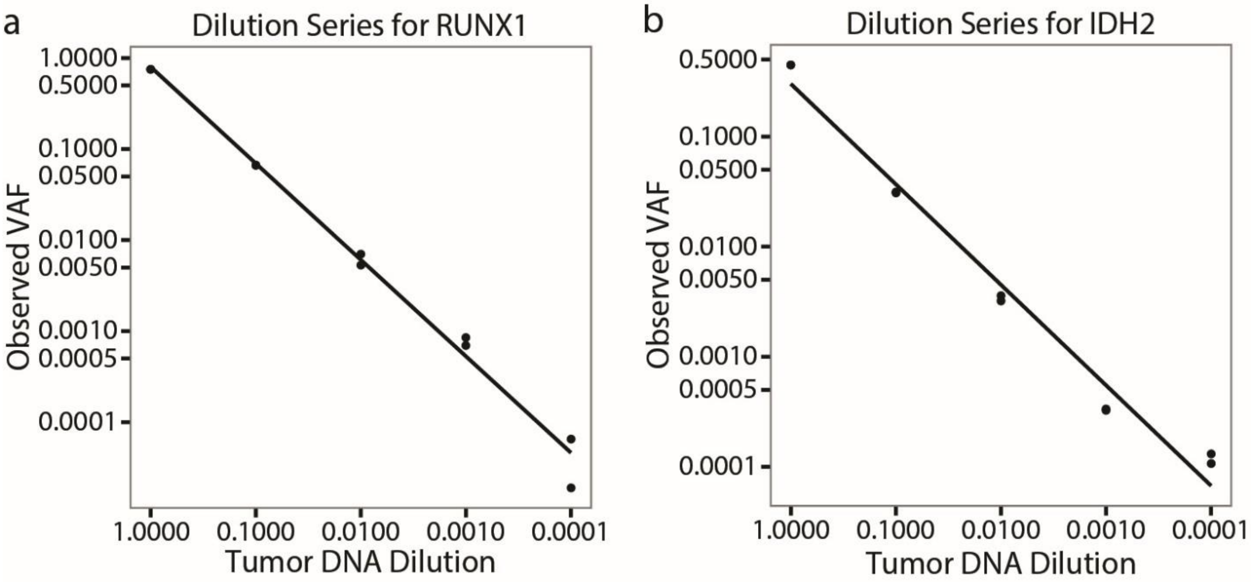

Figure 3.4 Benchmarking for ECS and the identification of rare pre-leukemic mutations. DNA extracted from a diagnostic leukemia sample with known mutations in a) *RUNX1* and b) *IDH2* was serially diluted into non-cancer, unrelated human DNA. Two replicates were analyzed per sample/dilution. The coefficient of determination (r^2) between diluted tumor concentration in the sample and VAF in the generated read families was 0.9999 and 0.9991 for *RUNX1* and *IDH2*, respectively.

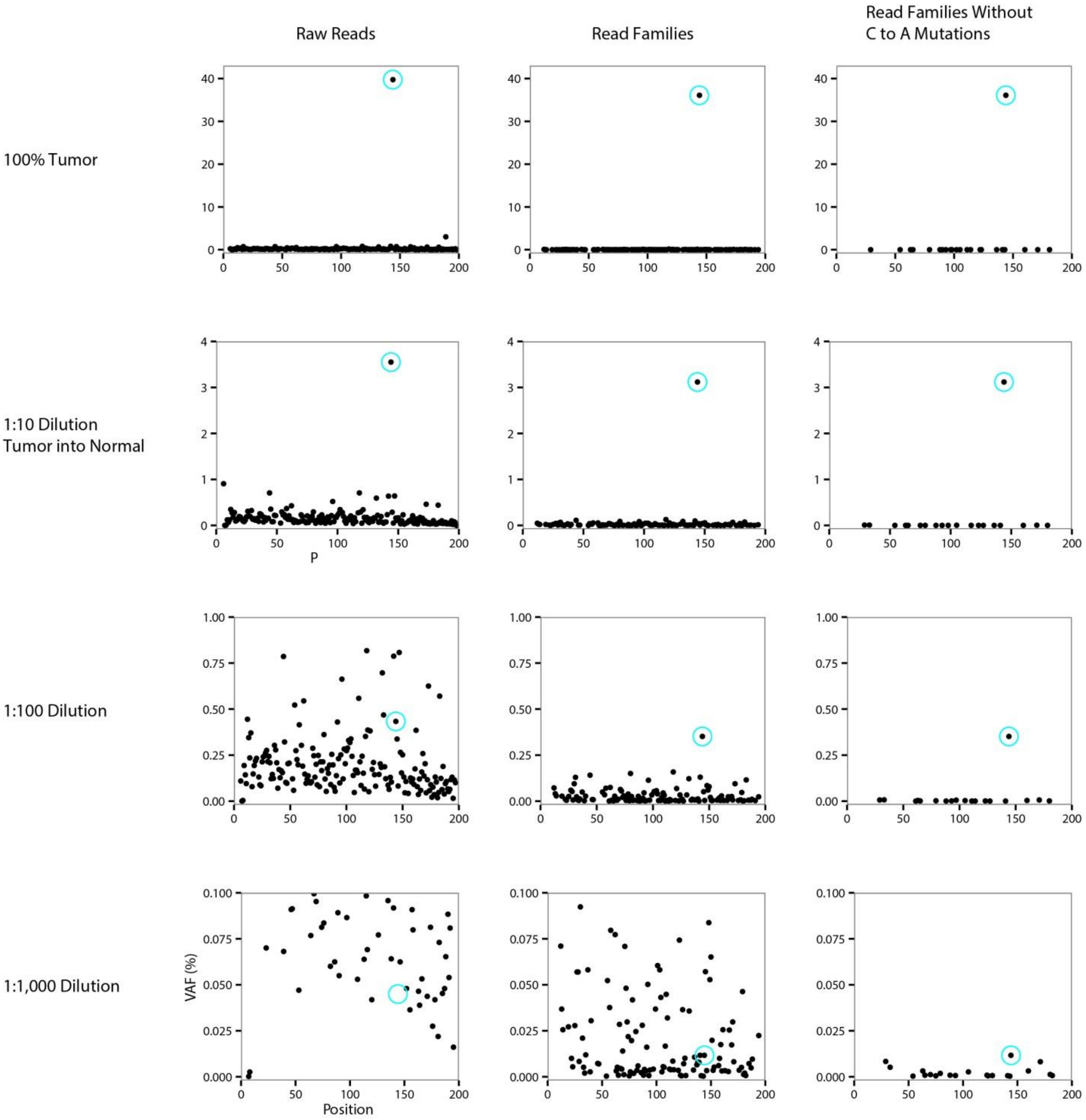

Figure 3.5 Second dilution series experiment. A leukemia sample with a somatic *TP53* H179L mutation at 0.37 VAF was serially diluted with normal genomic DNA as described by the labels on the left. The observed VAFs across the amplicon of interest with conventional sequencing (left panels) or error-corrected sequencing (middle and right panels) were plotted. Artifacts due to guanine oxidation lead to an increased rate of C to A (G to T) mutations. These data were also analyzed after removing C to A (G to T) substitutions as the variant of interest was a T to A substitution (right panel). The TP53 variant allele was circled in blue.

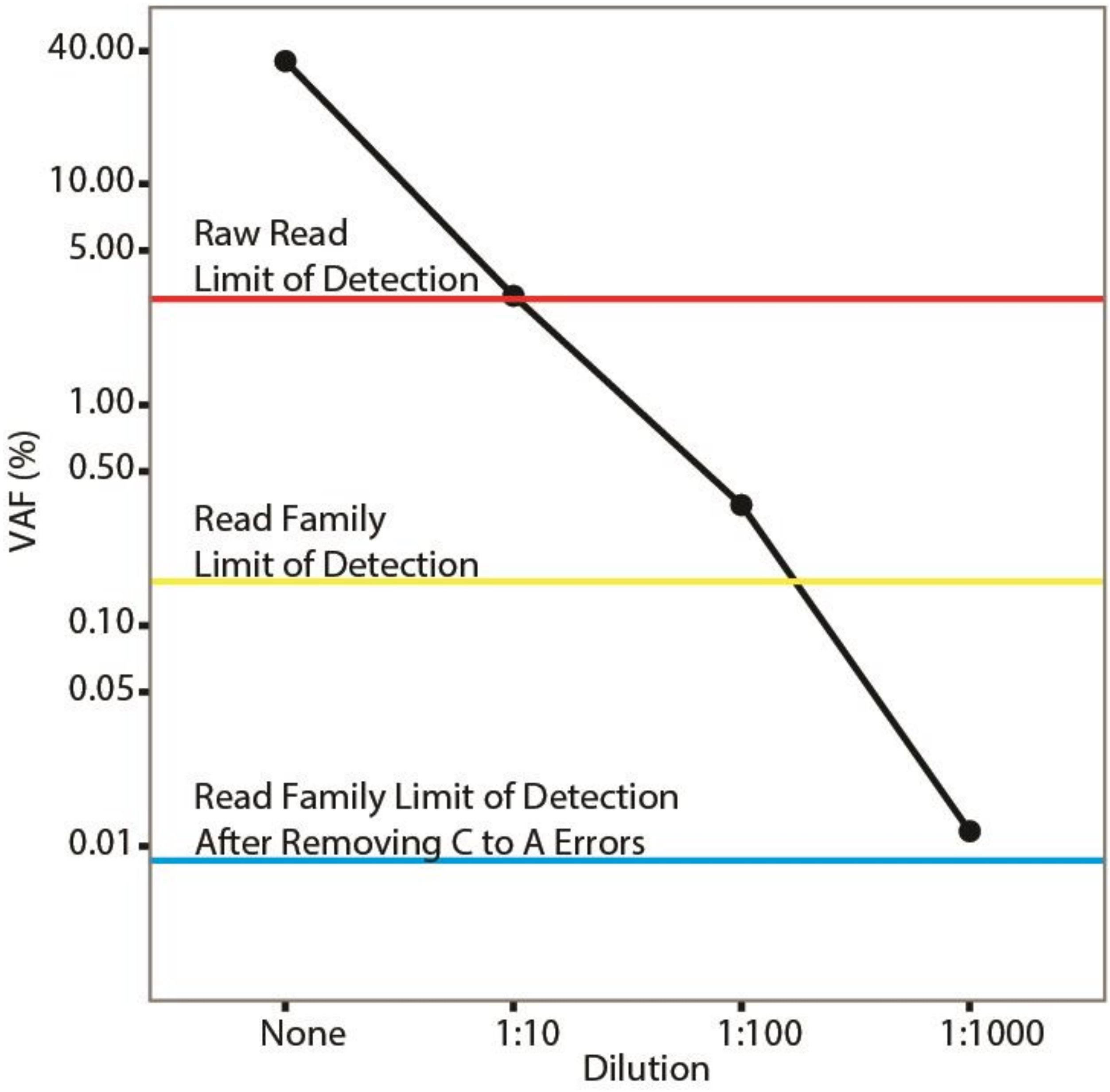

Figure 3.6 Threshold of variant detection for the second dilution series experiment. Given the second dilution series experiment, this was that range of detection for mutant alleles relative to the error rates of raw sequencing reads (red) and error-corrected read families (yellow). A DNA damage-specific C to A (G to T) error bias was observed in the read families. Sensitivity was further improved after removing C to A (G to T) substitutions (blue).

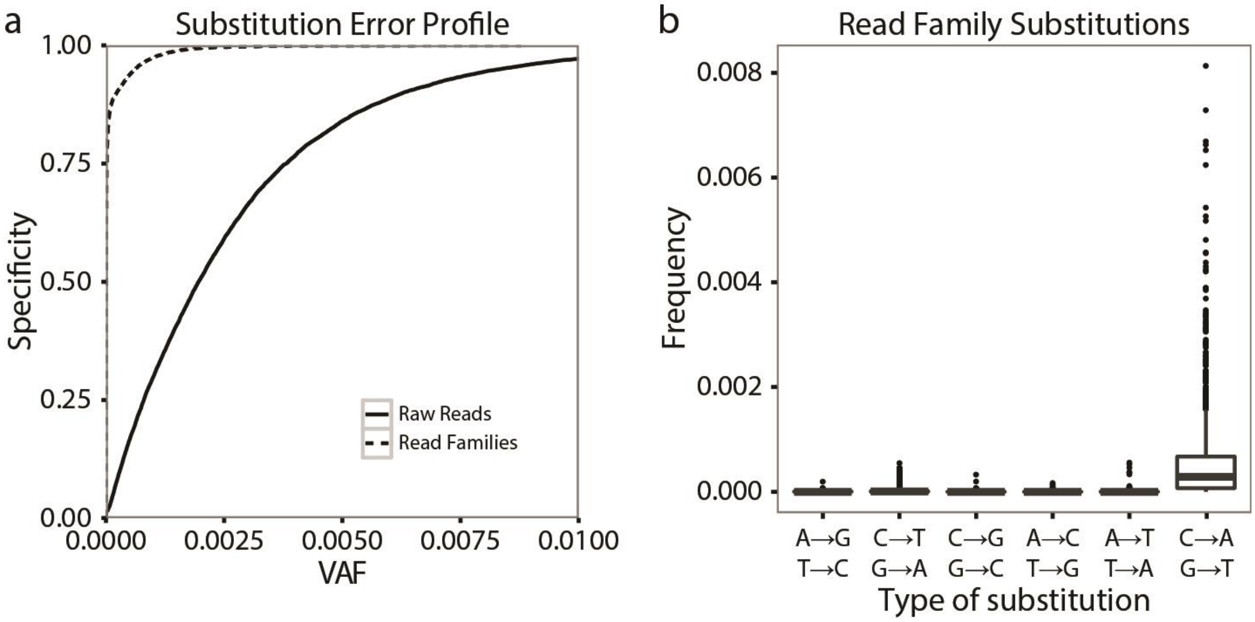

Figure 3.7 Characteristics of substitutions called from the error-corrected sequencing experiments. a) The VAF at every nucleotide not expected to contain somatic mutations in the first dilution series experiment were analyzed to determine the error profile of the error-corrected consensus sequences compared to conventional deep sequencing. A cumulative distribution function of VAF demonstrated a reduced error-profile in read families relative to conventional deep sequenced reads. b) The most frequent class of substitution seen in read families was the G to T (C to A) transversion, which was consistent with oxidative conversion of guanine to 8-oxo-guanine.

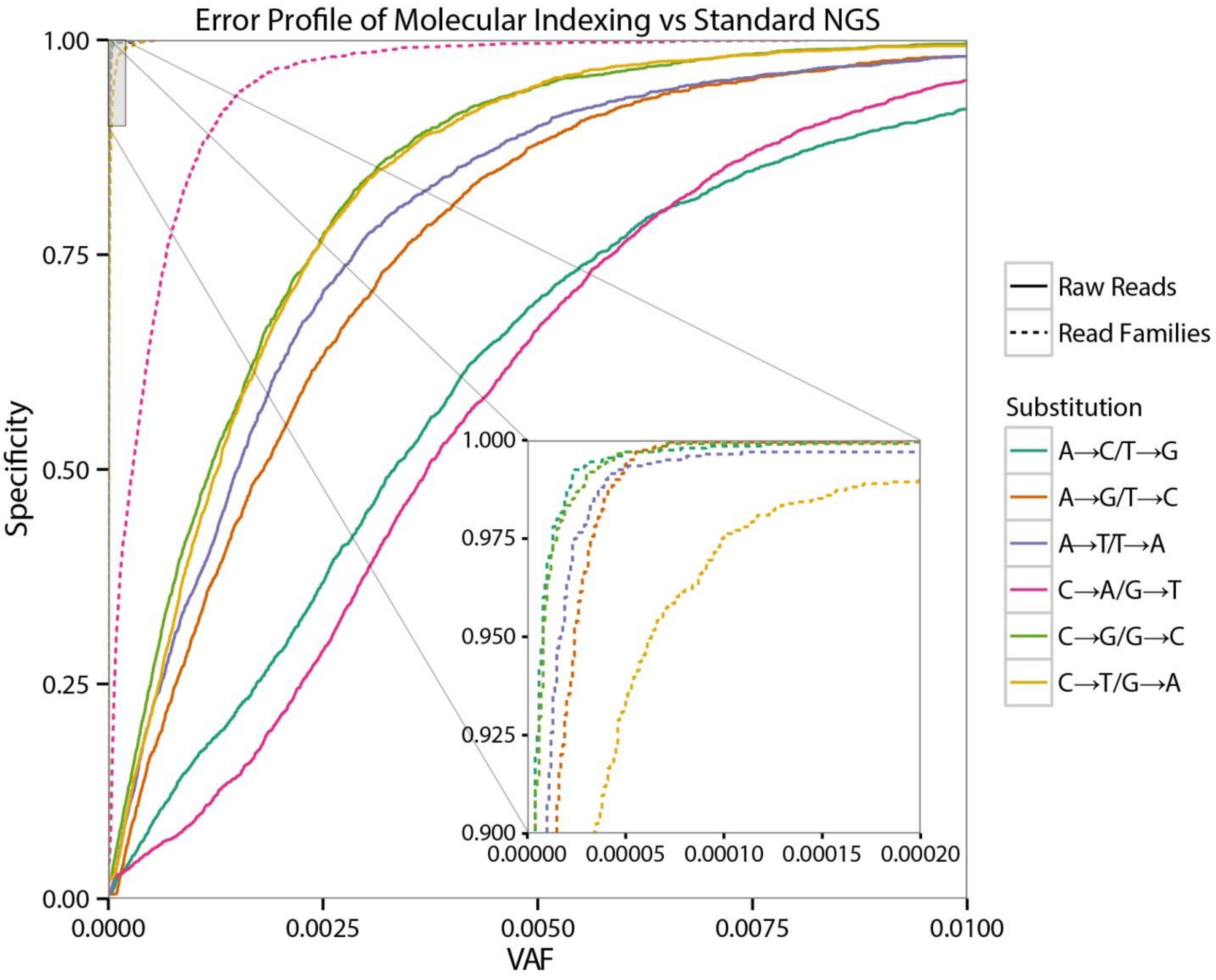

Figure 3.8 Cumulative distribution function of the error profile comparing error-corrected sequencing to conventional deep sequencing. The variant allele fraction for each non-variant position covered in the dilution series experiment was sorted and plotted cumulatively. The variant allele fractions of errors were higher in every nucleotide covered across all substitution types for the raw sequenced reads compared the error-corrected consensus sequences generated from read families.

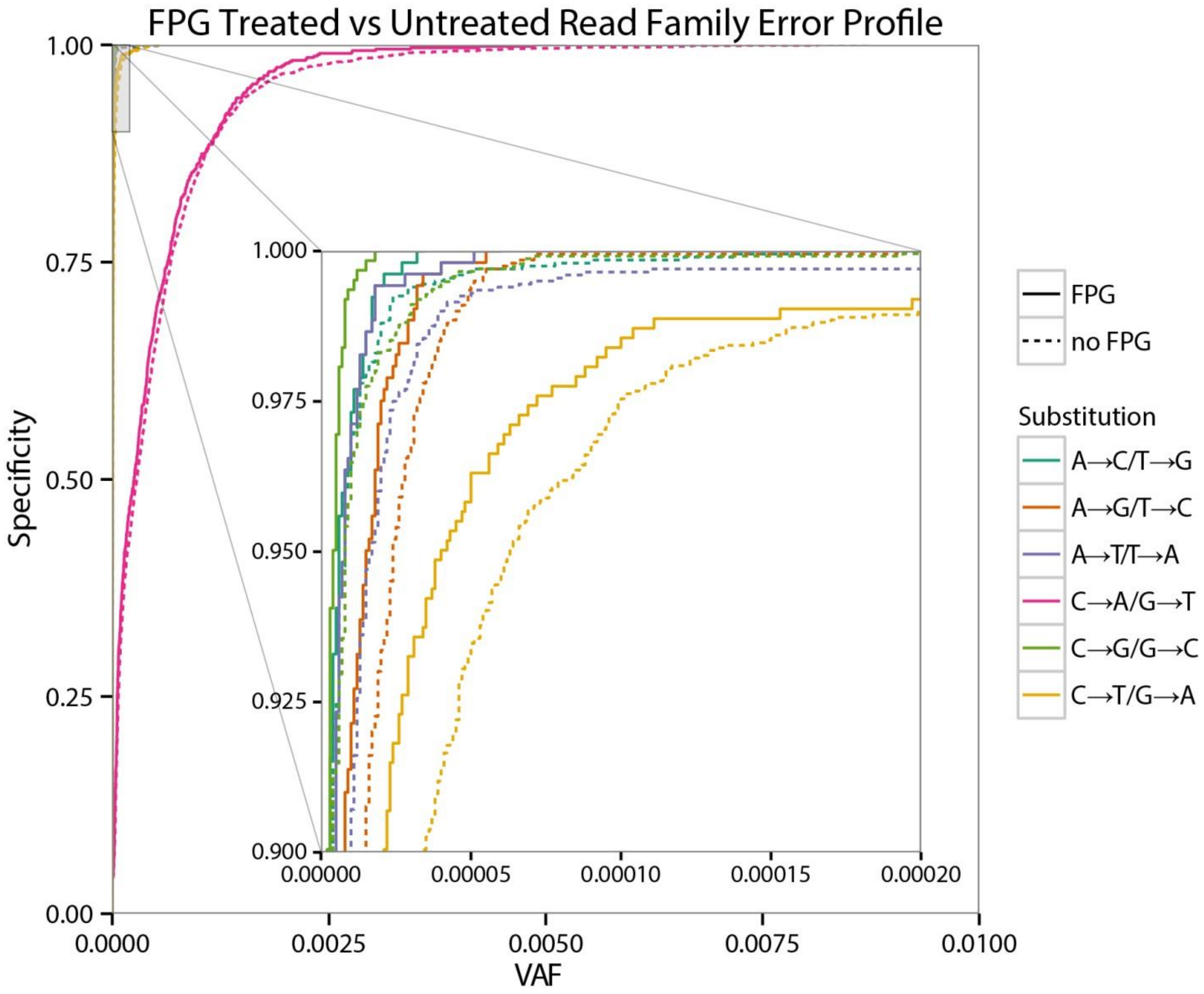

Figure 3.9 Cumulative distribution function of read family error profile per specific substitution type with and without FPG pretreatment. The error profile of G to T (C to A) substitutions, consistent with guanine oxidation to 8-oxo guanine, was higher than the other classes of mutations. The C to T (G to A) substitutions, consistent with cytosine deamination to uracil, was visible just over the error profile for the remaining 8 types of substitutions (inset). FPG pretreatment did not appreciably change the error profile.

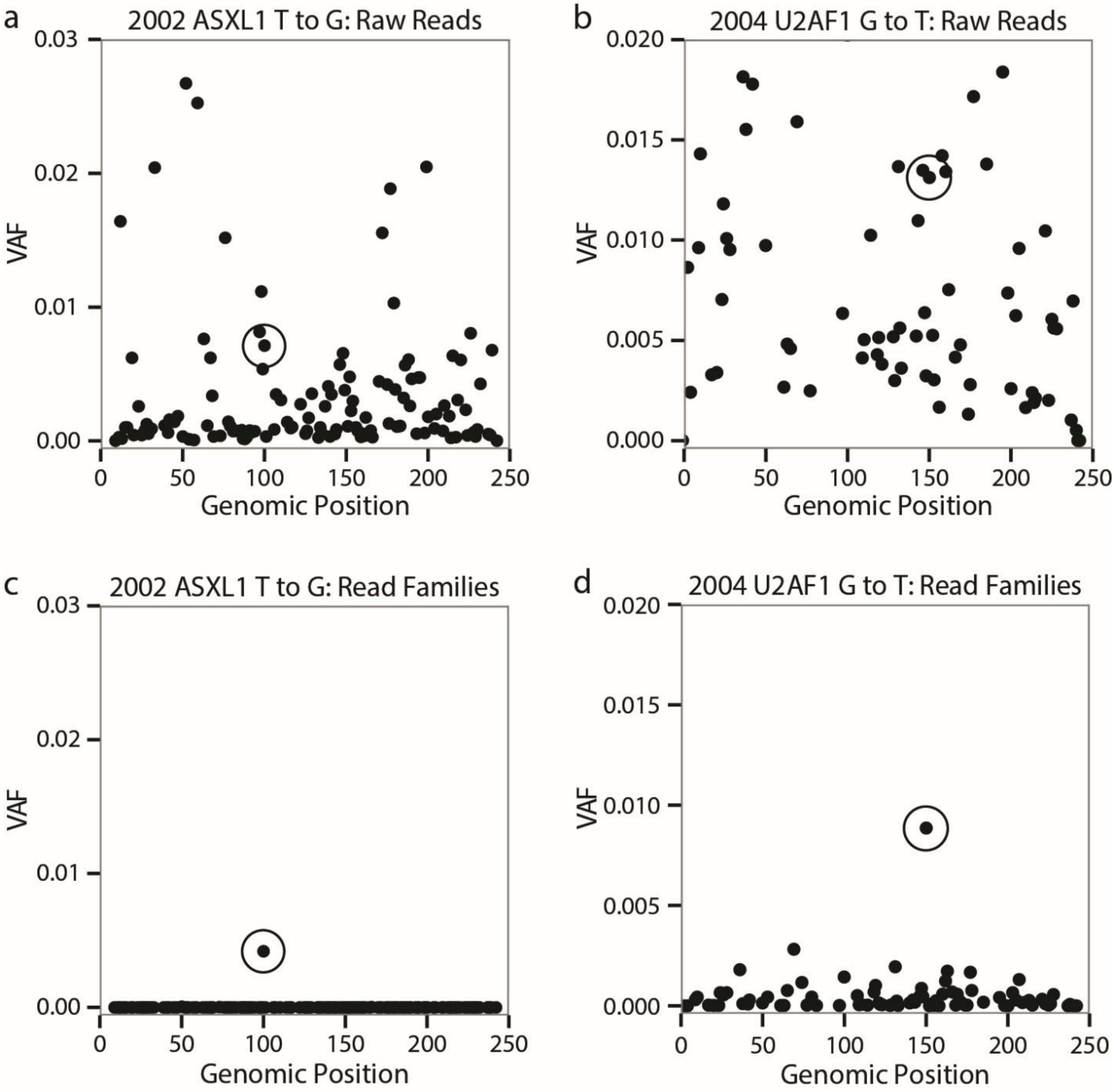

Figure 3.10 Identification of rare clonal *ASXL1* and *U2AF1* pre-leukemic mutations. a,b) The leukemia-specific variants identified in *ASXL1* and *U2AF1* at diagnosis (circled) were not distinguishable from sequencing errors in the same substitution class by conventional deep sequencing. c,d) Targeted error-corrected sequencing identified the *ASXL1* variant in the 2002 banked sample at 0.004 VAF and the *U2AF1* variant in the 2004 banked sample at 0.009 VAF.

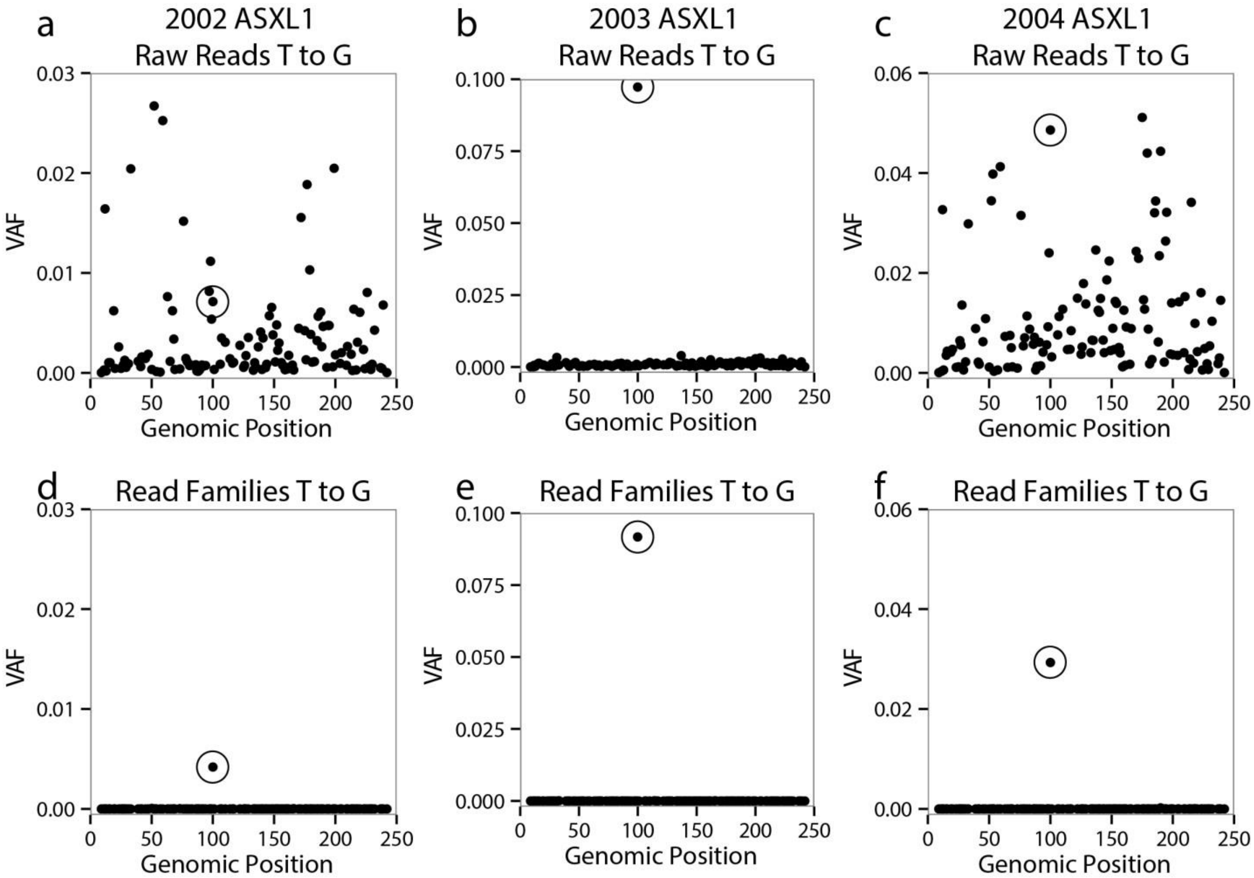

Figure 3.11 *ASXL1* mutations over time in UPN684949. Formalin-fixed paraffin-embedded bone marrow samples were banked over three years from this individual. a-c) Conventional deep sequencing only distinguished the *ASXL1* variant from the T to G sequencing errors in the 2003 banked sample at 0.097 VAF. d-f) Correcting the sequencing errors with ECS identified the *ASXL1* variant at 0.0042 VAF in 2002, 0.092 VAF in 2003 and 0.029 VAF in 2004.

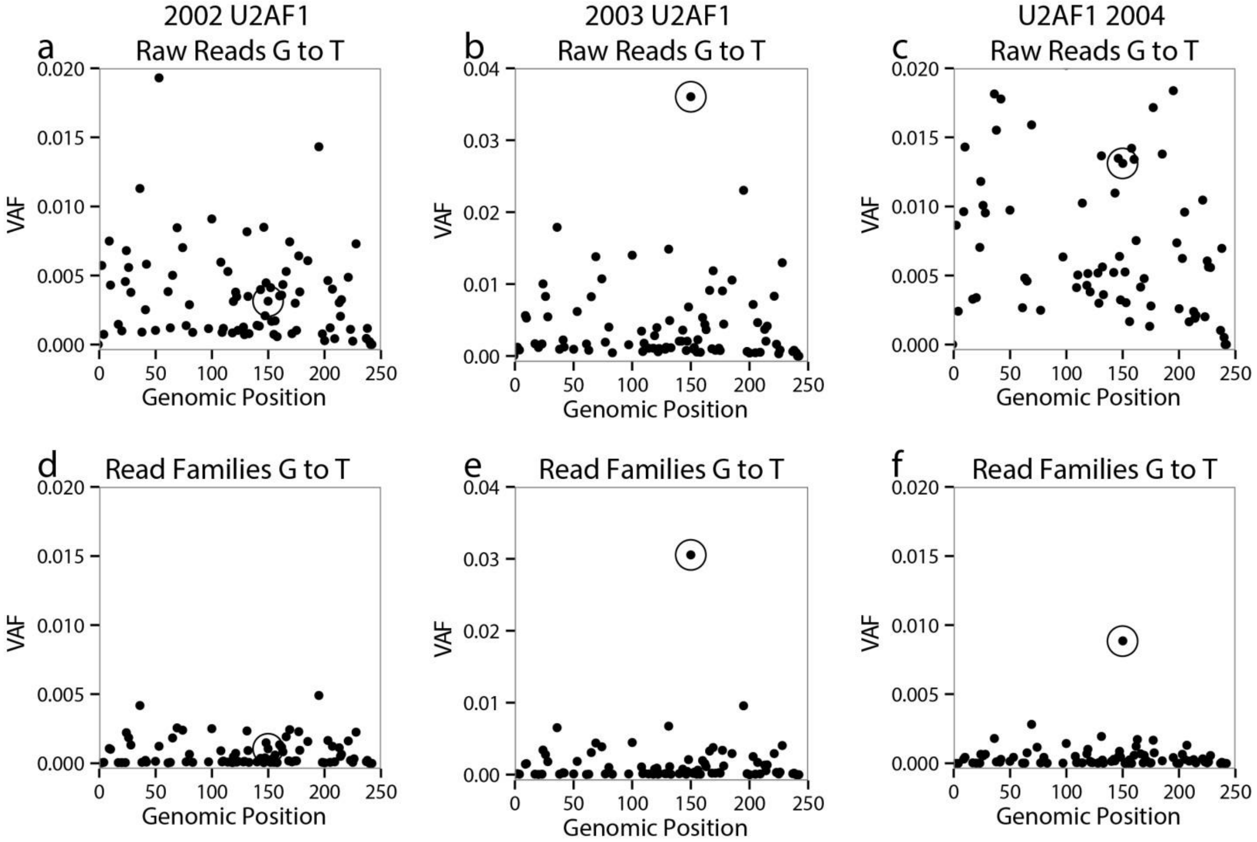

Figure 3.12 *U2AF1* mutations over time in UPN684949. Formalin-fixed paraffin-embedded bone marrow samples were banked over three years from this individual. a-c) Conventional deep sequencing only distinguished the *U2AF1* variant from the G to T sequencing errors in the 2003 banked sample at 0.036 VAF. d-f) Correcting the sequencing errors with molecular indexing did not identify the *U2AF1* variant in 2002, but did identify the *U2AF1* variant at 0.031 VAF in 2003 and 0.0089 VAF in 2004.

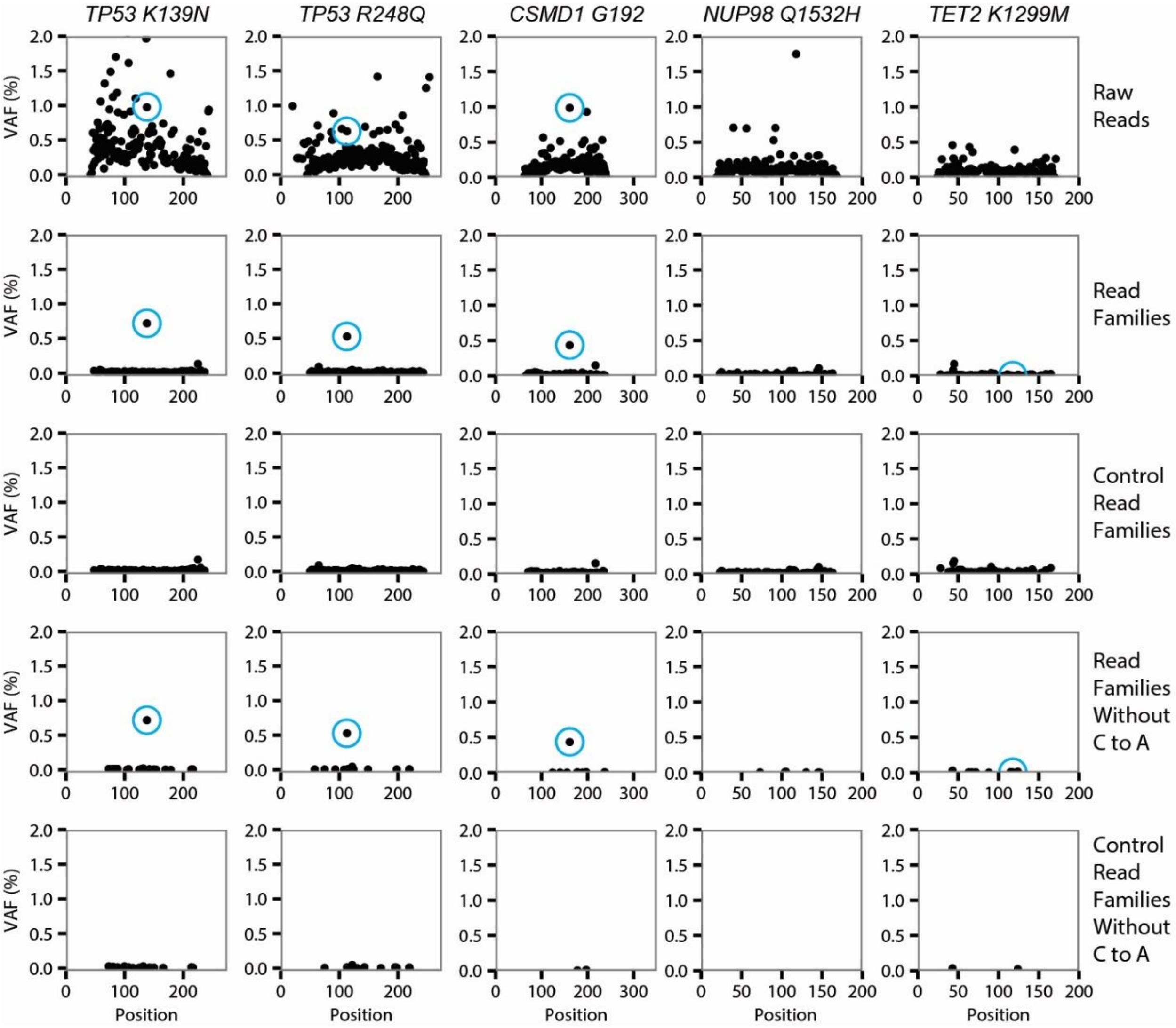

Figure 3.13 Rare clonal somatic mutations identified by error-corrected sequencing in individual UPN530447. Rare clonal *TP53* K139N and *TP53* R248Q mutations were detected at 0.007 VAF and 0.005 VAF, respectively (blue circles). These mutations were not distinguishable from the sequencing errors in the raw reads (row 1), but detectable in the error-corrected read families (row 2). The frequency of other mutations detected at t-AML diagnosis was also measured. The *CSMD1* mutation was observed at 0.004 VAF in the error-corrected read families and not distinguishable from sequencing noise in the raw reads. The *NUP98* Q1532H and *TET2* K1299M mutations were not detected.

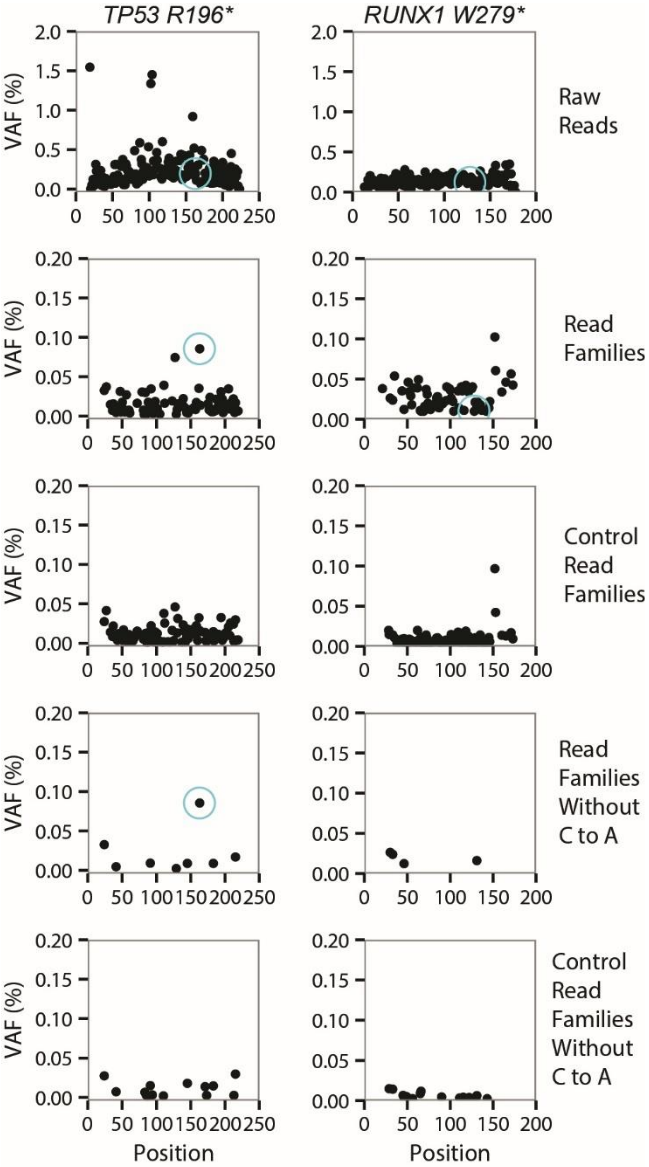

Figure 3.14 Rare clonal somatic mutations identified by error-corrected sequencing in individual UPN341666. A clonal *TP53* R196* mutation was identified at 0.0009 VAF (blue circle). This mutation was not distinguishable from the sequencing errors in the raw reads (row 1) or read families (row 2). However, removing systematic C to A (G to T) substitution errors enabled identification of the true mutation above the noise threshold (row 4). A RUNX1 W279* mutation that was detected at diagnosis was not detected in prior-banked sample. The raw sequencing results (row 3) and read family results (row 5) for a control sample without the mutations were included for comparison.

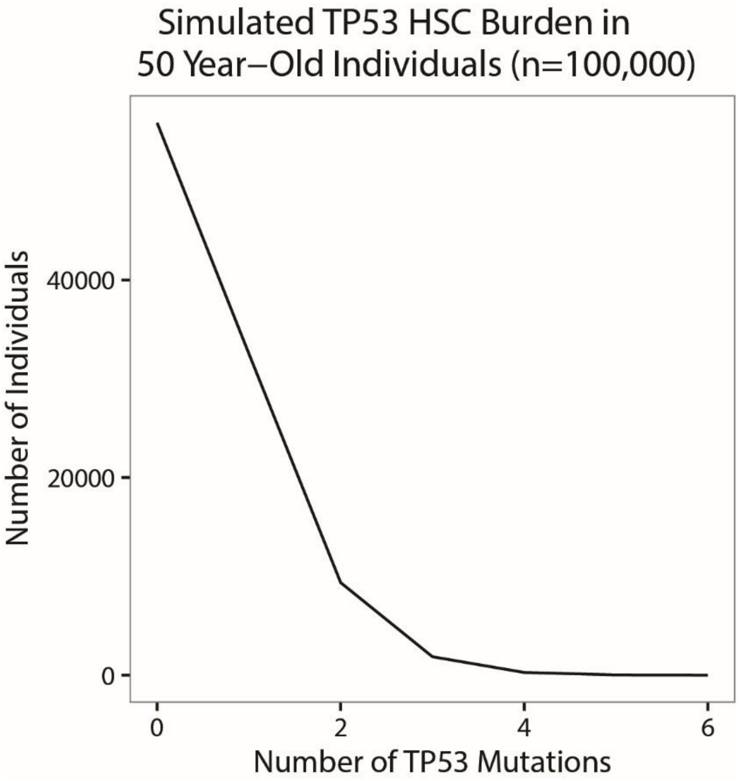

Figure 3.15 Simulated burden of predicted damaging *TP53* mutations in hematopoietic stem cells (HSCs) from healthy 50-year-old individuals. Using Monte Carlo simulation (n=100,000) with the observed exome-wide mutation rate and substitution distribution, we estimated that approximately 44% of healthy 50-year-old individuals had at least one HSC with a detrimental *TP53* mutation. The randomly distributed mutations were deemed detrimental if the SIFT score was less than 0.05 and the mutation was a putative *TP53* driver mutation.

Table 3.1 Primers targeting leukemia-specific variants. Primer sequences used to generate variant-specific amplicons from banked genomic DNA samples.

UPN	Gene	FWD Primer	Reverse Primer
341666	TP53	CCCAGGCCTCTGATTCCTCAC	GGCCACTGACAACCACCCTTAACC
	RUNX1	GGAAAGTTCTGCAGAGAGGGTTGTCAT	CCTTTCTGATTCTCTTCAGATACAAGGC
446294	OBSCN	GGAGCCTCTGACCCTGCATCCCTCC	CCCGCCTCACAGCTGTACTCCCCAG
	TP53	AGACCTCAGGCGGCTCATAGGGCAC	GGGGCTGGAGAGACGACAGGGCTG
499258	RUNX1	TCACTAGAATTTTGAAATGTGGGTTTGTTGCC	GCACTCTGGTCACTGTGATGGCTGGC
530447	TP53 K139N	AGTTGCTTTATCTGTTCACTTGTGC	CTCCGTCATGTGCTGTGACTGC
	TP53 R248Q	CCCTGCTTGCCACAGGTCTCC	AGTGTGCAGGGTGGCAAGTGG
	CSMD1	AAAGCATCTCCAAAACCATTGCCCTGCC	AAAATCCGGTACAGCTGCCTCCCTG
	NUP98	GCAGGAGGACAAAGATGGCCCAC	GACTACCGCCTAAGCTGGCACTTG
	TET2	TGGGTCATCCCCAAGCAGCTTAAAC	CAGGAGAACTTGCGCCTGTCAGG
574214	DMD	GGCGATGTTGAATGCATGTTCCAGT	AGGACTATGGGCATTGGTTGTCAAT
643006	ASXL1	GGACCCTCGCAGACATTAAAGCCCGT	GCCTCACCACCATCACCACTGCTGC
	GATA2	CCACAGGTGCCATGTGTCCAGCCAG	CTGTGGCGGGGTGGGAGGAATGTTG
	U2AF1	TGAACACAAATGGAAAATACAACTACGAGAGAAAA	CCCAGCAAAATAATCAGCTCTCATTTTCCC
684949	ASXL1	CACTATGAAGGATCCTGTAAATGTGACCCC	TGGTTTGGGCTGTTTCACTACCTCA
	U2AF1	TGAACACAAATGGAAAATACAACTACGAGAGAAAA	CCCAGCAAAATAATCAGCTCTCATTTTCCC
856024	S100A4	CCACGTGGGGACTCACTCAGGCA	AATAAGACGGTCTCTGTGCCTCCTG
	IGSF8	TGGTACACGCCTTCATCCTCGGG	GCTCAGCTCTGTCCCTGCCCAGCT
	PLA2R1	ACCCTGGTGTCTGTGGCATTCTCTG	AGTCACAGCATCATTCCTCTTGCGGT
	POU3F2	CAAATGCGCGGCTCCTTTAACCGGA	GCGTGGCTGAGCGGGTGTCC
	ANKRD18B	TACCACATTCGGGACTGGGAACTGC	CTCCCAGGGTCCCGGCGAACTCC
	ESR2	TGGCAATCACCCAAACCAAAGCATCGGT	AACCCAGATCACCTCGGAGCAGGCG
	FBN3	GGGGACACAGTTCGCAGGGGTC	GACTGGGGTGCGGGAGGTCACAGG
942008	IDH2	GGCGTGCCTGCCAATGGTGATGGG	CCGTCTGGCTGTGTTGTTGCTTGGGG
	RUNX1	ACATGGTCCCTGAGTATACCAGCCT	GGCCACCAACCTCATTCTGTTTTGT

Table 3.2 Random 16-mer molecular indexed adapters. A terminal 5-prime phosphorylation on the complementary adapter sequence was used to enable ligation (*****).

Label	**Sequence**
16N Index Adapter	AGACGGCATACGAGATNNNNNNNNNNNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
Complementary Adapter	*GATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT

Table 3.3 Primers used for amplification of *TP53* DNA binding domain in healthy individuals.

Primer	Sequence
4a fwd	GGATACGGCCAGGCATTGAAGTCTCA
4a rev	ACCCAGGTCCAGATGAAGCTCCCAG
5 fwd	GCCCTGTCGTCTCTCCAGCCCCAG
5 rev	GACTTTCAACTCTGTCTCCTTCCTCTTCCT
6 fwd	GGCCACTGACAACCACCCTTAACCCC
6 rev	GGCCTCTGATTCCTCACTGATTGCTCTT
7 fwd	CCCAGGGGTCAGAGGCAAGCAGAGG
7 rev	CTCCCCTGCTTGCCACAGGTCTCCC
8 fwd	CCTCCACCGCTTCTTGTCCTGCTTGC
8 rev	GGGTGGTTGGGAGTAGATGGAGCCTGG

Table 3.4 Whole-genome sequencing of diagnosis t-AML/t-MDS samples.

UPN	Gene	Chr	Position	Mutation	AA Change	Reference Reads	Variant Reads	VAF
341666	TP53	17	7578263	G to A	R196*	60	53	0.47
	RUNX1	21	36171728	C to T	W279*	41	36	0.47
446294	OBSCN	1	228461129	A to G	H1857R	3	5	0.63
	TP53	17	7578271	T to A	H193L	79	106	0.57
499258	RUNX1	21	36252865	C to G	R139P	122	17	0.12
530447	TP53	17	7578513	C to G	K139N	67	43	0.39
	TP53	17	7577538	C to T	R248Q	91	109	0.54
	CSMD1	8	3889461	A to C	G192	119	89	0.43
	NUP98	11	3707283	C to G	Q1532H	66	59	0.47
	TET2	4	106180868	A to T	K1299M	193	147	0.43
574214	DMD	X	32827676	G to A	R187*	103	73	0.41
643006	ASXL1	20	31022448	G to T	G645C	36	32	0.47
	ASXL1	20	31022442	del G	G645fs	33	32	0.49
	GATA2	3	128200135	del CTT	K390in_frame_del	8	10	0.56
	U2AF1	21	44524456	G to T	S34Y	24	27	0.53
684949	ASXL1	20	31023112	T to G	L866*	75	14	0.16
	U2AF1	21	44524456	G to T	S34Y	57	9	0.14
856024	S100A4	1	153517192	A to G	F27L	103	48	0.32
	IGSF8	1	160062252	G to A	P516S	28	42	0.60
	PLA2R1	2	160798389	A to G	L1431P	45	33	0.42
	POU3F2	6	99282794	C to A	S15R	15	15	0.50
	ANKRD18B	9	33524645	G to A	C53Y	26	20	0.43
	ESR2	14	64701847	G to A	A416V	40	22	0.35
	FBN3	19	8155081	G to A	P2029L	54	38	0.41
942008	IDH2	15	90631934	C to T	R88Q	10	10	0.50
	RUNX1	21	36231791	T to C	D171G	15	35	0.70

Table 3.5 Summary of patient information. The type of primary malignancy, the date of primary malignancy diagnosis, the date and type of blood/bone marrow banked prior to t-AML/t-MDS diagnosis and the date of t-AML/t-MDS diagnosis are included in the table below. At t-AML/t-MDS diagnosis, tumor/normal whole genome sequencing identified leukemia-specific mutations. Some of the prior banked blood/bone marrow samples showed evidence of clonal populations harboring those leukemia-specific mutations before the clinical detection of disease.

UPN	Primary Malignancy Diagnosis	Date Primary Malignancy	Banked Samples	Banking Type	Date Banked	t-AML/t-MDS Diagnosis	Evidence of Pre-Leukemic Subclones
341666	NHL	04/2002	22.04	Cryo	11/2002	2005 (t-MDS)	Yes
446294	Breast cancer	2002	75.02	FFPE	07/2005	2006 (t-MDS)	Yes
499258	Hodgkin’s lymphoma	1998	24.06	Cryo	02/2002	2004 (t-MDS)	No
530447	Hodgkin’s lymphoma	1993	25.01	Cryo	02/2001	2007 (t-AML)	Yes
574214	Breast cancer	1998	26.04	Cryo	01/2000	2007 (t-MDS)	No
643006	AML	1989	80.01	FFPE	04/1992	2004 (t-MDS)	Yes
684949	CLL	09/1991	91.01	FFPE	11/2002	2007 (t-MDS)	Yes
			92.02	FFPE	09/2003		Yes
			93.01	FFPE	10/2004		Yes
856024	NHL	11/2004	30.02	Cryo	03/2005	2006 (t-AML)	No
942008	NHL	08/1992	33.04	Cryo	09/1996	2005 (t-AML)	Yes
			107.01	FFPE	11/2005		Yes

Table 3.6 Patient-specific leukemia-associated somatic mutations identified by ECS. Two to seven mutations were queried per individual and the number of read families (RF) containing the variant allele or reference allele were reported and used to calculate the variant allele faction (VAF).

UPN	Sample	Yrs Prior	Gene	Chr	Position	Mut	AA Change	Variant RFs	Reference RFs	VAF
341666	22.04	3	TP53	17	7578263	G to A	R196*	431	500,828	0.0009
			RUNX1	21	36171728	C to T	W279*	2	99,421	0.0000
446294	75.02	1	OBSCN	1	228461129	A to G	H1857R	61,238	156,986	0.2806
			TP53	17	7578271	T to A	H193L	220,551	110,047	0.6671
499258	24.06	2	RUNX1	21	36252865	C to G	R139P	2	486,196	0.0000
530447	25.01	6	TP53	17	7578513	C to G	K139N	3,551	489,368	0.0073
			TP53	17	7577538	C to T	R248Q	3,377	632,791	0.0053
			CSMD1	8	3889461	A to C	G192	2472	555,704	0.0044
			NUP98	11	3707283	C to G	Q1532H	97	636,713	0.0002
			TET2	4	106180868	A to T	K1299M	17	451,219	0.0000
574214	26.04	7	DMD	X	32827676	G to A	R187*	7	199,945	0.0000
643006	80.01	12	ASXL1	20	31022448	G to T	G645C	7	85,781	0.0001
			ASXL1	20	31022442	del G	G645fs	2,898	82,245	0.0340
			GATA2	3	128200135	del CTT	K390in_fr_del	0	4,187	0.0000
			U2AF1	21	44524456	G to T	S34Y	85	414,613	0.0002
684949	91.01	5	ASXL1	20	31023112	T to G	L866*	3,583	853,598	0.0042
			U2AF1	21	44524456	G to T	S34Y	545	514,410	0.0011
	92.02	4	ASXL1	20	31023112	T to G	L866*	54,074	535,976	0.0916
			U2AF1	21	44524456	G to T	S34Y	11,195	355,276	0.0305
	93.01	3	ASXL1	20	31023112	T to G	L866*	17,319	573,629	0.0293
			U2AF1	21	44524456	G to T	S34Y	827	92,104	0.0089
856024	30.02	1	S100A4	1	153517192	A to G	F27L	0	211,512	0.0000
			IGSF8	1	160062252	G to A	P516S	0	22,614	0.0000
			PLA2R1	2	160798389	A to G	L1431P	2	338,616	0.0000
			POU3F2	6	99282794	C to A	S15R	8	201,240	0.0000
			ANKRD18B	9	33524645	G to A	C53Y	7	214,836	0.0000
			ESR2	14	64701847	G to A	A416V	10	135,861	0.0001
			FBN3	19	8155081	G to A	P2029L	0	152,304	0.0000
942008	33.04	9	IDH2	15	90631934	C to T	R88Q	23,170	236,587	0.0892
			RUNX1	21	36231791	T to C	D171G	40	253,168	0.0002
	107.01	<1	IDH2	15	90631934	C to T	R88Q	138,180	161,371	0.4613
			RUNX1	21	36231791	T to C	D171G	368,438	50,796	0.8788

Table 3.7 Rare clonal *TP53* mutations identified in healthy individuals. Mutations were identified by ECS in exons 4-8 in *TP53* in 9/19 healthy 50-year-old individuals. Most of the mutations had been observed previously in malignancy as reported in the COSMIC database. Three variants were also assayed by ddPCR and the expected variants were observed at similar VAFs (last column).

ID	Chr	Pos	Ref	Var	AA Change	COSMIC	Var RF	Total RF	VAF (ECS)	VAF (ddPCR)
34	17	7577505	T	G	D259A		13	33,085	0.0004	-
99	17	7577124	C	T	V272M	10891	26	81,015	0.0003	-
99	17	7577548	C	T	G245S	6932	18	41,836	0.0004	-
269	17	7577120	C	T	R273H	10660	489	420,026	0.0012	-
271	17	7577209	C	T	Intronic		36	333,996	0.0001	-
271	17	7578413	C	T	V173M	11084	177	182,809	0.0010	0.0008
271	17	7578449	C	T	A161T	10739	25	164,591	0.0002	-
271	17	7579310	A	T	Splicing	1522474	23	165,672	0.0001	-
273	17	7578265	A	G	I195T	11089	57	15,540	0.0037	0.0028
300	17	7578190	T	C	Y220C	10758	91	316,765	0.0003	0.0003
324	17	7577094	G	A	R282W	10704	51	86,090	0.0006	-
335	17	7577539	G	C	R248G	11564	245	218,077	0.0011	-
338	17	7577539	G	A	R248W	10656	188	51,001	0.0037	-

Chapter 4: Clonal Hematopoiesis in Healthy Individuals

4.1 Introduction

The advent of cost-effective, next-generation sequencing (NGS) has permitted in-depth analysis of the spectrum of somatic mutations driving clonal evolution in malignancy[3,54,56]. Subsequently, benign clonal hematopoiesis has been identified in healthy individuals[136,139,143,167]. Recent studies revealed that malignant and benign hematopoietic clones frequently harbor mutations in the epigenetic modifiers *DNMT3A* and *TET2*[56,144–147]. Benign clones were rarely detected before 60 years old, but were detected in 10-20% of individuals older than 70 years old[144–147]. While compelling, these previous studies could only detect common clonal mutations—greater than 0.02 variant allele fraction (VAF)—due to the NGS error-rate. Hematopoietic clones detected above this 0.02 VAF threshold have been termed clonal hematopoiesis of indeterminate potential (CHIP) and are associated with an increased risk of developing hematological malignancy[149].

Recently, the development of error-corrected sequencing (ECS) using single molecule tagging with unique molecular identifiers (UMIs) has permitted the detection of rare variants below the error-rate of NGS[117–120,134,135]. Here, we combined ECS with targeted capture for 54 genes recurrently mutated in acute myeloid leukemia (AML) to enable the detection of clonal mutations at VAFs two orders of magnitude lower than the detection limit of NGS. Using these methods, we sought to thoroughly describe the prevalence and mutation profile of rare hematopoietic clones in healthy individuals, determine if these clones are stable longitudinally, and determine if clonal mutations arise in long-lived hematopoietic stem and progenitor cells (HSPCs) or in more committed progenitors. We studied clonal hematopoiesis in longitudinally banked blood samples from middle-aged healthy participants in the Nurses' Health Study. We found clonal hematopoiesis, predominantly harboring mutations in *DNMT3A* and *TET2*, in 95% of individuals studied. Many clonal mutations were stable longitudinally and detected in both myeloid and lymphoid lineages, suggesting they arose in long-lived HSPCs.

4.2 Methods

4.2.1 Study Population

The Nurses' Health Study (NHS) began in 1976 with 121,701 female United States registered nurses age 30 to 55 years old who returned an enrollment questionnaire, which queried medical history, anthropometric measures, and lifestyle/environmental risk factors[168]. Since enrollment the participants have returned biennial follow-up questionnaires to update information on potential exposures and diagnoses of chronic disease. To date, follow-up rates have been consistently high (>90%). In 1989-1990, 32,826 women provided a heparinized whole blood sample by methods described previously[169]. In 2000-2001, 18,743 of the women who had provided a sample in 1989-1990 provided a second whole blood sample using the same protocol[170]. Briefly, participants willing to provide blood samples received kits that included all supplies necessary for their collection and overnight return (including a chill pack), and a brief questionnaire. Upon receipt, specimens were separated into plasma, buffy coat and red blood cell fractions and frozen in liquid nitrogen. Informed consent to participate in the NHS was implied by return of the enrollment and follow-up questionnaires; written informed consent was obtained for biomarker studies at time of blood collection.

Among women who provided blood samples in 1989-1990 and 2000-2001, we identified 20 with no history of cancer or other major chronic disease. De-identified aliquots from those 40 buffy coat samples were prepared and shipped overnight to Washington University for the detection of persistent rare hematopoietic clones harboring AML-associated somatic mutations as described below. Since each sample was de-identified and the capture space for targeted genomic DNA sequencing was not enough to enable individual identification (141 kb per person), the Washington University Human Research Protection Office deemed this study as non-human research.

4.2.2 Sample Preparation for Error-Corrected Sequencing

Genomic DNA was extracted from 50 uL of purified buffy coat from each sample using the QIAmp DNeasy Blood and Tissue Kit (Qiagen) with MinElute columns (Qiagen) instead of standard columns to facilitate elution in a lower volume (three 30 uL elutions). The concentration of extracted genomic DNA was measured using the Qubit dsDNA HS Assay Kit (Life Technologies). Enrichment of 568 amplicons in 54 genes (141 kb) commonly mutated in AML was performed using 250 ng of genomic DNA via the Illumina TruSight Myeloid Panel (Illumina). Technical replicates were prepared for each sample (80 libraries total). Following extension-ligation, the amplified fragments were eluted in 50 mM NaOH. Recovered fragments were amplified using the Q5 High-Fidelity 2x Master Mix (New England Biolabs) in a 75 uL reaction (37.5 uL 2x master mix, 20 uL DNA in 50 uM NaOH, 2 uL Tris-HCl pH 7.5, 0.4 uM i5/i7 primers). Illumina's standard i7 primers were used to enable sample multiplexing. The i5 primer was redesigned to contain a random 16 nucleotide index to facilitate error-corrected sequencing (Table 5.1). The following conditions were used for amplification: 98C for 30s; 6 cycles of 98C for 10s, 66C for 30s, 72C for 30s; 72C for 2m; hold 10C. The PCR reaction was purified using a modified Ampure bead (Beckman Coulter) cleanup to purify the amplified fragments (>400 bp). A modified poly-ethylene glycol (PEG) solution was made containing 382.5 uL 50% wt/vol PEG (Sigma), and 562.5 uL 5M NaCl and 555 uL ddH2O. 100 uL of beads were washed twice with ddH2O to remove the stock PEG solution. 150 uL of the modified PEG solution was added to the washed Ampure beads with the 75 uL PCR reaction and otherwise purified using the standard Ampure protocol. The fragments were eluted in 20 uL ddH_2O and the concentration of each library was quantified with Qubit (Life Technologies).

4.2.3 Quantification by ddPCR

Our goal was to generate each error-corrected sequencing (ECS) library from 4M uniquely tagged molecules. We quantified each library's concentration using the QX200 droplet digital PCR (ddPCR) platform (Bio-Rad). A 2 uL aliquot of each library was diluted 1000-fold and quantified in duplicate wells. Each well contained the following reaction mixture: 10 uL 2x EvaGreen 2x ddPCR master mix (Bio-Rad), 5 uL 1:1,000 diluted ECS library, 100 nM P5/P7 primers (Table 4.1), and ddH_2O added to 20 uL total. Droplets were generated using the standard Bio-Rad protocol. Amplification was completed using the following conditions: 95C for 5m; 40 cycles of 95C for 30s, 66C for 1m; 4C for 5m; 90C for 5m; 4C hold. With the calculated concentration, we aliquoted the appropriate volume of each library to introduce 4M molecules into the subsequent amplification step.

4.2.4 Amplification and Normalization

Following ddPCR quantification, 4M molecules for each library were amplified using Q5 High-Fidelity 2x Master Mix (New England Biolabs) using 1 uM P5/P7 primers (Table 4.1) in a 50 uL reaction under the following conditions: 98C for 30s; 16 cycles of 98C for 10s, 66C for 30s, 72C for 30s; 72C for 2m; 4C hold. The PCR reaction was purified using the modified Ampure bead cleanup. 100 uL of beads were washed twice with ddH2O and replaced with 100 uL of the modified PEG solution described above. The PCR reaction was then added to the mixture and purified using the standard protocol. The fragments were eluted in 20 uL ddH2O. A

2 uL aliquot of each library was diluted 10-fold and quantified on the Agilent 2200 Tape Station. Libraries were then pooled in equimolar groups of eight. Once pooled, each library was again quantified on the Tape Station and submitted for sequencing.

4.2.5 Sequencing

Each library was sequenced on the Illumina NextSeq platform using a 300 cycle high output kit as specified by the manufacturer. Approximately 5-10% of PhiX control DNA was spiked into each sequencing experiment. Each sequencing run contained roughly 400M paired-end 144 bp reads with corresponding 16bp unique molecular index (UMI) and 8bp sample-specific index sequences. Sequenced reads were demultiplexed by sample-specific index allowing for at most one mismatch in the index sequence (Table 4.2). Raw sequence reads were aligned to the PhiX genome using Bowtie 2[159]. Sequence reads that did not align to PhiX were retained for subsequent analysis (below).

4.2.6 Error Corrected Sequencing Analysis

The first 30 nucleotides of each sequenced read were hard clipped to remove the primer sequences from the TruSight Myeloid panel. Next, the sequenced read pairs tagged with the same random index sequence (originating from the same uniquely tagged DNA molecule) were aligned to each other to generate read families in a fashion similar to previously published methods[117,120,134,135]. Read families were required to have five or more reads sharing the same index sequence. Paired-end reads within a read family were error corrected to generate an error-corrected consensus sequence (ECCS) in a stepwise fashion. First, at every position, the nucleotides called by each sequence read were compared and a consensus nucleotide was called if there was at least 90% agreement between the reads. If there was less than 90% agreement, an N was called in the consensus sequence at that position. Errors that occurred during library preparation and sequencing were corrected or removed because they were not shared between different reads within a read family. Second, an ECCS was discarded if >10% of the 228 nucleotides comprising the paired-end read were reported as N nucleotides. ECCSs were then locally aligned to UCSC hg19/GRCh37 using Bowtie2 and realigned with GATK's Indel Realigner[171]. Next, the aligned ECCSs were processed with Mpileup using the parameters -BQ0 -d 10000000000000. This removed the coverage thresholds to ensure that all of the pileup output was returned regardless of variant allele fraction (VAF) or coverage. The parsed pileup output was further filtered to ignore positions with less than 1000x ECCS coverage or outside of the Illumina TruSight Myeloid target space. Additionally, germline variants identified by the 1000 Genomes Project above a 0.01 minor allele fraction were excluded from analysis.

We implemented a position-specific binomial error model to improve rare clonal single nucleotide variants (SNVs) calling as described previously[134]. For each sample, we generated a nucleotide position-specific error profile using all sequenced libraries that were not from the same individual. A variant was called if the binomial p-value was: a) less than 0.05 after Bonferroni correction, b) the variant was observed in at least 5 ECCSs, c) the VAF was greater than 0.0001, and d) the variant was identified with criteria a-c in at least two replicates from one of the two time points. Likely clonal SNVs (<0.2 VAF) were reported and annotated using Annovar[172] with the COSMIC 68[37] and 1000 Genomes (Oct 2014 release)[173] databases. The amino acid substitutions were predicted based on the canonical transcript reported in the GENCODE (v22)[174] as retrieved from the UCSC Table Browser[175].

We identified potential insertion/deletion (indel) events using VarScan 2[176], from the filtered Mpileup output (described above), with the following parameters --min-coverage 1000 --min-reads2 5 --min-var-freq 0.001 --strand-filter 0 --output-vcf 1. We filtered out single nucleotide indels in homopolymer runs at least four nucleotides long and indels that were observed in multiple samples to remove technical artifacts in the variant calling. We reported likely clonal indels (<0.2 VAF) that were detected in at least two replicates from one of the collection time points. Reported indels were annotated with Annovar[172] as described previously.

4.2.7 Droplet Digital PCR Validation

We validated 21 SNVs and 1 indel using the droplet digital PCR (ddPCR) probe assay (Bio-Rad)[154]. Probes were designed by Bio-Rad based on MIQE guidelines for quantitative digital PCR[177]. All reagents were purchased from Bio-Rad. For each sample and control, 45 ng of genomic DNA was aliquoted per well of generated droplets. We generated between 8 and 32 wells of droplets for each validation experiment, depending on the expected VAF for the assayed mutation. Each control sample was assayed with the same number of wells as the corresponding sample. Droplets were generated on the QX200 Droplet Generator (Bio-Rad) and assayed on the QX200 droplet reader (Bio-Rad) using the standard protocol[154]. The VAF was estimated from droplets lacking the reference allele and the Poisson-estimated number of singleton droplets as described previously[134].

4.2.8 Flow cytometry

Cells were sorted from buffy coat samples using a Sony iCyt Synergy SY3200 BSC 17-color, 5-laser cell sorter (Sony Biotechnology Inc.) and analyzed with FlowJo (Treestar) using standard protocols (Figure 4.1). Cells were stained with the following antibodies (BioLegend): CD45 (BV-421), CD33 (APC), CD19 (FITC), CD3 (PE-CY7) per the manufacturer’s instructions. Variants were detected in purified cell populations using the ddPCR assay described previously.

4.2.9 Data Availability

The sequencing data have been deposited into the NCBI Sequence Read Archive under accession number SRP078948. All other relevant data are included in the article or supplementary files, or available from the authors upon request.

4.3 Results

4.3.1 Variant Quantification in Rare Hematopoietic Clones

We obtained paired buffy coat blood samples, banked approximately 10 years apart, from 20 healthy participants in the Nurses' Health Study (Table 4.3)—a cohort of 121,701 female registered nurses longitudinally studied since 1976[168–170]. The median ages at sample collection were 56.6 and 68.1 years old. This facilitated the investigation of benign clonal hematopoiesis in younger individuals previously thought to only rarely harbor hematopoietic clones[144–147,149]. To identify hematopoietic clones, we combined ECS with targeted-capture for 568 amplicons in 54 genes frequently mutated in AML (Methods)[117,120,134,135]. This enabled us to sequence a tractable subset of the genome, while still querying loci associated with clonal hematopoiesis and AML. Samples were prepared and sequenced in duplicate. We generated an average of 47.7 million paired-end sequencing reads, which yielded an average of 3.4 million error-corrected consensus sequences (ECCSs), per library (Table 4.2).

We modeled position-specific errors in the ECCSs using binomial statistics to identify clonal mutations (Methods). We identified 109 clonal single nucleotide variants (SNVs) in at least one time point below 0.2 VAF in 95% (19/20) of individuals. We detected 1-17, mostly exonic, SNVs per individual at 0.0003-0.1451 (median 0.0024) VAF (Figure 4.2a, Table 4.4). Of note, the median VAF we observed was 10-fold less than the detection limit governing previous studies of clonal hematopoiesis[144–146]. Separately, we identified 9 clonal insertion/deletion variants (indels) in 6 individuals (Table 4.5). Indels were identified by ECCS coverage alone, as indel errors were not appropriately modeled by the same statistical framework implemented to identify SNVs.

We were initially concerned that most of the identified rare variants were artifacts introduced during library preparation or sequencing. We first determined that SNV calls were not biased by coverage per amplicon (Figure 4.3) or by the number of bases captured per gene (Figure 4.4). Next, we validated these findings using droplet digital PCR (ddPCR)—an orthogonal non-sequencing-based technique for VAF quantification. We designed ddPCR assays for 21 SNVs that had been previously observed in malignancy[37] and for one indel (Figure 4.5). The VAFs measured by ECS and ddPCR were highly correlated (R^2=0.98, Figure 4.6, Table 4.6), consistent with the previously observed accuracy of ECS[134].

We next compared the mutation profile observed in these rare hematopoietic clones to previous findings in CHIP and AML. We detected 88 exonic clonal SNVs with 58 missense SNVs, 17 nonsense SNVs, 9 synonymous SNVs, 3 splicing SNVs, and 1 SNV in a 3'UTR (Figure 4.2b). While exonic variants were detected in only 18 of the 54 genes in the panel, 64% (56/88) occurred in the epigenetic regulators *DNMT3A* and *TET2* (Figure 4.2c). We frequently detected multiple *DNMT3A* and *TET2* mutations in the same individual, which were not necessarily in the same clone (Figure 4.7). The *DNMT3A* SNVs were predominantly nonsense mutations in the 5' end of the gene or missense mutations in the three functional domains (Figure 4.8). For comparison, *TET2* SNVs were primarily missense mutations in the functional domains or nonsense mutations throughout the gene (Figure 4.9), consistent with previous observations of AML[178]. While less prevalent, intronic clonal SNVs were observed in 12 genes with 29% (6/21) detected in *DNMT3A* and 5% (1/21) detected in *TET2* (Figure 4.10, Figure 4.11). The most common type of exonic substitution was the cytosine to thymine (C to T) transition (Figure 4.2d), as previously observed in CHIP[144–146]. Conversely, intronic SNVs were evenly distributed across substitution types.

4.3.2 Temporal Stability of Rare Clonal Mutations

We characterized the temporal stability of these clones by tracking clonal mutations longitudinally within our 20 study participants. Variants were called independently from paired samples banked approximately 10 years apart (Figure 4.12). Of the 109 clonal SNVs identified, 27.5% (30/109) were detected at both time points, 13.8% (15/109) were detected at only the first time point, and 58.7% (64/109) were detected at only the second time point (Table 4.4). The stability of VAFs observed here was consistent with previous observations at higher VAFs in a few instances of CHIP[145]. The presence of the same clonal mutations longitudinally suggested that these mutations arose in long-lived HSPCs or committed progenitors.

4.3.3 Clonal Mutations Arise in Hematopoietic Stem and Progenitors

To further elucidate the cell of origin for clonal hematopoiesis, we sorted 26 samples from 13 individuals into B lymphocyte (CD45+CD33-CD19+), T lymphocyte (CD45+CD33-CD3+) and myeloid (CD45+CD33+) compartments using flow cytometry (Figure 4.1). While cell recovery was variable per sample, we observed the same clonal SNVs in both myeloid and lymphoid compartments in 10/13 individuals (Figure 4.13, Table 4.7). Frequently, the VAF measured in the bulk sample was approximately equal to the VAF measured in each compartment. These observations were unlikely to have arisen due to contamination, given that variants were often detected at similar VAFs in different sorted compartments.

4.4 Discussion

These findings suggest that clonal hematopoiesis harboring mutations in AML-associated genes is nearly ubiquitous (95%) in 50-70 year olds—an age group in which previous studies identified hematopoietic clones in only 5% of individuals[144–147]. Clonal mutations were detected in both the myeloid and lymphoid compartments in samples banked a decade apart in these healthy individuals, clearly demonstrating that these mutations arose in long-lived HSPCs. However, these clonal mutations conferred only a modest proliferation advantage, as most clonal mutations were rare (median 0.0024 VAF) and stable longitudinally. One possible explanation for these observations was that these mutations, often in epigenetic regulators, augmented self-renewal capacity without a concomitant increase in proliferation. This hypothesis may also explain why HSPC number increases and quiescence decreases as a function of age[179,180]. As HSPCs gradually senesce throughout life, the acquisition of these mutations may allow benign clonal hematopoiesis to maintain ostensibly normal blood production years after it would otherwise decline[148]. This hypothesis is supported by work in mice demonstrating that *DNMT3A* loss-of-function mutations in HSCs are associated with an increase in HSC self-renewal without increasing proliferation[181]. Comparably, *TET2* loss-of-function mutations in mice increase HSC self-renewal and proliferation[182].

While *DNMT3A* mutations are frequently observed in CHIP and AML[56,144–147], we observed a different distribution of *DNMT3A* mutations, specifically at the arginine 882 (R882) residue. Previous studies showed that mutations in *DNMT3A* R882 comprised approximately two-thirds of total *DNMT3A* mutations in AML[45] and one-sixth of *DNMT3A* mutations in CHIP[145,146]. However, we observed only a single *DNMT3A* R882H variant. These findings suggest that *DNMT3A* R882 mutations potently drive clonal expansion, explaining their prior detection in common CHIP clones (median 0.11 VAF)[145] and their rarity in these lower frequency clones.

The detection limit of ECS was approximately 1:10,000 cells. Thus, given an estimated 11,000 hematopoietic stem cells (HSCs) in adults, of which only a fraction actively contribute to hematopoiesis at any given time[164], we expected to observe unique somatic mutations marking each active HSC (a random distribution of synonymous and nonsynonymous mutations throughout the 54 AML-associated genes captured). Instead, over half of the detected mutations were in *DNMT3A* or *TET2*. This observation alone could have occurred if *DNMT3A* and *TET2* were hotspots of somatic mutation. However, 89% (78/88) of the detected exonic mutations were nonsynonymous, truncating or splicing mutations. Given this skew towards presumed functional mutations, it was more likely that these hematopoietic clones were enriched by selection.

Due to technical limitations of our methods, we likely underreported the number of clonal mutations present in each individual. Specifically, we likely underreported the number of C to T (G to A) substitutions present in these rare hematopoietic clones due to the stringency of the binomial variant calling strategy and the background rate of cytosine deamination, which is a predominant artifact of error-corrected sequencing[117,122,135]. Here 38/109 substitutions identified were C to T (G to A) substitutions. Conversely, in previous studies of CHIP and AML, C to T (G to A) substitutions comprised approximately 50-60% of identified substitutions[62,145,146]. Additionally, the binomial statistical framework underreported hotspot mutations occurring in multiple individuals. However, in our raw data we only observed a single likely instance of an uncalled hotspot mutation—a *DNMT3A* R882H variant in individual 5 observed at a lower VAF than the variant reported in individual 13. Additionally, we could not routinely co-localize mutations within the same hematopoietic clone. However, we co-localized mutations in three instances where they co-occurred in the same sequenced reads (PID 2, *TET2* R1216G/A1217A; PID 13, *DNMT3A* G498V/C494F; PID 14, *KRAS* A66A/S65S). Future adaptations of this technology could address these limitations by targeting a larger capture panel and implementing single-cell sequencing approaches.

In summary, we demonstrate that clonal hematopoiesis, originating in long-lived HSPCs, is far more common than previously thought in healthy middle-aged adults. Despite its prevalence, clonal hematopoiesis shares many mutations with AML, raising additional questions regarding the sequence of mutation acquisition and cooperating events necessary for malignant transformation. Furthermore, in previous studies of CHIP the detection of a hematopoietic clone (at any age) was associated with an 11-13 fold increased risk of developing a hematological malignancy[145,146]. These earlier findings suggested that CHIP was comparable to monoclonal gammopathy of undetermined significance and monoclonal B-cell lymphocytosis, which are benign clonal proliferative conditions that occasionally progress to hematological malignancy[143,149,167]. Conversely, our findings suggest that clinically silent clonal hematopoiesis is present in almost all individuals by middle age, and that progression to hematological malignancy is exceptionally rare. Given the current public interest in precision medicine[183], these findings have practical implications for sequencing-based screening of nascent malignancy or recurrence. Future research must focus on reliably distinguishing benign clonal hematopoiesis, however rare, from malignant clonal hematopoiesis that could drive transformation and relapse. This imperative extends to sequencing-based non-invasive screening[184], which will require even finer discrimination between nascent malignancy and benign clonal expansion.

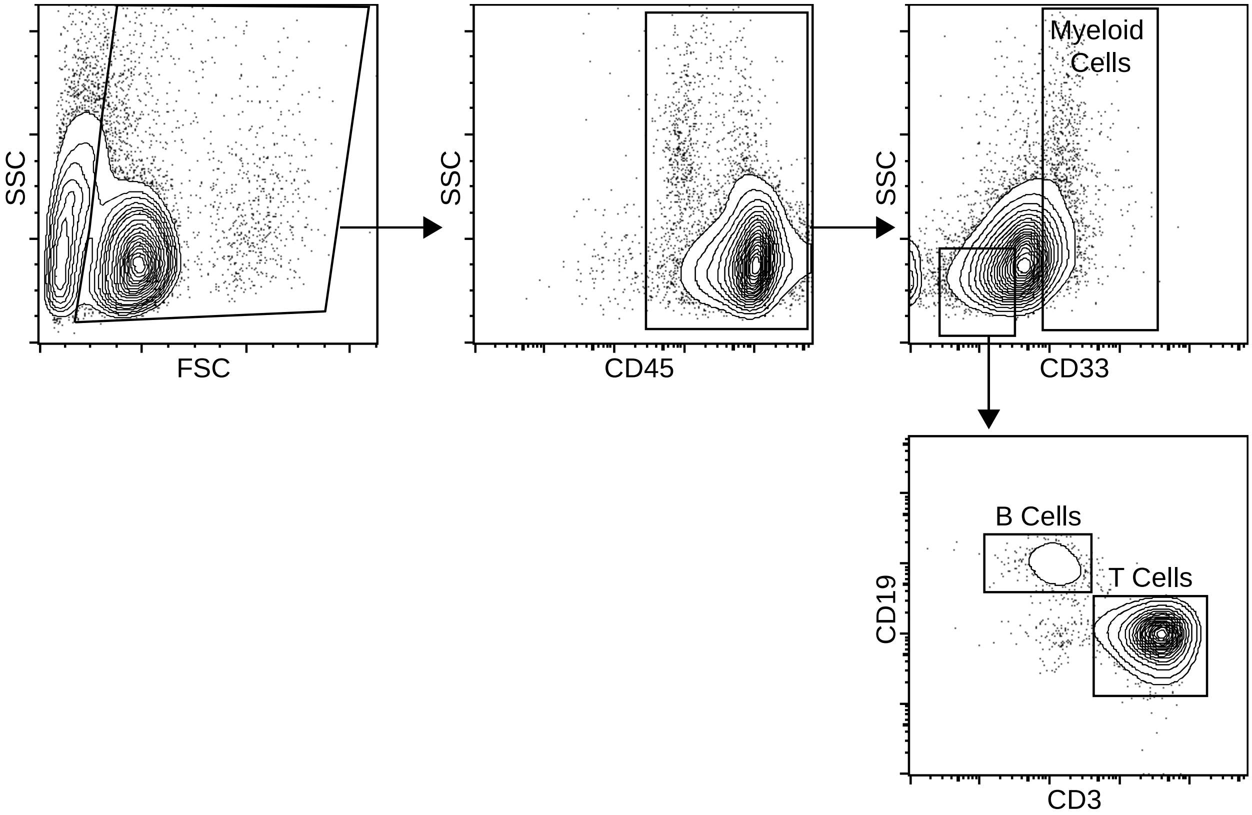

Figure 4.1 Representative flow cytometry gating strategy. These results were generated from NHS participant 5, time point 1 to isolate B lymphocytes, T lymphocytes and myeloid cells.

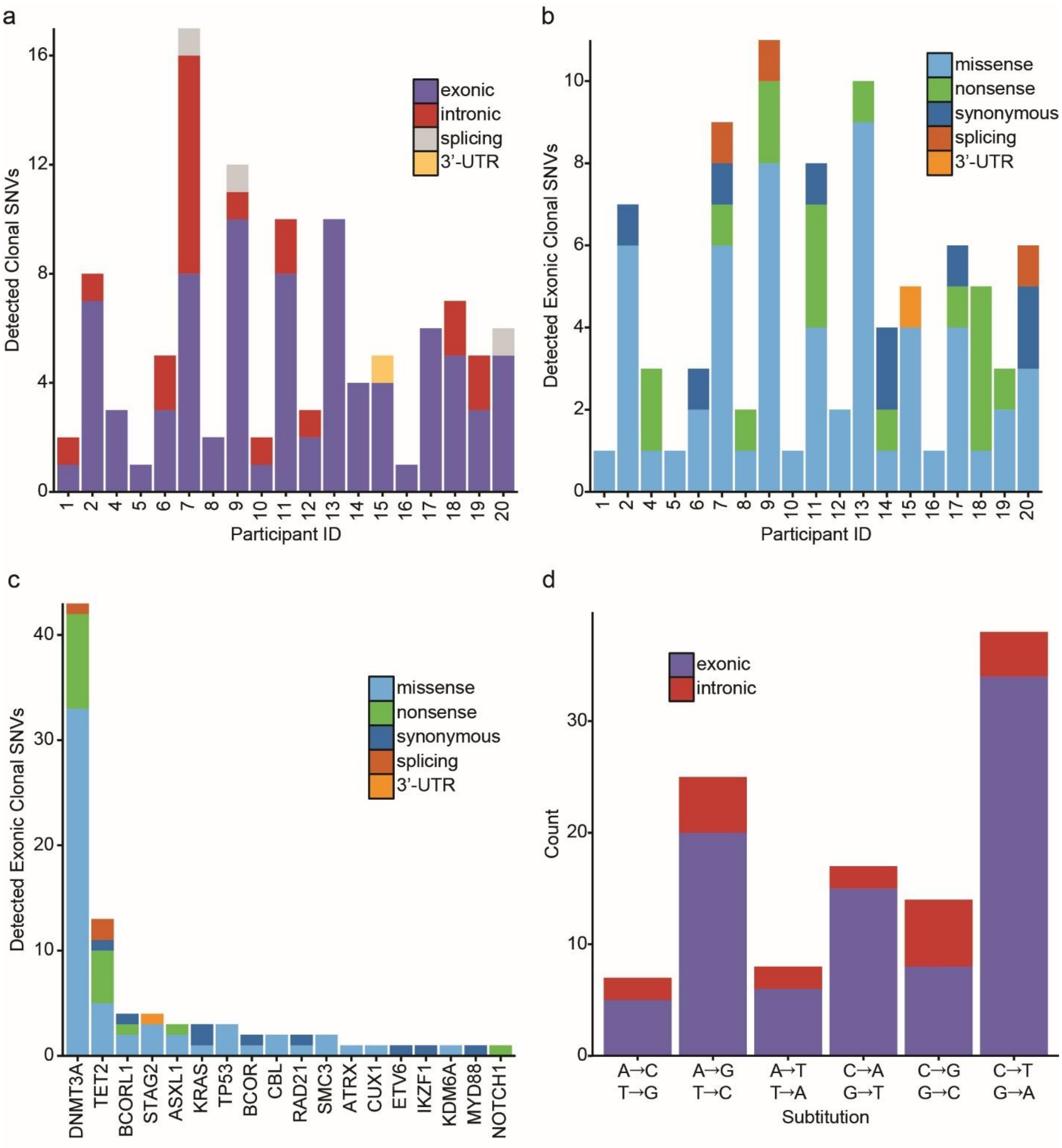

Figure 4.2 Number and characteristics of clonal SNVs detected by ECS in the peripheral blood of healthy adult nurses. a) Clonal SNVs detected in each individual, color-coded by annotation. b) Exonic clonal SNVs detected in each individual, color-coded by predicted effect. c) Detected exonic clonal SNVs organized by gene, color-coded by predicted effect. d) Distribution of substitution types observed in clonal SNVs.

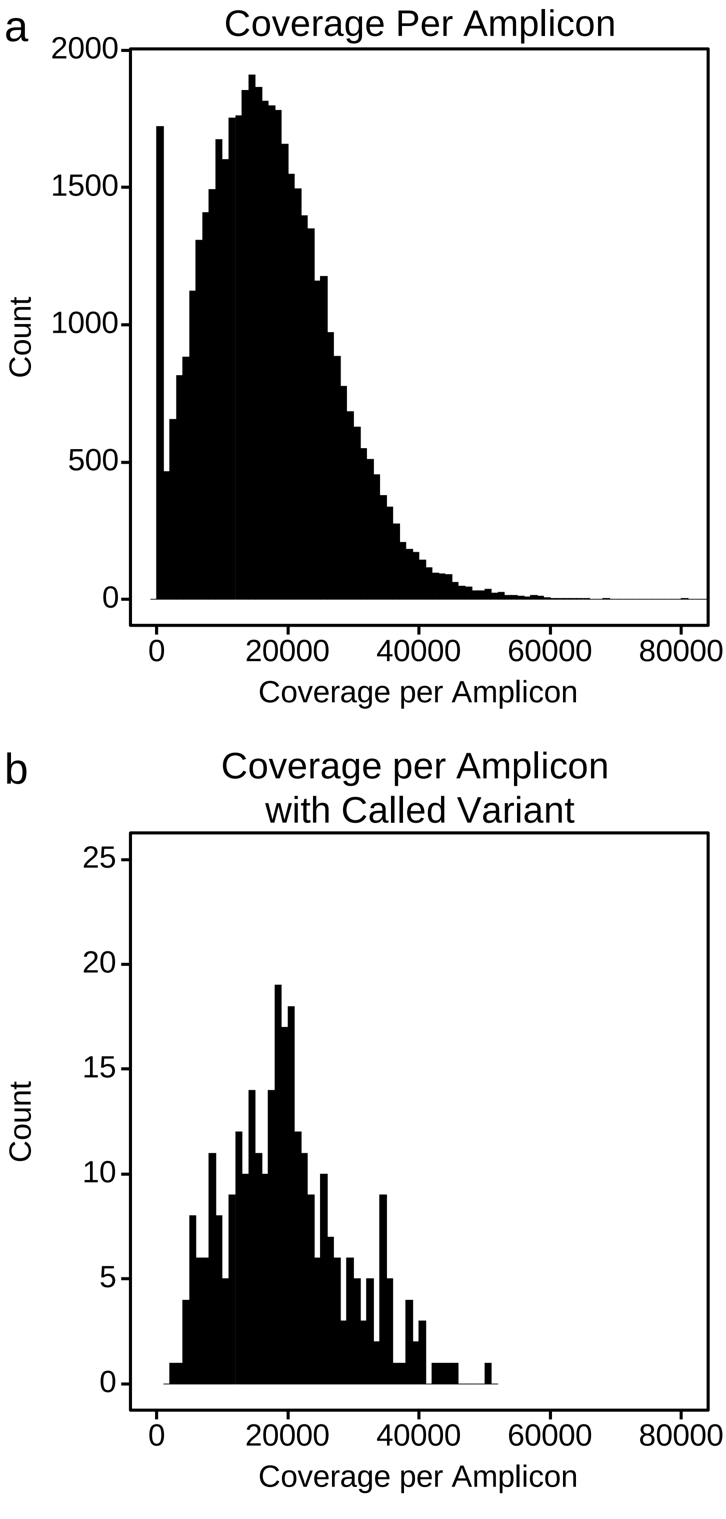

Figure 4.3 Coverage per amplicon for error-corrected sequencing experiments. Error-corrected consensus sequence (ECCS) coverage was calculated for each of the 568 amplicons in the capture panel. a) Histogram of ECCS coverage for all amplicons. b) Histogram of ECCS coverage in amplicons in which a variant was detected.

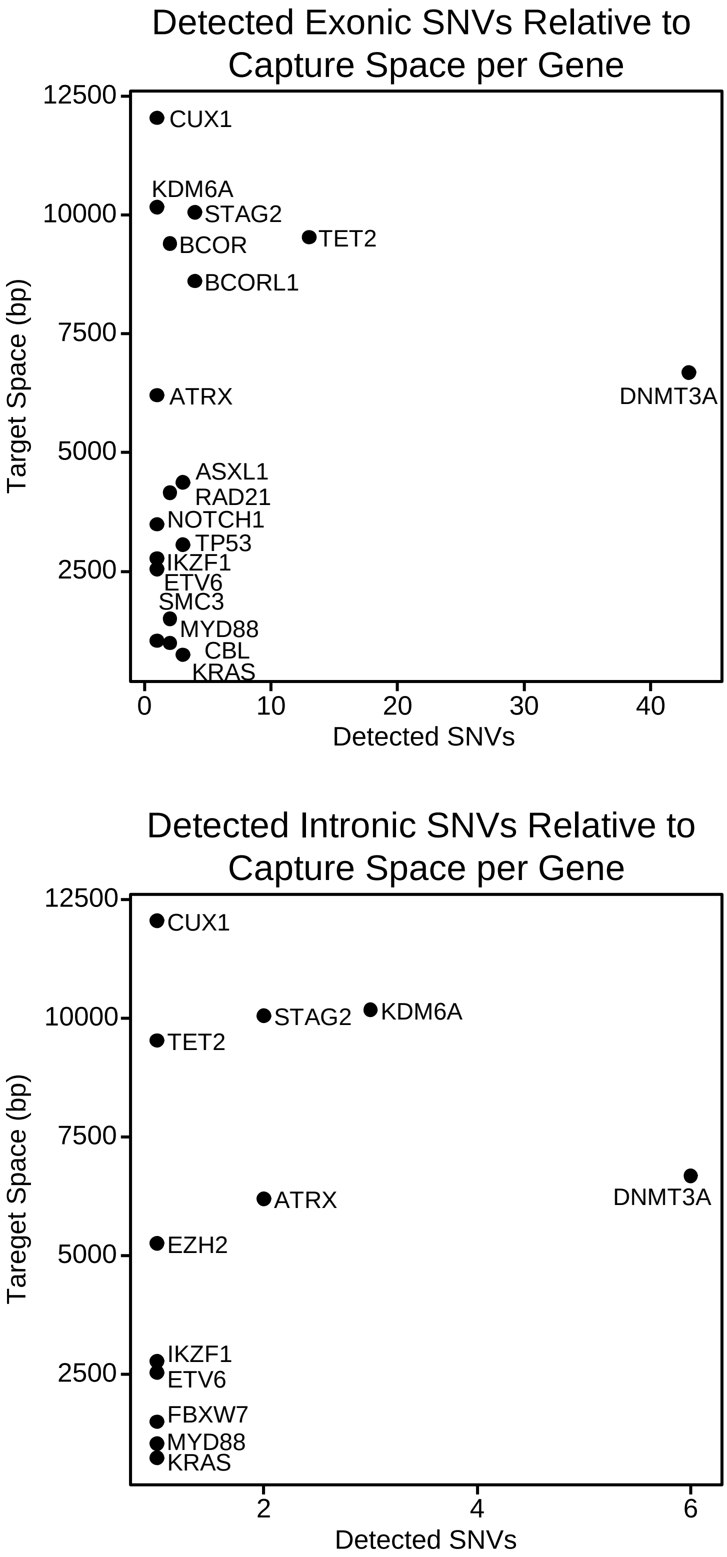

Figure 4.4 Number of mutations detected compared to target space per gene. Mutations detected in exons (top panel) and introns (bottom panel) were plotted relative to the capture space (bp = base pairs) targeting that gene in the panel.

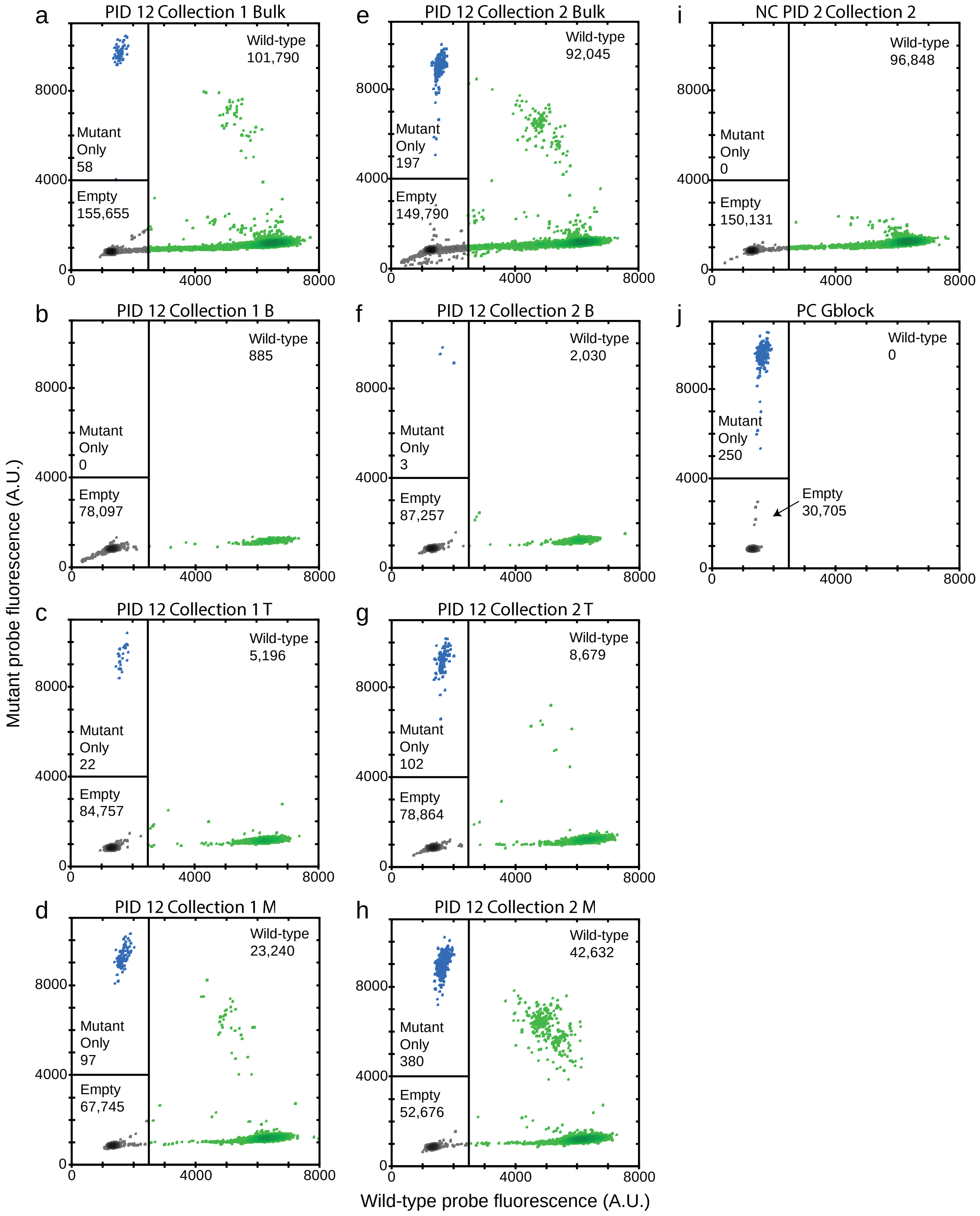
a
PID 12 Collection 1 Bulk
Wild-type
101,790
Mutant
Only
58
Empty
155,655
e
PID 12 Collection 2 Bulk
Wild-type
92,045
Mutant
Only
197
Empty
149,790
i
NC PID 2 Collection 2
Wild-type
96,848
Mutant
Only
0
Empty
150,131
b
PID 12 Collection 1 B
Wild-type
885
Mutant
Only
0
Empty
78,097
f
PID 12 Collection 2 B
Wild-type
2,030
Mutant
Only
3
Empty
87,257
j
PC Gblock
Wild-type
0
Mutant
Only
250
Empty
30,705
c
PID 12 Collection 1 T
Wild-type
5,196
Mutant
Only
22
Empty
84,757
g
PID 12 Collection 2 T
Wild-type
8,679
Mutant
Only
102
Empty
78,864
d
PID 12 Collection 1 M
Wild-type
23,240
Mutant
Only
97
Empty
67,745
h
PID 12 Collection 2 M
Wild-type
42,632
Mutant
Only
380
Empty
52,676
0
4000
8000
Mutant probe fluorescence (A.U.)
Wild-type probe fluorescence (A.U.)

Figure 4.5 Representative droplet digital PCR (ddPCR) results. These results originated from NHS participant 12 for the detected *DNMT3A* G543A clonal variant. The wild-type probe intensity in arbitrary units (A.U.) was plotted relative to the *DNMT3A* G543A (mutant) probe intensity for each droplet. a-d) Variant quantification at the first time point for a) all cells; b) B lymphocytes; c) T lymphocytes; and d) myeloid cells. e-h) Variant quantification at the second time point for e, all cells; f) B lymphocytes; g) T lymphocytes; and h) myeloid cells. i) The *DNMT3A* G543A variant was not detected in the negative control sample from participant 2, time point 2. j) Only DNMT3A G543A positive (or empty) droplets were detected in the gblock positive control.

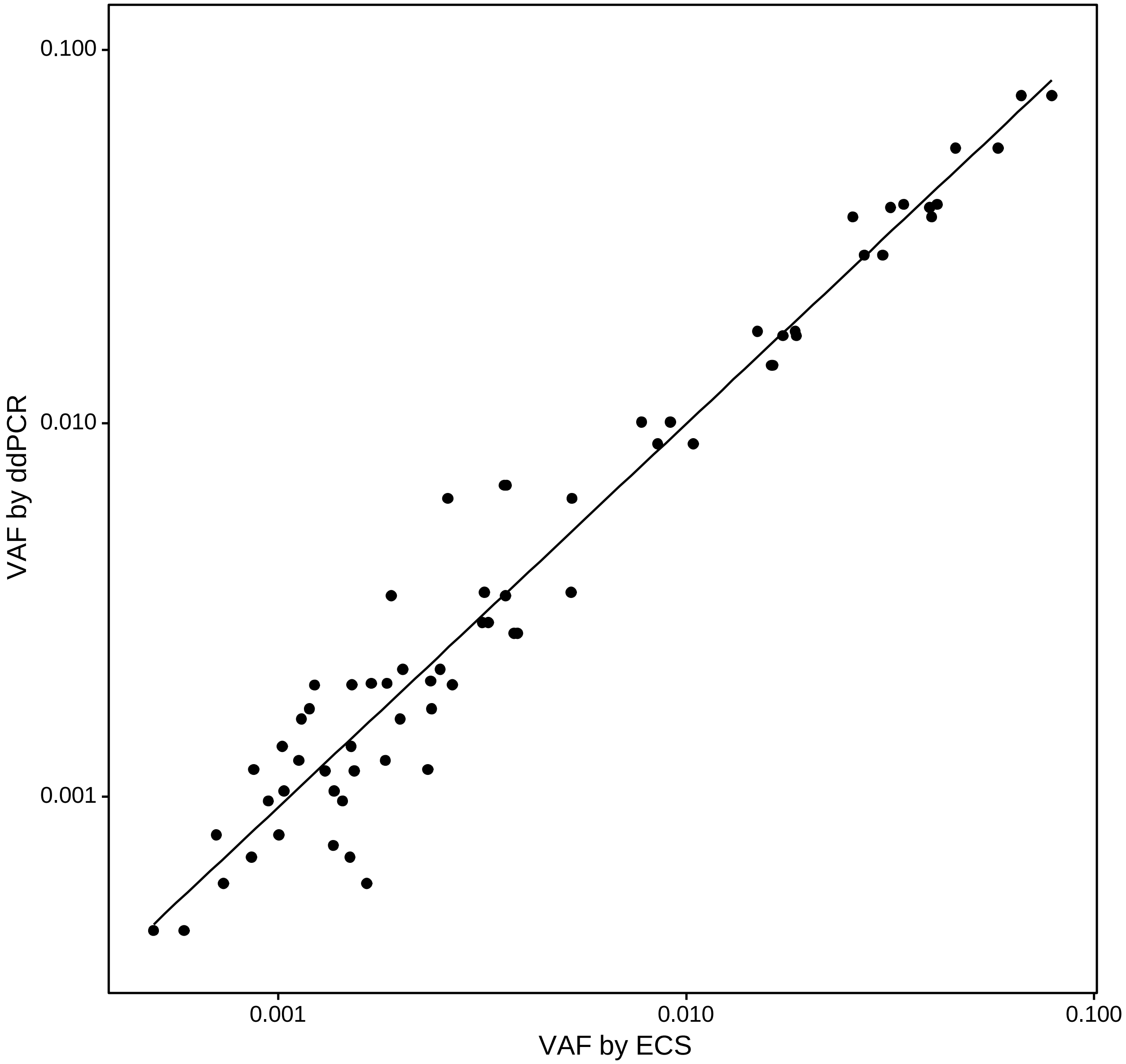

Figure 4.6 Concordance of variant allele fraction (VAF) measured by error-corrected sequencing (ECS) and droplet digital PCR (ddPCR). Several mutations identified by ECS were verified using ddPCR. The variant allele fractions (VAFs) identified ECS and ddPCR were highly correlated (R^2=0.98).

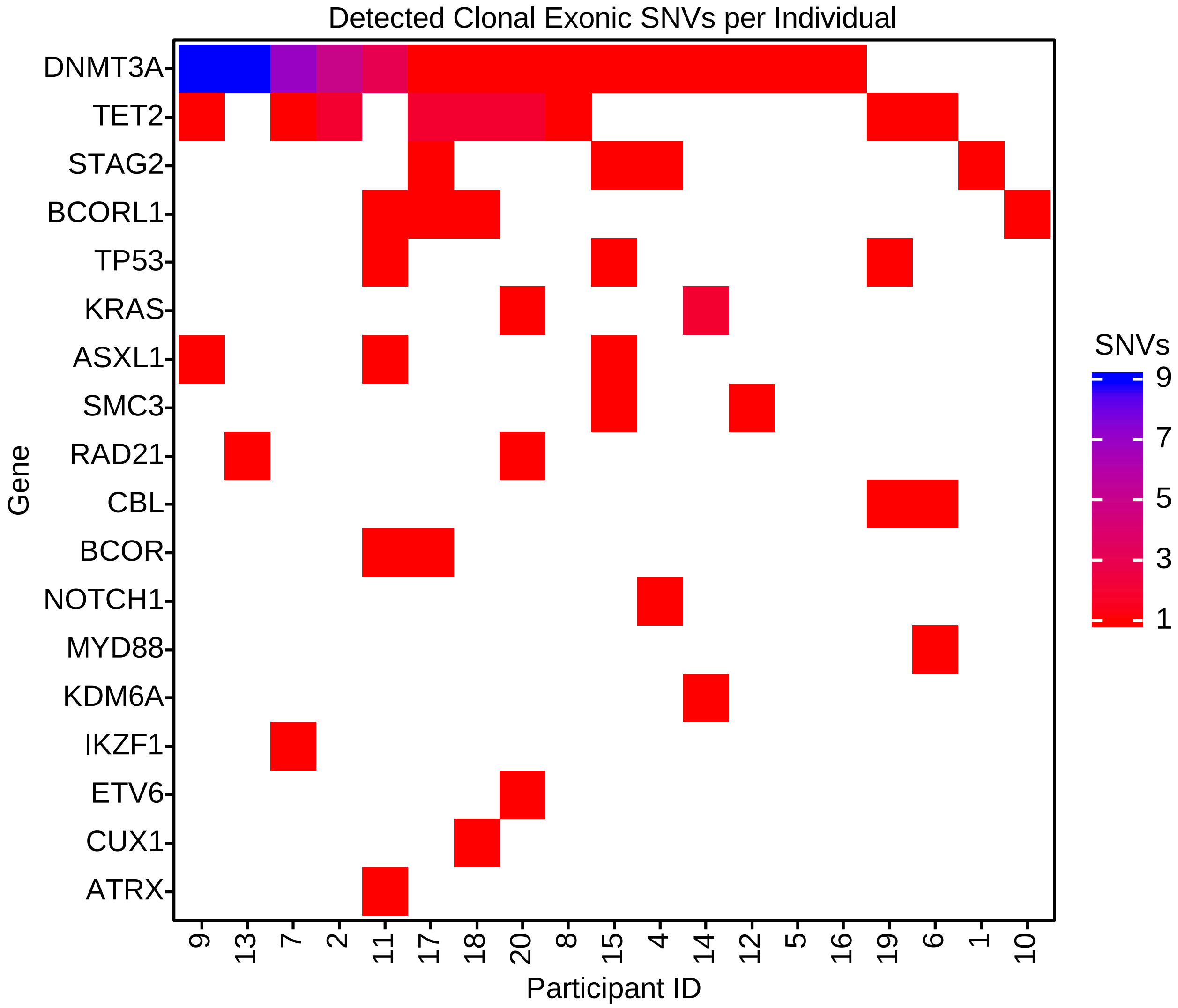

Figure 4.7 Heat map depicting the number of exonic single nucleotide variants (SNVs) detected in each gene per study participant.

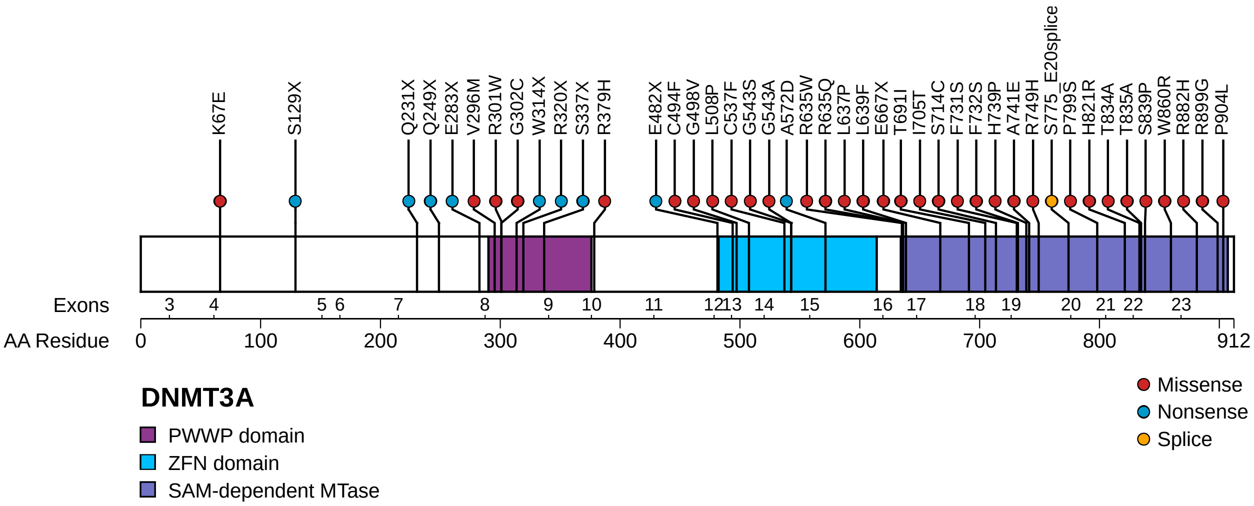

Figure 4.8 Detected exonic clonal single nucleotide variants (SNVs) in *DNMT3A*. The detected SNVs were predominantly nonsense mutations (blue) in the first half of the gene or missense mutations (red) in the three functional domains—a proline-tryptophan-tryptophan-proline (PWWP) chromatin targeting domain, a zinc finger nuclease (ZFN) domain and a S-adenosylmethionine (SAM) dependent methyltransferase (MTase) domain.

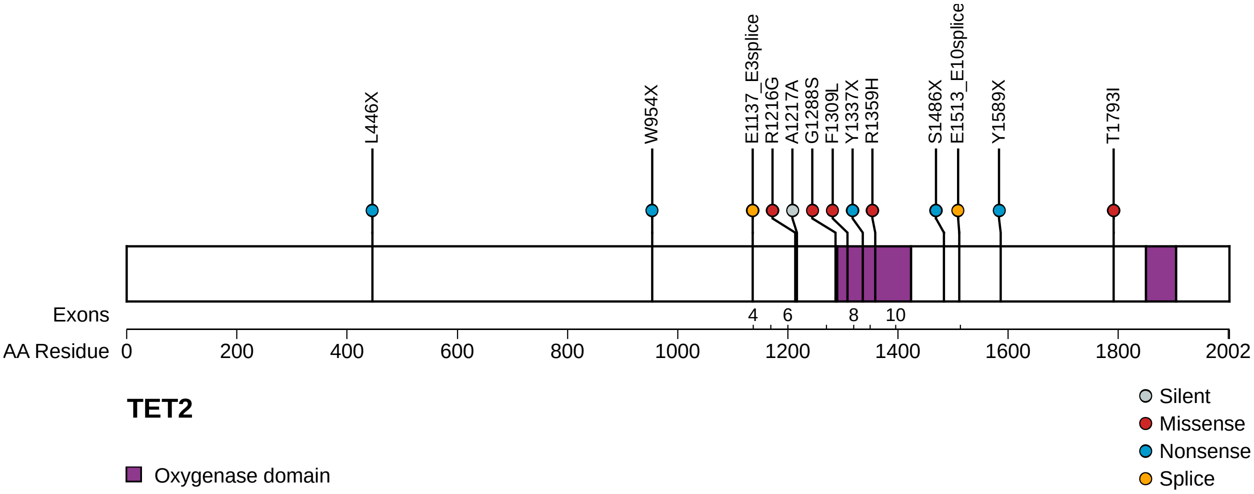

Figure 4.9 Detected exonic clonal single nucleotide variants in *TET2*.

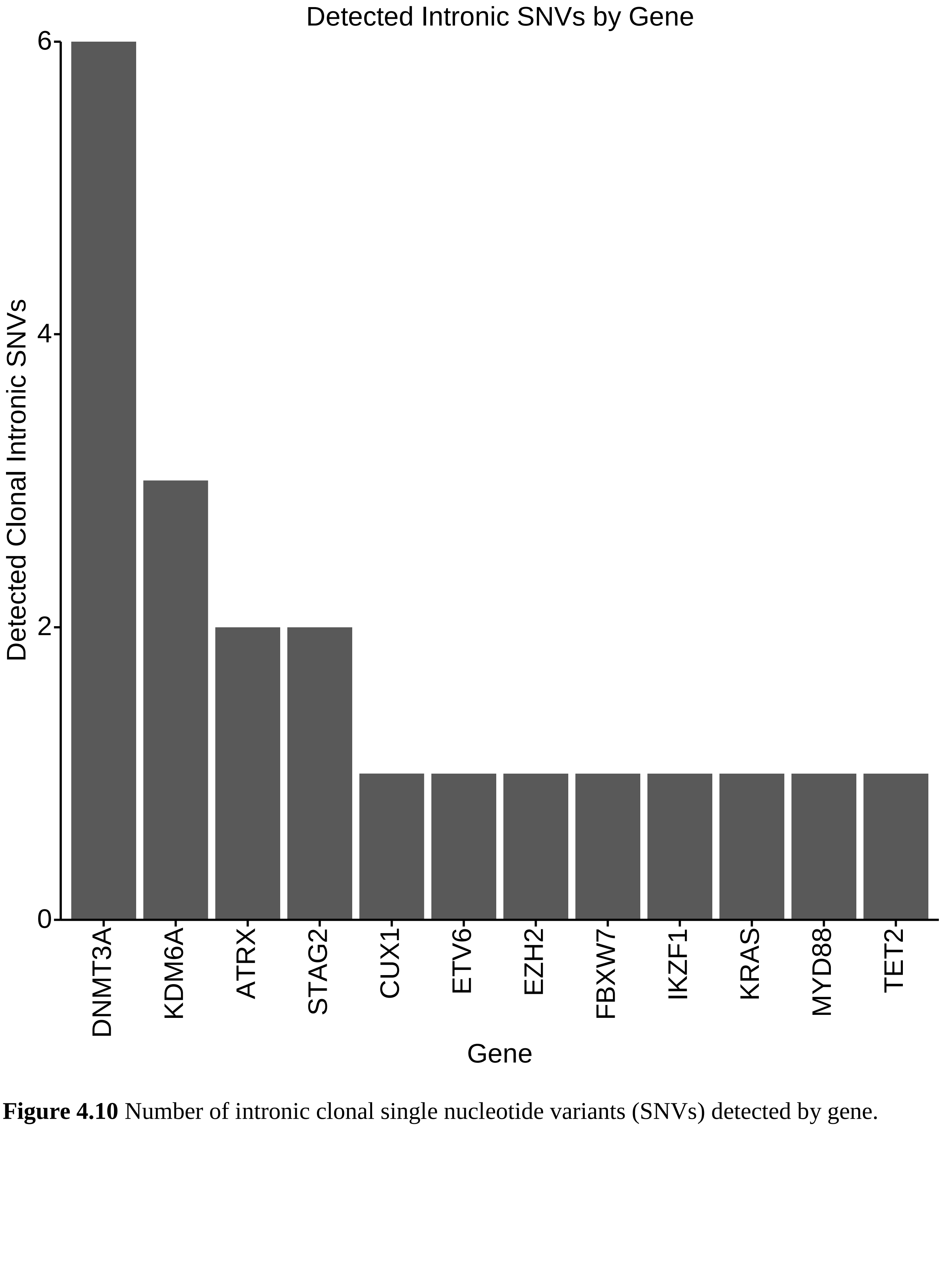

Figure 4.10 Number of intronic clonal single nucleotide variants (SNVs) detected by gene.

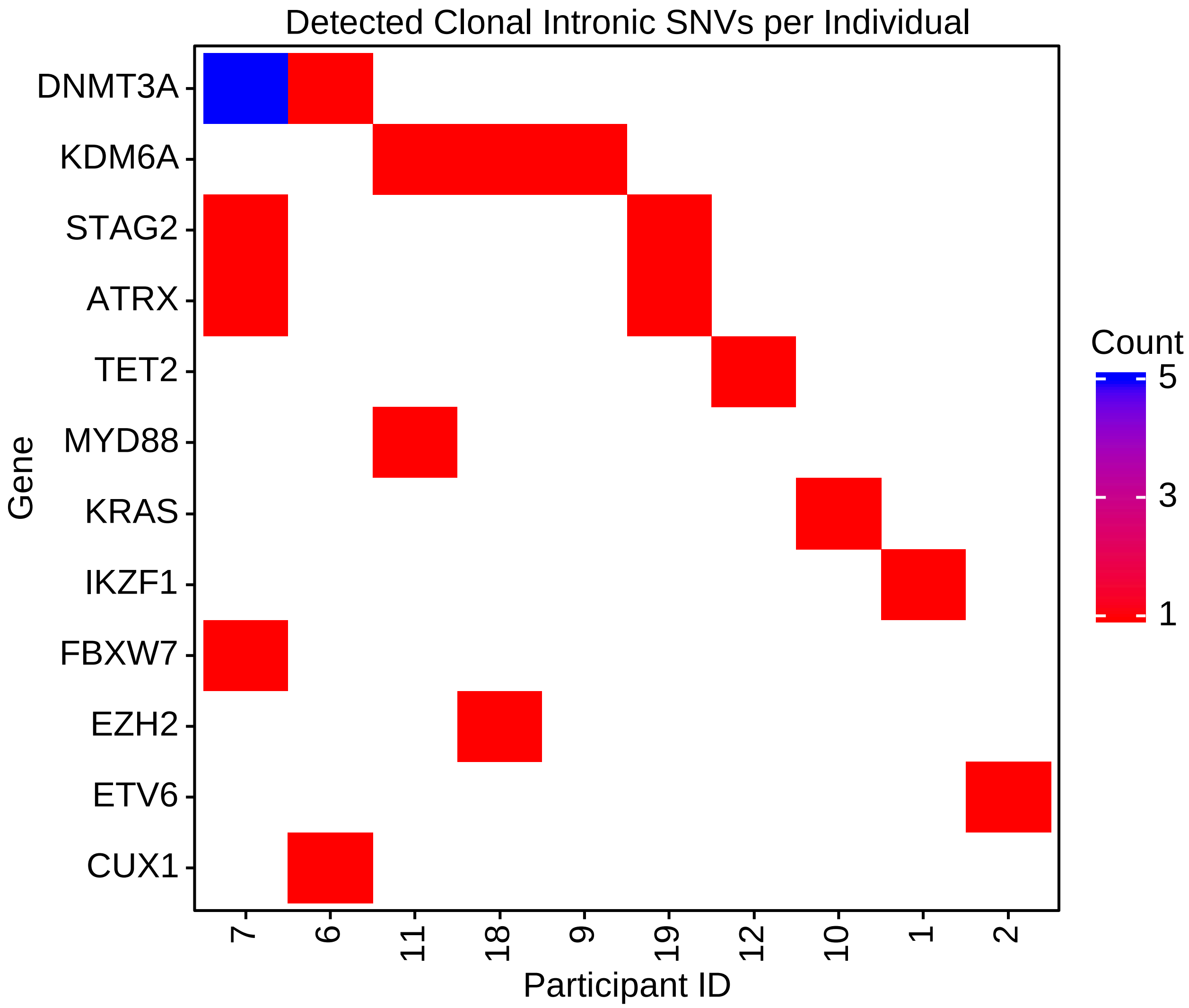

Figure 4.11 Heat map depicting the number of intronic single nucleotide variants (SNVs) detected in each gene per individual.

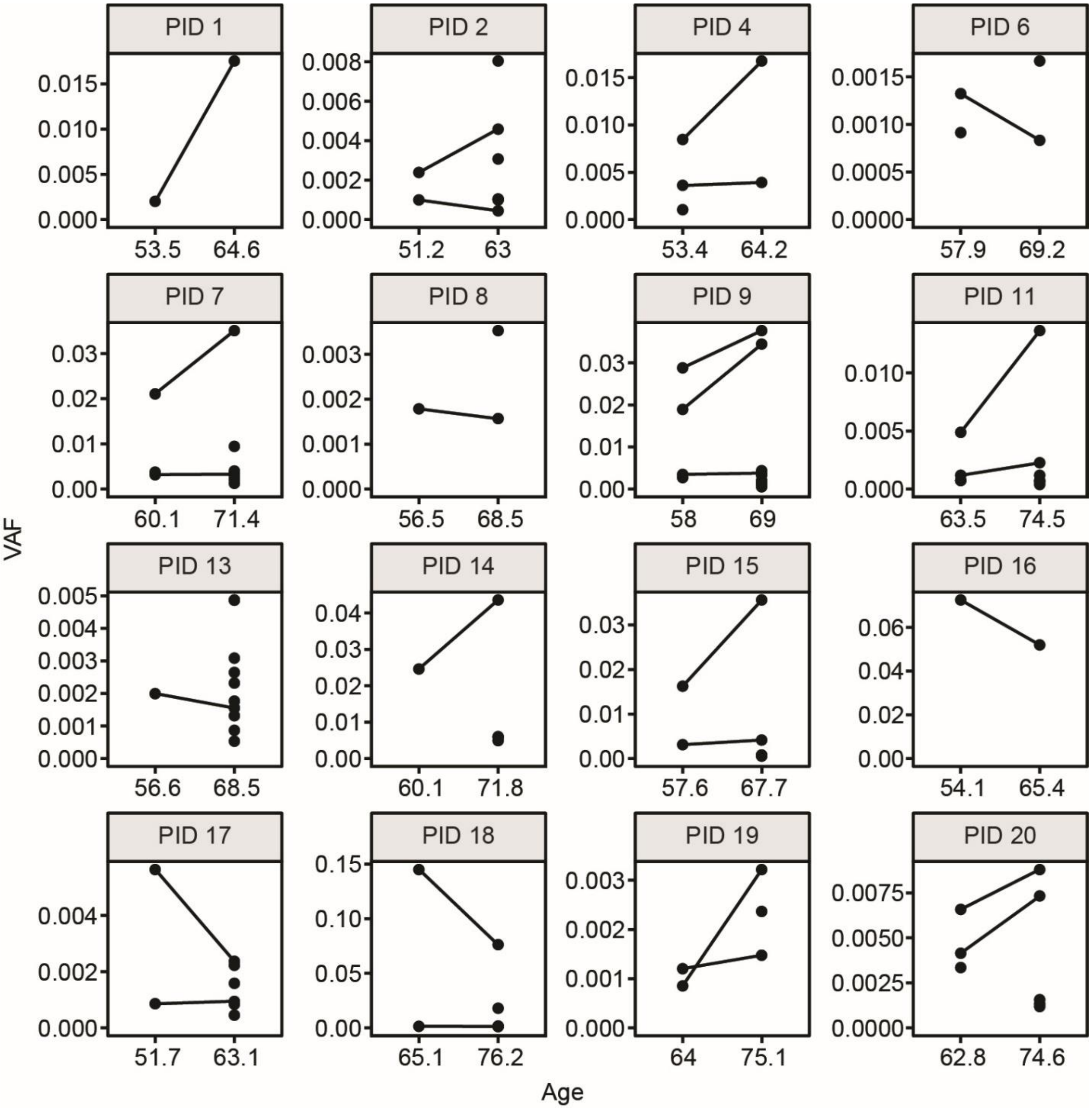

Figure 4.12 Longitudinal detection of clonal SNVs in NHS participants. Clonal SNVs were detected by ECS in both time points for 16/20 NHS participants. For each participant ID (PID), the VAF measured by ECS was plotted relative to the age at sample collection. Variants detected in both time points were connected with a line.

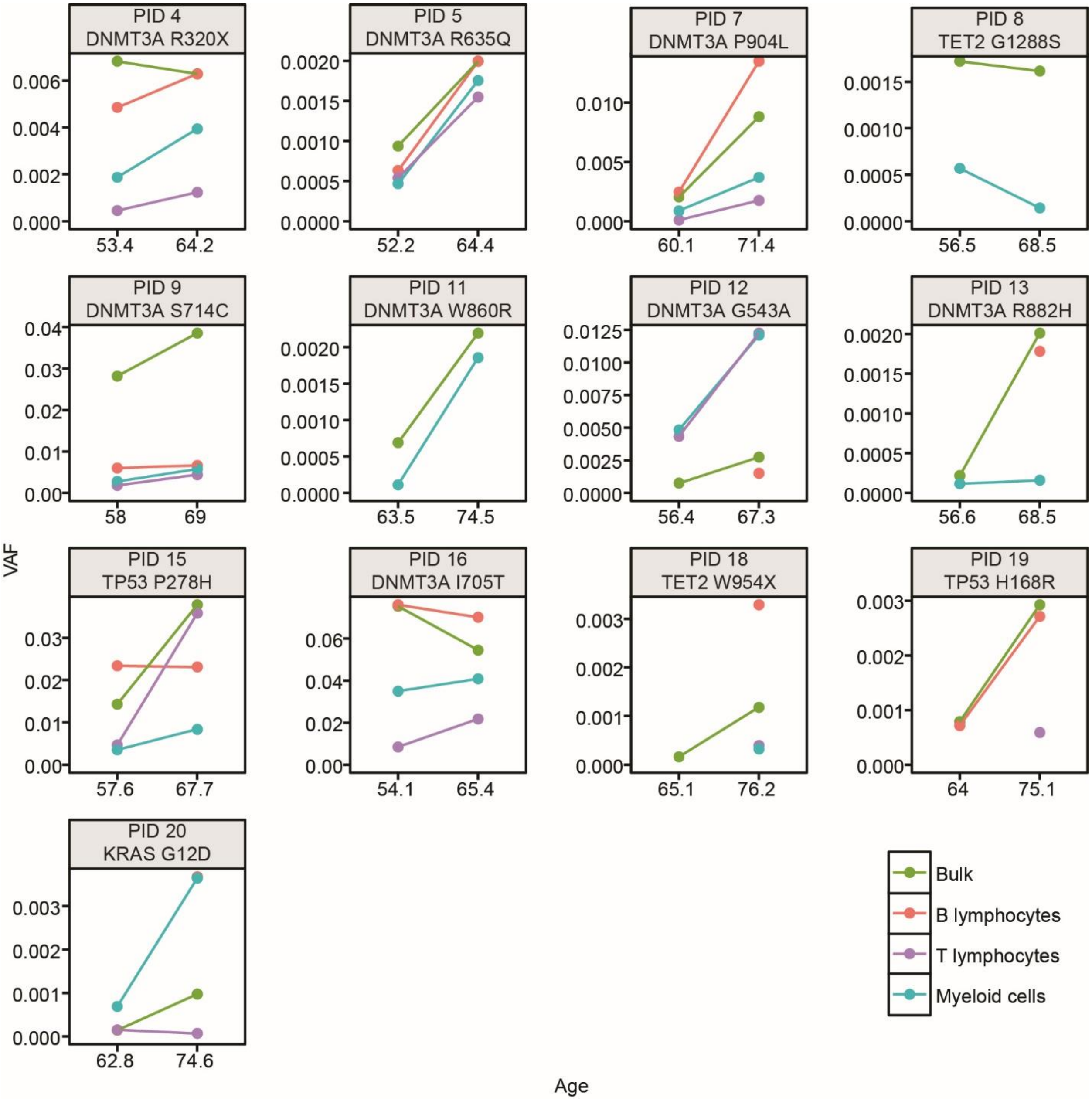

Figure 4.13 Hematopoietic compartment-specific detection of clonal SNVs in NHS participants. Paired buffy coat samples from 13 individuals were sorted into B lymphocyte (pink), T lymphocyte (purple), and myeloid (blue) compartments using flow cytometry. For each NHS participant (PID), a single SNV, detected by ECS, was selected for compartment-specific quantification by ddPCR. Variants detected in both time points were connected with a line. The VAF measured by ddPCR in the bulk sample (green) was included for comparison.

Table 4.1 Primer sequences for library preparation.

Primer Name	Sequence
i5 16N Random	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNNNNNNNNACACTCTTTCCCTACACGACGCTCTTCCGATCT
P5	AATGATACGGCGACCACCGA
P7	CAAGCAGAAGACGGCATACGA

Table 4.2 Sequenced reads and error-corrected consensus sequences (ECCS) generated for each library.

	Collection 1 Replicate 1		Collection 1 Replicate 2		Collection 2 Replicate 1		Collection 2 Replicate 2	
PID	**Raw Reads**	**ECCS**	**Raw Reads**	**ECCS**	**Raw Reads**	**ECCS**	**Raw Reads**	**ECCS**
1	38,483,255	3,333,626	33,448,218	2,804,567	35,427,539	2,987,274	33,392,978	2,784,229
2	39,268,072	3,318,173	41,984,812	3,558,516	35,157,811	3,027,607	33,657,209	2,860,180
3	32,603,819	2,581,039	36,107,671	2,959,152	48,998,142	3,584,046	40,599,238	3,291,215
4	30,932,163	2,212,764	30,623,846	2,501,632	39,579,433	3,254,544	48,529,452	3,503,765
5	35,011,143	2,727,030	34,151,207	2,411,821	52,302,285	3,759,106	55,049,072	4,017,037
6	34,207,169	2,863,690	35,084,657	2,946,669	50,852,817	3,682,303	48,351,486	3,514,019
7	41,658,678	2,663,917	42,508,068	2,714,869	45,885,262	3,233,300	44,468,353	3,548,708
8	44,771,597	2,734,288	41,632,517	2,528,357	50,072,031	3,399,553	50,378,471	3,698,270
9	39,449,116	2,531,229	41,067,140	2,599,127	60,014,462	4,197,532	50,347,145	3,993,077
10	40,492,765	2,554,060	38,729,489	2,400,500	59,870,612	4,034,423	58,962,293	3,996,550
11	48,940,303	3,684,038	44,034,692	3,456,949	64,520,183	4,096,893	56,501,287	3,797,404
12	57,115,177	4,245,185	48,446,875	3,692,857	61,813,583	4,322,748	59,452,070	4,110,288
13	39,368,839	3,059,660	41,269,631	3,343,408	59,327,495	4,213,628	52,689,305	4,173,008
14	38,837,743	3,076,601	37,306,017	2,976,419	60,366,370	4,053,326	58,109,532	4,501,213
15	54,605,075	3,407,283	44,547,457	2,490,226	58,101,101	3,908,542	51,539,869	4,104,014
16	52,226,742	2,986,829	60,744,391	3,907,372	45,532,881	3,414,608	60,632,143	4,027,633
17	57,985,852	4,079,429	51,835,232	3,373,746	55,355,241	3,721,052	59,501,527	4,096,858
18	54,185,217	3,337,380	54,495,083	3,388,504	57,117,288	4,147,180	58,064,485	3,999,074
19	52,213,946	2,947,031	51,028,264	3,343,682	53,983,868	3,723,161	48,268,843	3,796,565
20	51,521,626	3,376,026	41,417,619	2,806,990	58,651,106	4,092,482	56,692,728	3,793,691

Table 4.3 Age at sample collection for each NHS participant.

Participant ID	Collection 1 Age	Collection 2 Age
1	53.5	64.6
2	51.2	63.0
3	52.3	64.4
4	53.4	64.2
5	52.2	64.4
6	57.9	69.2
7	60.1	71.4
8	56.5	68.5
9	58.0	69.0
10	54.7	66.9
11	63.5	74.5
12	56.4	67.3
13	56.6	68.5
14	60.1	71.8
15	57.6	67.7
16	54.1	65.4
17	51.7	63.1
18	65.1	76.2
19	64.0	75.1
20	62.8	74.6

Table 4.4 Clonal SNVs detected by error-corrected sequencing.

PID	Chr	Pos	Ref	Alt	Gene	AA Change	COSMIC	VAF1.1	VAF1.2	VAF2.1	VAF2.2
1	7	50444517	T	C	IKZF1	intronic		0.0009	-	0.0010	0.0029
1	X	123185174	G	T	STAG2	V376L		0.0028	0.0012	0.0180	0.0170
2	2	25463271	G	T	DNMT3A	A741E		0.0011	-	0.0010	0.0010
2	2	25463277	T	G	DNMT3A	H739P		0.0017	0.0003	0.0004	0.0005
2	2	25466793	A	G	DNMT3A	L637P		0.0014	0.0034	0.0035	0.0056
2	2	25468153	A	G	DNMT3A	L508P		-	0.0012	0.0076	0.0085
2	2	25470573	G	A	DNMT3A	R301W		-	-	0.0026	0.0035
2	4	106164778	C	G	TET2	R1216G		-	-	0.0011	0.0010
2	4	106164783	T	A	TET2	A1217A		-	-	0.0011	0.0010
2	12	11803160	G	C	ETV6	intronic		0.0012	0.0006	-	-
4	2	25470516	G	A	DNMT3A	R320X	133724	0.0036	0.0036	0.0052	0.0026
4	9	139391179	G	A	NOTCH1	Q2338X		0.0011	0.0010	-	-
4	X	123184056	G	T	STAG2	R305L	254953	0.0091	0.0078	0.0149	0.0185
5	2	25466799	C	T	DNMT3A	R635Q	1583088	-	-	0.0015	0.0027
6	2	25471183	C	G	DNMT3A	intronic		0.0016	0.0015	-	-
6	3	38181952	C	T	MYD88	I192I		-	-	0.0018	0.0016
6	4	106180899	T	G	TET2	F1309L		0.0014	0.0013	0.0009	0.0008
6	7	101843537	T	G	CUX1	intronic		0.0040	0.0042	0.0017	0.0052
6	11	119148929	A	G	CBL	I383M		0.0014	0.0004	-	-
7	2	25457176	G	A	DNMT3A	P904L	87007	0.0024	-	0.0085	0.0104
7	2	25458673	T	C	DNMT3A	T834A		0.0029	0.0035	0.0026	0.0040
7	2	25462012	G	A	DNMT3A	P799S		-	-	0.0023	0.0048
7	2	25463372	G	A	DNMT3A	intronic		0.0042	0.0039	0.0034	-
7	2	25463384	G	A	DNMT3A	intronic		0.0066	0.0082	0.0029	-
7	2	25463385	C	G	DNMT3A	intronic		-	-	0.0066	0.0087
7	2	25463387	C	G	DNMT3A	intronic		0.0067	0.0079	0.0032	-
7	2	25463389	G	A	DNMT3A	intronic		0.0068	0.0077	0.0038	-
7	2	25464441	G	A	DNMT3A	T691I		-	-	0.0036	0.0025
7	2	25464514	C	A	DNMT3A	E667X		-	-	0.0011	0.0032
7	2	25466788	G	A	DNMT3A	L639F		0.0216	0.0206	0.0407	0.0295
7	2	25467449	C	T	DNMT3A	G543S		-	-	0.0048	0.0033
7	4	106158509	G	A	TET2	splicing	87117	-	-	0.0010	0.0015
7	4	153249632	T	C	FBXW7	intronic		0.0004	0.0003	-	-
7	7	50367256	C	T	IKZF1	S21S		0.0035	0.0041	0.0021	-
7	X	76874262	A	T	ATRX	intronic		-	-	0.0005	0.0003
7	X	123199914	G	T	STAG2	intronic		-	-	0.0008	0.0009
8	2	25468919	C	A	DNMT3A	E482X		-	-	0.0031	0.0040
8	4	106180834	G	A	TET2	G1288S	110780	0.0012	0.0024	0.0020	0.0011
9	2	25459821	T	C	DNMT3A	H821R		-	0.0022	0.0017	0.0025

PID	Chr	Pos	Ref	Alt	Gene	AA Change	COSMIC	VAF1.1	VAF1.2	VAF2.1	VAF2.2
9	2	25462086	T	C	DNMT3A	splicing		0.0026	-	0.0015	0.0015
9	2	25463247	C	T	DNMT3A	R749H		0.0203	0.0177	0.0333	0.0356
9	2	25463541	G	C	DNMT3A	S714C	87011	0.0273	0.0303	0.0412	0.0341
9	2	25466800	G	A	DNMT3A	R635W	87012	0.0019	0.0036	-	0.0012
9	2	25467160	G	T	DNMT3A	A572D		-	-	0.0010	0.0021
9	2	25469632	C	T	DNMT3A	R379H		-	-	0.0029	0.0057
9	2	25470588	C	T	DNMT3A	V296M		-	-	0.0035	0.0032
9	2	25471016	G	A	DNMT3A	Q249X		0.0014	-	0.0013	0.0013
9	4	106193995	C	G	TET2	S1486X	211625	-	-	0.0005	0.0006
9	20	31023091	A	G	ASXL1	N859S		0.0041	0.0029	0.0044	0.0032
9	X	44733267	T	G	KDM6A	intronic		-	-	0.0017	0.0013
10	12	25380459	G	C	KRAS	intronic		-	-	0.0010	0.0009
10	X	129162659	C	G	BCORL1	H1376Q		0.0049	0.0023	0.0018	-
11	2	25458595	A	G	DNMT3A	W860R	231568	0.0009	0.0015	0.0020	0.0025
11	2	25470914	C	A	DNMT3A	E283X		0.0042	0.0056	0.0129	0.0143
11	2	25505372	G	C	DNMT3A	S129X		-	-	0.0007	0.0005
11	3	38182245	T	A	MYD88	intronic		-	-	0.0019	0.0022
11	17	7577129	A	G	TP53	F270S	11305	-	-	0.0007	0.0016
11	20	31023606	A	T	ASXL1	K1031X		0.0007	0.0008	-	-
11	X	39933358	G	A	BCOR	A414V		-	-	0.0006	0.0009
11	X	44733249	G	T	KDM6A	intronic		-	-	0.0009	0.0008
11	X	76814150	A	T	ATRX	F2165Y		0.0007	0.0007	-	-
11	X	129148158	A	G	BCORL1	L470L		-	-	0.0003	0.0005
12	2	25467448	C	G	DNMT3A	G543A	256033	0.0014	-	0.0038	0.0039
12	4	106155048	C	G	TET2	intronic		-	-	0.0010	0.0020
12	10	112342324	C	T	SMC3	S243F		-	-	0.0014	0.0018
13	2	25457192	G	C	DNMT3A	R899G		0.0025	0.0015	0.0013	0.0018
13	2	25457242	C	T	DNMT3A	R882H	52944	-	-	0.0018	0.0017
13	2	25458658	A	G	DNMT3A	S839P		0.0003	-	0.0011	0.0006
13	2	25463298	A	G	DNMT3A	F732S		0.0015	-	0.0030	0.0032
13	2	25467466	C	A	DNMT3A	C537F		-	-	0.0019	0.0034
13	2	25468183	C	A	DNMT3A	G498V		0.0009	-	0.0037	0.0060
13	2	25468195	C	A	DNMT3A	C494F		0.0009	-	0.0037	0.0060
13	2	25470532	C	T	DNMT3A	W314X		-	-	0.0021	0.0026
13	2	25470570	C	A	DNMT3A	G302C		-	-	0.0016	0.0010
13	8	117859842	T	C	RAD21	Y598C		-	-	0.0005	0.0006
14	2	25470464	G	C	DNMT3A	S337X		0.0252	0.0240	0.0423	0.0448
14	12	25380260	T	G	KRAS	A66A		-	-	0.0063	0.0055
14	12	25380263	A	G	KRAS	S65S		-	-	0.0064	0.0056
14	X	44911015	C	T	KDM6A	A239V		-	-	0.0043	0.0056
15	2	25458670	T	C	DNMT3A	T835A		-	-	0.0005	0.0005

PID	Chr	Pos	Ref	Alt	Gene	AA Change	COSMIC	VAF1.1	VAF1.2	VAF2.1	VAF2.2
15	10	112342321	T	C	SMC3	L242P		0.0028	0.0034	0.0036	0.0047
15	17	7577105	G	T	TP53	P278H	43755	0.0162	0.0163	0.0317	0.0395
15	20	31024085	G	T	ASXL1	R1190S		-	-	0.0011	0.0005
15	X	123234540	T	A	STAG2	UTR3		-	-	0.0007	0.0006
16	2	25463568	A	G	DNMT3A	I705T	1583102	0.0663	0.0788	0.0457	0.0582
17	2	25463301	A	G	DNMT3A	F731S		0.0012	-	0.0019	0.0026
17	4	106156436	T	G	TET2	L446X		-	-	0.0007	0.0009
17	4	106197045	C	T	TET2	T1793I		-	-	0.0013	0.0019
17	X	39932643	G	A	BCOR	P652P		0.0055	0.0058	0.0019	0.0029
17	X	123224536	A	T	STAG2	K1130I		0.0008	0.0009	0.0009	0.0010
17	X	129173203	G	A	BCORL1	G1522S		-	-	0.0004	0.0004
18	2	25471070	G	A	DNMT3A	Q231X		0.0015	0.0014	0.0019	0.0009
18	4	106157961	G	A	TET2	W954X	87110	-	-	0.0023	0.0009
18	4	106182972	T	A	TET2	Y1337X	87145	-	-	0.0015	0.0013
18	7	101840496	G	A	CUX1	R602H		0.1479	0.1424	0.0828	0.0693
18	7	148515272	A	G	EZH2	intronic		-	-	0.0013	0.0013
18	X	44938634	A	G	KDM6A	intronic		0.0138	0.0121	0.0146	0.0137
18	X	129149098	C	T	BCORL1	R784X	1319521	-	-	0.0186	0.0173
19	4	106196434	T	G	TET2	Y1589X		-	0.0021	0.0023	0.0025
19	11	119148922	G	A	CBL	C381Y	34073	0.0010	0.0014	0.0011	0.0018
19	17	7578427	T	C	TP53	H168R	43545	0.0007	0.0010	0.0033	0.0032
19	X	76813170	A	G	ATRX	intronic		-	-	0.0007	0.0010
19	X	123179344	C	T	STAG2	intronic		0.0190	0.0160	0.0202	0.0222
20	2	25505559	T	C	DNMT3A	K67E		0.0052	0.0080	0.0076	0.0100
20	4	106190798	G	A	TET2	R1359H		0.0035	0.0032	-	-
20	4	106194076	G	A	TET2	splicing		-	-	0.0015	0.0017
20	8	117878873	C	A	RAD21	V32V		0.0043	0.0040	0.0075	0.0071
20	12	12022854	C	A	ETV6	V320V		-	-	0.0010	0.0016
20	12	25398284	C	T	KRAS	G12D	521	-	-	0.0009	0.0014

Table 4.5 Clonal insertion/deletion variants detected by error-corrected sequencing. We reported the variant allele fraction (VAF) for each variant detected in two replicates from at least one of the two time points.

PID	Chr	Start	End	Ref	Alt	Gene	AA	VAF1.1	VAF1.2	VAF2.1	VAF2.2
7	2	25463381	25463381	-	GTG	DNMT3A	intronic	0.0068	0.0070	0.0030	-
7	4	106155858	106155861	CAGT	-	TET2	N253fs	0.0063	0.0063	0.0133	0.0126
7	4	106164895	106164895	-	A	TET2	Y1255_G1256delinsX	0.0072	0.0052	0.0074	0.0090
9	2	25463567	25463567	A	-	DNMT3A	I705fs	0.0331	0.0229	0.0601	0.0801
9	2	25467528	25467547	AGCAGCGGGAAGGGTCAGAA	-	DNMT3A	intronic	-	-	0.0038	0.0041
11	2	25468168	25468168	-	T	DNMT3A	T503fs	-	-	0.0030	0.0031
15	8	117862892	117862895	TCTC	-	RAD21	E528fs	0.0068	0.0065	0.0142	0.0128
18	X	123179310	123179318	ATTAATTTT	-	STAG2	intronic	0.0277	0.0255	0.1193	0.1207
20	4	106190864	106190864	C	-	TET2	A1381fs	0.0052	0.0032	0.0400	0.0256

Table 4.6 Summary of droplet digital PCR validation experiments. The variant allele fraction (VAF) determined by error-corrected sequencing (ECS) was included for comparison. For each experiment, a control sample was selected where the variant of interest was not observed in the ECS data. For the control samples, the same number of genome equivalents were analyzed as the experimental sample.

PID	Gene	AA	ECS VAF 1.1	ECS VAF 1.2	ddPCR VAF 1	ECS VAF 2.1	ECS VAF 2.2	ddPCR VAF 2	Control	Control VAF
4	DNMT3A	R320X	0.0036	0.0036	0.0068	0.0052	0.0026	0.0063	1.2	0.0000
4	STAG2	R305L	0.0091	0.0078	0.0101	0.0149	0.0185	0.0176	1.1	0.0000
5	DNMT3A	R635Q	-	-	0.0009	0.0015	0.0027	0.0020	15.1	0.0000
7	DNMT3A	P904L	0.0024	-	0.0020	0.0085	0.0104	0.0088	3.1	0.0000
7	TET2	splice	-	-	0.0009	0.0010	0.0015	0.0014	3.2	0.0000
8	TET2	G1288S	0.0012	0.0024	0.0017	0.0020	0.0011	0.0016	2.1	0.0000
9	DNMT3A	S714C	0.0273	0.0303	0.0282	0.0412	0.0341	0.0385	2.2	0.0000
9	DNMT3A	R635W	0.0019	0.0036	0.0035	-	0.0012	0.0020	10.2	0.0000
9	TET2	S1486X	-	-	-	0.0005	0.0006	0.0004	6.1	0.0000
11	TP53	F270S	-	-	0.0005	0.0007	0.0016	0.0006	2.1	0.0000
11	DNMT3A	W860R	0.0009	0.0015	0.0007	0.0020	0.0025	0.0022	6.2	0.0000
12	DNMT3A	G543A	0.0014	-	0.0007	0.0038	0.0039	0.0027	2.2	0.0000
13	DNMT3A	R882H	-	-	0.0002	0.0018	0.0017	0.0020	14.1	0.0000
15	TP53	P278H	0.0162	0.0163	0.0143	0.0317	0.0395	0.0378	10.2	0.0001
16	DNMT3A	I705T	0.0663	0.0788	0.0753	0.0457	0.0582	0.0545	14.1	0.0000
18	TET2	W954X	-	-	0.0002	0.0023	0.0009	0.0012	10.1	0.0000
18	TET2	Y1337X	-	-	0.0000	0.0015	0.0013	0.0012	6.1	0.0000
18	BCORL1	R784X	-	-	0.0031	0.0186	0.0173	0.0171	14.2	0.0000
19	CBL	C381Y	0.0010	0.0014	0.0010	0.0011	0.0018	0.0013	17.2	0.0000
19	TP53	H168R	0.0007	0.0010	0.0008	0.0033	0.0032	0.0029	6.2	0.0000
20	KRAS	G12D	-	-	0.0001	0.0009	0.0014	0.0010	17.1	0.0000
20	TET2	A1381fs	0.0052	0.0032	0.0035	0.0400	0.0256	0.0357	13.1	0.0000

Table 4.7 Summary of variant allele fractions (VAF) detected by droplet digital PCR in sorted hematopoietic compartments.

PID	Gene	AA	COSMIC	Collection	Bulk	B	T	M
4	DNMT3A	R320X	133724	1	0.0068	0.0048	0.0005	0.0019
				2	0.0063	0.0063	0.0012	0.0039
5	DNMT3A	R635Q	1583088	1	0.0009	0.0006	0.0005	0.0005
				2	0.0020	0.0020	0.0015	0.0018
7	DNMT3A	P904L	87007	1	0.0020	0.0025	0.0001	0.0009
				2	0.0088	0.0135	0.0018	0.0037
8	TET2	G1288S	110780	1	0.0017	0.0000	0.0000	0.0006
				2	0.0016	0.0000	0.0000	0.0001
9	DNMT3A	S714C	87011	1	0.0282	0.0060	0.0018	0.0028
				2	0.0385	0.0066	0.0044	0.0058
11	DNMT3A	W860R	231568	1	0.0007	0.0000	0.0000	0.0001
				2	0.0022	0.0000	0.0000	0.0019
12	DNMT3A	G543A	256033	1	0.0007	0.0000	0.0043	0.0048
				2	0.0027	0.0015	0.0123	0.0121
13	DNMT3A	R882H	52944	1	0.0002	0.0000	0.0000	0.0001
				2	0.0020	0.0018	0.0000	0.0002
15	TP53	P278H	43755	1	0.0143	0.0234	0.0046	0.0035
				2	0.0378	0.0231	0.0358	0.0084
16	DNMT3A	I705T	1583102	1	0.0753	0.0760	0.0084	0.0350
				2	0.0545	0.0701	0.0218	0.0408
18	TET2	W954X	87110	1	0.0002	0.0000	0.0000	0.0000
				2	0.0012	0.0033	0.0004	0.0003
19	TP53	H168R	43545	1	0.0008	0.0007	0.0000	0.0000
				2	0.0029	0.0027	0.0006	0.0000
20	KRAS	G12D	521	1	0.0001	0.0000	0.0002	0.0007
				2	0.0010	0.0037	0.0001	0.0036

Chapter 5: Discussion

5.1 Change

Change is the only constant in genomics. Every day as our cells divide, differentiate and replenish our body, the underlying genomic blueprint of each cell changes. Every cell division propagates these new genetic and epigenetic modifications. Fortunately, the genome is large and most of these changes are inconsequential. Now the technology exists to deeply study these genomic changes within an organism. Change is also a central tenet of the sequencing technology that drives our growing understanding of genomics. During the Sanger-sequencing era, the cost of sequencing decreased yearly at a rate paralleling Moore's Law, culminating in the three billion dollar effort to sequencing the first human genome[22,185,186]. Since the development of NGS, sequencing costs have decreased even more rapidly and a human genome can now be sequenced for approximately one thousand dollars[185]. This democratization of sequencing has fostered a genomics renaissance illuminating the fundamentals of life and evolution. Ironically, this explosion of knowledge did not immediately provide cures for the panoply of human ailments. Instead, it exposed the vast complexity of human biology and malignancy. Regardless, the foundation of knowledge is growing and slowly being translated into targeted therapeutics, sensitive diagnostic and screening tools, and novel immunotherapies. It is exciting to imagine how future advancements will enhance this knowledge and improve treatment for patients battling cancer.

Here, we developed novel experimental and computational methods to improve rare variant detection with standard NGS technology. We characterized previously undetectable rare clonal populations in pre-chemotherapy blood and bone marrow samples from individuals who later developed t-AML/t-MDS and in the peripheral blood of healthy individuals. We were the first to describe the ubiquitous presence of rare hematopoietic clones harboring leukemia-associated mutations in healthy individuals. While these were interesting findings, it is exciting to look forward and wonder when this cutting edge technology will become obsolete. While the cost of sequencing has dramatically decreased over the past decade, the accuracy of sequencing has not significantly improved. We were able to circumvent the NGS error rate using our methods for ECS. However, future advancements in this field must improve sequencing accuracy, throughput and seamless integration with single-cell techniques.

5.2 Future Exploration in Clonal Hematopoiesis

Presented here is the first application of targeted-ECS to study clonal hematopoiesis in healthy middle-aged individuals. However, the causes and effects of clonal hematopoiesis are still not well understood. Given our current understanding of leukemogenesis and the age-dependent increase in risk of leukemia, we hypothesize that clonal expansion rarely occurs in younger individuals. To directly answer this question, we designed a study using targeted-ECS to identify rare hematopoietic clones in cord blood samples. The capture panel for this study was expanded to include genes recurrently mutated in pediatric AML. Future work could expand the study to include samples collected in adolescence and young adulthood. While difficult to obtain, longitudinal studies of healthy individuals are essential to understand the stability of these hematopoietic clones over time. Additionally, similar studies in individuals with germline mutations in cancer predisposition genes would help describe the transition from benign clonal hematopoiesis to fulminant leukemia[187]. These studies would be well suited to unveil the discrete steps that transform benign clonal expansion into a founding leukemic clone.

Another feasible study, given the current ECS technology, is the examination of rare clones in different hematopoietic compartments. Using flow cytometry, we were able to detect rare clonal mutations in both myeloid and lymphoid cells. Thus, these clonal mutations most likely arose in hematopoietic stem and progenitor cells (HSPCs). However, the functional effect of these leukemia-associated mutations in HSPCs is largely unknown. Previous work in mice demonstrated that loss-of-function mutations in *DNMT3A* resulted in increased hematopoietic stem cell (HSC) self-renewal, but not proliferation[181]. In a separate study, *TET2* loss-of-function mutations increased HSC self-renewal and proliferation[182]. Now the technology exists to examine the characteristics of clonal HSPC expansion in healthy adults who spontaneously acquire these mutations, but do not acquire leukemia. Future work could quantify compartment-specific expansion and skewing during differentiation. In this type of study, clonal mutations that block differentiation would appear as clonal expansion in a specific progenitor population.

So far, these proposed studies would examine the cell intrinsic characteristics of leukemia-associated mutations on clonal expansion. However, HSPCs do not exist in isolation from their environment. Instead there is a rich milieu of cytokines, circulating hematopoietic cells and the bone marrow niche that directly modulate the quiescence and activity of HSPCs. The interplay of these extrinsic factors and HSPCs harboring leukemia-associated mutations is totally unknown. One puzzling observation from our study of clonal hematopoiesis in healthy individuals was the spectrum of somatic mutations. Given the sensitivity of our technology and the estimated number of HSCs in humans, we expected to observe a random pattern of mutations mirroring the rate of somatic mutation and genetic drift in individual HSPCs. Instead, two-thirds of the observed mutations were in the epigenetic modifiers *DNMT3A* and *TET2*, suggesting that selection was already acting to enrich these clonal mutations that were in as few as 1:10,000 peripheral blood cells. Understanding the source of this selection could potentially explain why human HSCs increase in frequency, become more myeloid biased and decrease quiescence as a function of age[179]. It is possible that these two phenomenon are related. Perhaps a lifetime of infections, inflammatory processes and environmental exposures select for HSPCs that acquired these mutations in epigenetic modifiers. As a consequence, an oligoclonal marrow harboring leukemia-associated mutations may be necessary to maintain ostensibly normal blood production late in life. Using standard flow sorting techniques and ECS, it is currently possible to study the effect of inflammation and infection on clonal selection in the hematopoietic compartment.

5.3 Residual disease detection in AML with ECS

Current work in the lab will produce the first comparison of targeted-ECS to multiparameter flow cytometry (MPFC) for residual disease detection in AML. This collaboration with the Children's Oncology Group will enable us to test ECS-based residual disease detection in a large cohort of pediatric AML cases. There are several possible outcomes from this study. One outcome is that detected leukemic clones or subclones persist following treatment and predict a poor outcome. Alternatively, pre-leukemic clones harboring a subset of the leukemia-associated mutations may predominate post-induction, repopulate the marrow, and reconstitute an ostensibly normal hematopoietic compartment. Yet another possibility, therapy may select for a clonal population unrelated to the leukemia that expands post-induction. Regardless, this study will explore the characteristics of clonal expansion and mutation clearance post-induction that predict patient outcomes. Already, in adult AML the clearance of all leukemia-associated mutations to a detection limit of 0.025 variant allele fraction (VAF) was associated with better event-free survival and overall survival[111]. However, 70% of individuals who cleared all of their mutations at that limit of detection relapsed by 40 months. Using ECS, the limit of detection for clonal mutations is much lower. Likewise, this study will determine if mutation clearance at a lower limit of detection corresponds with longer event free survivals.

Complicating this analysis is the observation that non-leukemic hematopoietic clones expand following induction therapy in adult AML[112]. While a challenging task, future work will help discriminate between benign clonal expansion following treatment and malignant clonal expansion that precedes relapse.

5.4 Clonal evolution in solid organs

The hematopoietic compartment is the ideal system to study clonal evolution for a variety of reasons. Samples are routinely acquired from healthy individuals and often serially banked during the natural history of hematological disease. Conversely, it is difficult to study clonal evolution in solid organs. Focusing on malignancy, two pioneering studies described the exquisite geographical diversity of clonal mutations in renal cell carcinoma and pancreatic cancer[60,61]. However, a similar study of benign clonal expansion in a disease-free kidney or pancreas is not feasible. The only sequencing-based clonal evolution study of a solid organ characterized somatic mutations in eyelid epidermis, one of the only disease-free tissues routinely removed from healthy individuals[188]. Interestingly, they observed a high burden of somatic mutations, similar to many cases of skin cancer, that reflected the spectrum of mutations introduced by UV light. It would be fascinating to explore clonal evolution in other organ systems that do not experience the same level of DNA damage as sun-exposed skin. Studying somatic mutation acquisition and clonal evolution in a variety of organ systems would help clarify the effect of “bad luck”—spontaneous random mutations arising in disease-free stem cells—on the organ-specific risk of developing malignancy[189]. Regardless, the characterization of rare clonal expansion in non-hematopoietic organ systems is an important and necessary undertaking in order to understand the process of malignant transformation.

5.5 Predicting solid tumor development and relapse with circulating cell-free DNA

One challenge when treating solid malignancies is the accurate assessment of response to treatment and risk of relapse. While neoadjuvant chemotherapy and surgery can often effectively remove the primary tumor, distant metastases often escape therapy and may spawn recurrence. Currently, response to treatment is measured with the Response Evaluation Criteria in Solid Tumours (RECIST) criteria, which assesses disease burden using imaging modalities such as X-ray computed tomography (CT) and fluorodeoxyglucose-positron emission tomography (FDG-PET)[190,191]. Unfortunately, these methods are insensitive to detect occult lesions that frequently seed relapse. Interestingly, circulating tumor cells (CTCs) have been detected in the blood of individuals with metastatic breast cancer, prostate cancer, colorectal cancer, and a variety of other carcinomas[192–194]. Detecting as few as five CTCs per 7.5 mL vial of whole blood was associated with shorter progression free survival and lower overall survival. This method routinely detected residual malignant cells that were invisible using imaging modalities. However, by targeting epithelial cell surface markers, these methods miss many malignant cells such as tumor stem cells and CTCs that have undergone epithelial to mesenchymal transition[195]. These methods have improved over the years, but the primary challenge will always concern the identification of a sensitive and specific cell surface marker of disease.

More recent advancements circumvent this limitation by targeting the genomic DNA from malignant cells directly in the circulating cell-free fraction of the peripheral blood (cfDNA)[196]. In an early study of a few cases of metastatic breast cancer, tumor-specific mutations were detected in cfDNA and correlated with disease course[197]. More recent studies have utilized targeted sequencing of cancer-associated hot spots and whole genome sequencing to identify malignancy-associated mutations in cfDNA[156,198]. Interestingly, these methods could detect malignancy-associated mutations not observed in the primary tumor. While these methods were sensitive to detect malignancy-associated somatic mutations, the prognostic value of mutation burden in the cfDNA is still largely unknown. While the relative abundance of tumor-specific cfDNA may not directly correlate with disease burden or risk of disease progression, the absence of tumor-specific cell-free DNA may reflect disease clearance. Likewise, a binary readout of tumor DNA burden may be the most efficacious application of cfDNA sequencing. In that case, detecting rare tumor-specific mutations will be a critical component of disease assessment. Our ECS methods could improve the limit of detection for these rare malignancy-associated somatic mutations in cfDNA. While proposed, ECS has not yet been applied to rare tumor-associated mutations detection in cfDNA[196]. These proposed studies focus on cfDNA in peripheral blood; however, any medium is suitable for detecting the genomic signature of occult disease. One pioneering study retrospectively identified cases of ovarian and endometrial cancer by identifying tumor-specific mutations in Papanicolaou (Pap) smear samples using targeted-ECS[127]. In general, there is great potential for the application of ECS to quantify response to treatment for solid malignancies.

Another exciting application of cfDNA analysis is for the identification of malignancy before clinical presentation. Currently, only the sequencing company Illumina is pursuing this goal with the creation of a separate entity, Grail. This is a high risk venture for several reasons. First, the amount of sequencing required per person is nontrivial. Tumor-specific cfDNA will likely make up only a small portion of the total cfDNA and will only be detectable with ECS. Additionally, the target-space for capture must be large to recover a broad set of driver mutations across multiple cancer types. Second, the relationship between detecting malignancy-associated mutations in cfDNA and the development of cancer are totally unknown. While initially optimistic about this venture, our study of clonal hematopoiesis in healthy individuals demonstrated that virtually all healthy individuals have detectable leukemia-associated somatic mutations in their peripheral blood by late adulthood. The shear breadth of sequencing and low limit of detection will invariably lead to the detection of benign clonal expansions and rare mosaic populations in healthy individuals. The discrimination between these benign processes and actionable pre-clinical malignant disease will be a monumental challenge.

5.6 Leveraging improvements in sequencing technology to study clonal evolution

Now, we are on the cusp of several advancements in sequencing technology. It is exciting and humbling to think that our cutting edge ECS technology will likely be obsolete in only a few years. It will be exciting to how improving this fundamental tool will enable future studies of clonal evolution and the development of malignancy. Two emerging technologies are Single-Molecule Real-Time (SMRT) sequencing and nanopore sequencing. The two major limitations of NGS are that sequenced reads are short (300-600 bp long) and the error-rate is approximately 1%. While we can circumvent the error rate with ECS, the short reads are an intrinsic limitation of NGS. Conversely, SMRT sequencing can generate much longer reads (>1,000 bp long), but have a high error rate (up to 14%)[199]. This technology works by observing replication of a single DNA strand in a restricted space called a zero-mode waveguide. So far, this technology has been instrumental in genome assembly, which is nearly impossible in repetitive regions using short NGS-generated reads[200]. Likewise, these long reads have improved structural variant mapping in the human genome[201]. Nanopore sequencing generates long sequence reads by monitoring changes in ionic current as a DNA molecule passes through a membrane-bound nanopore[202].

While these long nanopore-generated reads also have a high intrinsic error rate, their combination with NGS-reads and statistical modelling have generated accurate genome assemblies[203,204]. More recent advancements allowed polarity reversal in individual pores to reject molecules that have already been sequenced[205]. Together these technologies are currently well suited for *de novo* genome assembly and structural variant identification.

While a fascinating group of technologies, currently, they are ill-suited for rare variant detection. If future versions of these technologies can achieve an error-rate similar ECS, then this tool will enable key experiments regarding clonal evolution. Specifically, this tool would enable the phasing of mutations occurring in the same gene and the reliable detection of rare clonal translocations and rearrangements. It is not farfetched to envision a future technology, which could phase mutations along an entire chromosome. However, those advancements may be a way off. Even if current nanopore technology could sequence an entire chromosome, it would take 41 days at 70bp/s to read chromosome 1, which is 249 million bp long[205]. Despite these hypothetical applications for long-read sequencing, their development alone will not likely dramatically improve our understanding of clonal evolution and malignant transformation. However, combining these technologies with future advancements in single-cell partitioning and sequencing will revolutionize our understanding of malignancy, clonal evolution and fundamental cellular biology.

5.7 Future applications of single-cell sequencing.

Single-cell sequencing technologies are rapidly improving. Already, single-cell RNA sequencing approaches can tag transcripts from individual cells and distinguish cell types based on commonly expressed transcripts[206]. Single-cell genomic DNA sequencing approaches enabled the identification of the clonal architecture within a few cases of pediatric ALL[79]. This approach queried individual loci, identified by bulk sequencing at diagnosis, to uncover the clonal architecture and infer the clonal evolution of these tumors. More recent advancements leverage microfluidic partitioning of genomic DNA to phase and haplotype inherited germline variants and cancer-associated somatic mutations[207]. Whereas previous studies targeted specific loci in the genome, this study surveyed the entire genome. Soon this technology will enable the partitioning of single-cell genomes into microfluidic droplets for sequencing. Small but necessary advancements in single-cell partitioning and accurate long-read sequencing would enable fascinating studies into the evolution of single cells within an organism.

Ideally, pairing these technologies would reveal the somatic mutation profile of each individual cell isolated from any biological sample (solid tumor, metastatic lesion, leukemia). With this information, it would be possible to reconstruct the exact phylogenetic tree in a single tumor and infer the step-by-step acquisition of mutations during the development of disease. This type of study will be possible with several key improvements to the current technology. First, the throughput of single-cell partitioning and genomic DNA isolation must be improved. While current reports have targeted a handful of loci in hundreds of cells, moving to whole genome sequencing for thousands of cells is a necessity. Second, allelic dropout during sample preparation is a key limitation of current single-cell sequencing platforms. Currently, this limitation is partially addressed by targeting many cells. Addressing these two limitation would permit several insightful studies into clonal evolution and the development of malignancy.

Already, we understand that AML is an oligoclonal disease with multiple subclones often present at diagnosis[62]. Currently, MPFC enables the detection of these leukemic cells based on leukemia-specific cell surface markers that are not present on healthy cells[92]. The prevailing notion is that the founding AML mutations drive the leukemia-associated immunophenotype

(LAIP) and subclonal mutations do not alter that phenotype. This theory may need revision based on a couple of observations. First, LAIP almost universally changes between diagnosis and relapse for AML, even though relapse often arises directly from the founding clone[64,102]. Additionally, single-cell sequencing studies of ALL demonstrated that clonal heterogeneity for VDJ recombination at the immunoglobulin heavy chain (IgH) locus was linked to specific subclonal somatic mutations[79]. In one case, a subclone-specific *EYA4* mutation allowed precursor B-cell populations to develop further resulting in VDJ recombination at the IgH locus, while other clones without the mutation were arrested in earlier stages of B-cell development. These B-cell maturation steps are tightly regulated and would result in subclone-specific immunophenotypes within the same tumor. The same process likely occurs in AML. Single-cell sequencing of AML samples could co-localize subclone-specific mutations and may predict immunophenotypic heterogeneity. A separate study could employ MPFC to partition various cellular compartments of an AML sample. Targeted ECS could then identify the subclonal mutations that govern each subclonal LAIP. These types of studies would further elucidate the clonal structure in AML.

Future advancements in single-cell and long-read sequencing could also dramatically improve residual disease detection in hematological malignancies. In the future, following treatment for AML, residual disease could be assessed with the following hypothetical protocol. The patient’s bone marrow is flow sorted to isolate individual HSCs and any persistent leukemic stem cells. Those sorted cells are partitioned into microfluidic chambers bound by a membrane studded with nanopores. Each cell is lysed and the entire genome is read, one chromosome at a time, through the nanopores. Bioinformatics analysis identifies the cell-specific somatic mutations and reconstructs the clonal architecture of the sample. The likelihood of relapse is assessed based on the specific somatic mutations present and their clonal co-localization. If additional treatment is warranted, the specific somatic mutations present would inform therapeutic selection. While purely hypothetical, the tools will soon exist to make this type of personalized care a reality.

The cytotoxic therapy given to eradicate a hematological malignancy wreaks havoc on the healthy hematological compartment. A robust single-cell sequencing approach would be able to examine the kinetics and genetics driving clonal expansion during recovery after chemotherapy exposure. Already, clonal expansion of nonleukemic clones has been observed following treatment for AML[112]. This phenomenon is probably closely related to the development of t-AML/t-MDS, which we already demonstrated can arise from pre-existing clones harboring *TP53* mutations[134]. A sensitive single-cell sequencing approach could identify these clones before chemotherapy exposure, track them longitudinally, and identify the co-operating mutations within a single cell that initiate leukemic transformation.

5.8 Normal hematopoietic stem cell biology

These techniques would also enhance our understanding of clonal hematopoiesis in healthy individuals. We have demonstrated that clonal expansion harboring leukemia-associated mutations are a common phenomenon in healthy elderly individuals. While we know these clones are pervasive, stable longitudinally and accumulate as a function of age, it is unknown what specific events are necessary to develop AML. While we identified many mutations per individual, we are unable to determine if they arise in the same HSPCs or occur in isolation. A whole genome single-cell sequencing approach could co-localize mutations. Expanding these techniques with longitudinal banking of pre-leukemic samples, would elucidate the stepwise process of mutation acquisition within single cells that drive leukemic transformation. This information will be essential for developing a sequencing-based screening tool for hematological malignancy. Thoroughly understanding this phenomenon in liquid tumors may someday permit screening for nascent solid malignancies in cfDNA.

One interesting observation made in our study of benign clonal hematopoiesis was that rare somatic mutations were not randomly distributed across the target space. We initially hypothesized that with a limit of detection of 0.0001 VAF and an estimated 10,000 HSCs in an adult, we would observe the private passenger mutations present in the specific HSCs contributing to hematopoiesis at the time. Instead we observed strong selection for mutations in the epigenetic modifiers *DNMT3A* and *TET2*. However, even in the first WGS study of a single case of AML, most of the somatic mutations identified were passenger mutations that reflected the life history of the initiating cell[43]. If single-cell sequencing can approach the accuracy of ECS, it would be possible to identify the progeny of each HSC based on their unique somatic mutation fingerprint. This would enable us to study the clonal dynamics of the hematopoietic compartment in healthy individuals. Currently, HSC dynamics are studied using viral barcoding, which perturbs the hematopoietic compartment and is confounded by the mutagenic effects of barcode insertion[208,209]. More recent approaches in mice have leveraged the Sleeping Beauty transposase to avoid transplantation to demonstrate that most blood production originates from long-lived progenitor cells, rather than HSCs[210]. Still, the mutagenic effect of random transposon insertion is unknown and none of these studies have examined native hematopoiesis in humans. All of these limitations could be addressed by a study using accurate single-cell whole-genome sequencing.

5.9 Targeting therapy in the AML genome

Currently, treatment for AML (except acute promyelocytic leukemia) consists primarily of chemotherapy that targets rapidly dividing cells. The primary drugs used for induction therapy are cytarabine (a nucleoside analog) and an anthracycline drug (an intercalating agent). Following induction, consolidation therapy with cycles of high dose cytarabine attempts to eradicate persistent disease. Many of these patients will receive a stem cell transplant to replace their hematopoietic compartment with healthy hematopoietic stem cells from a donor or themselves. Regardless, many patients go on to relapse and the five-year survival for AML is 26%[211]. Outcomes are particularly bleak in older individuals who cannot tolerate intensive chemotherapy or stem cell transplant, and most relapse within two years. The primary limitation of these treatments is that they do not target the leukemic cells directly, but rather all rapidly dividing cells.

In contrast, targeted therapeutics interact with specific molecular targets that directly disrupt tumor growth. The most famous example, imatinib mesylate, directly targets the BCR-ABL translocation, which is the initiating lesion in CML[212]. Unfortunately, there are currently few targeted therapeutics for treating AML, but many are currently in development[213]. Optimistically, in the future, AML may be treated as a chronic illness instead of a death sentence. At diagnosis, rapid, accurate sequencing of the tumor could inform selection of targeted therapeutics to eradicate the primary tumor and mitigate the clinical symptoms of disease. Following initial treatment, sensitive single-cell sequencing (described previously) could identify the persistent leukemic clones and their assortment of co-occurring somatic mutations. Here, multiple targeted therapeutics could target these specific residual clones. The process would be repeated months to years later to again selectively target and remove persistent disease.

This treatment approach has several key advantages over the current standard of care. First, therapy would not target healthy rapidly dividing cells, reducing the side effects of treatment. Second, targeted therapy would remove the severe selective pressure on the HSC compartment introduced by current chemotherapies that drive clonal expansion of pre-leukemic and non-leukemic clones[112,134]. Third, a lower side effect profile would enable treatment in older individuals that cannot tolerate current induction therapy. Advancements in AML residual disease detection, prognostication and therapy will be applicable to other types of malignancy. Together, these advancements will empower future clinician scientists to personalize effective care for these devastating malignancies.

5.10 Conclusion

Upon reflection, it is humbling to realize the small contribution that all of this work has made to our understanding of leukemia and normal hematopoiesis. We developed novel methods that enhanced current sequencing technology to characterize previously undetectable rare clonal mutations. However, the technology is always improving. Soon the protocols developed here will be hopelessly outdated. I welcome that day and optimistically look forward to the technological advancements that will embellish and challenge our conception of biology, and improve the health and survival of our species.

References

1. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2016. *CA. Cancer J. Clin.* **66,** 7–30 (2016).
2. Chin, L., Andersen, J. N. & Futreal, P. A. Cancer genomics: from discovery science to personalized medicine. *Nat. Med.* **17,** 297–303 (2011).
3. Nowell, P. The clonal evolution of tumor cell populations. *Science (80-.).* **194,** 23–28 (1976).
4. Rowley, J. D. Identificaton of a translocation with quinacrine fluorescence in a patient with acute leukemia. *Ann. génétique* **16,** 109–12 (1973).
5. Rowley, J. D. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature* **243,** 290–3 (1973).
6. Rowley, J. D. Nonrandom chromosomal abnormalities in hematologic disorders of man. *Proc. Natl. Acad. Sci. U. S. A.* **72,** 152–6 (1975).
7. Rowley, J., Golomb, H. & Dougherty, C. 15/17 translocation, a consistent chromosomal change in acute promyelocytic leukaemia. *Lancet* **309,** 549–550 (1977).
8. Fialkow, P. J. Use of genetic markers to study cellular origin and development of tumors in human females. *Adv. Cancer Res.* **15,** 191–226 (1972).
9. Fialkow, P. J., Gartler, S. M. & Yoshida, A. Clonal origin of chronic myelocytic leukemia in man. *Proc. Natl. Acad. Sci. U. S. A.* **58,** 1468–71 (1967).
10. Fialkow, P. J., Sagebiel, R. W., Gartler, S. M. & Rimoin, D. L. Multiple cell origin of hereditary neurofibromas. *N. Engl. J. Med.* **284,** 298–300 (1971).
11. Fialkow, P. J., Klein, G. & Clifford, P. Second malignant clone underlying a Burkitt-tumor exacerbation. *Lancet (London, England)* **2,** 629–31 (1972).
12. Linder, D. & Gartler, S. M. Glucose-6-Phosphate Dehydrogenase Mosaicism: Utilization as a Cell Marker in the Study of Leiomyomas. *Science (80-.).* **150,** 67–69 (1965).
13. Fialkow, P. J. Clonal Origin of Human Tumors. *Annu. Rev. Med.* **30,** 135–143 (1979).
14. Friedman, J. M., Fialkow, P. J., Greene, C. L. & Weinberg, M. N. Probable clonal origin of neurofibrosarcoma in a patient with hereditary neurofibromatosis. *J. Natl. Cancer Inst.* **69,** 1289–92 (1982).
15. Rowley, J. D. Chromosome abnormalities in cancer. *Cancer Genet. Cytogenet.* **2,** 175–198 (1980).
16. Testa, J. R., Mintz, U., Rowley, J. D., Vardiman, J. W. & Golomb, H. M. Evolution of karyotypes in acute nonlymphocytic leukemia. *Cancer Res.* **39,** 3619–27 (1979).
17. Shapiro, J. R., Yung, W. K. A. & Shapiro, W. R. Isolation, Karyotype, and Clonal Growth of Heterogeneous Sub-Populations of Human-Malignant Gliomas. *Cancer Res.* **41,** 2349–2359 (1981).
18. Reddy, E. P., Reynolds, R. K., Santos, E. & Barbacid, M. A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature* **300,** 149–152 (1982).
19. Tabin, C. J. *et al.* Mechanism of activation of a human oncogene. *Nature* **300,** 143–149 (1982).
20. Maxam, a M. & Gilbert, W. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 560–4 (1977).

21. Sanger, F., Nicklen, S. & Coulson, a R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5463–7 (1977).
22. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
23. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).
24. Sjöblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314,** 268–74 (2006).
25. Wood, L. D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* **318,** 1108–13 (2007).
26. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455,** 1061–1068 (2008).
27. Bardelli, A. *et al.* Mutational analysis of the tyrosine kinome in colorectal cancers. *Science* **300,** 949 (2003).
28. Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455,** 1069–1075 (2008).
29. Jones, S. *et al.* Core Signaling Pathways in Human Pancreatic Cancers Revealed by Global Genomic Analyses. *Science (80-.).* **321,** 1801–1806 (2008).
30. Davies, H. *et al.* Mutations of the BRAF gene in human cancer. *Nature* **417,** 949–954 (2002).
31. Dalgliesh, G. L. *et al.* Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. *Nature* **463,** 360–363 (2010).
32. Levine, R. L. *et al.* Activating mutation in the tyrosine kinase JAK2 in polycythemia vera, essential thrombocythemia, and myeloid metaplasia with myelofibrosis. *Cancer Cell* **7,** 387–397 (2005).
33. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446,** 153–8 (2007).
34. Pfeifer, G. P. & Besaratinia, A. Mutational spectra of human cancer. *Hum. Genet.* **125,** 493–506 (2009).
35. Futreal, P. A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4,** 177–183 (2004).
36. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458,** 719–724 (2009).
37. Forbes, S. a *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **44,** 1–7 (2014).
38. Forbes, S. a *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39,** D945-50 (2011).
39. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26,** 1135–1145 (2008).
40. Balasubramanian, S. Sequencing nucleic acids: from chemistry to medicine. *Chem. Commun.* **47,** 7281 (2011).
41. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).
42. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376–80 (2005).
43. Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456,** 66–72 (2008).

44. Mardis, E. R. *et al.* Recurring mutations found by sequencing an acute myeloid leukemia genome. *N. Engl. J. Med.* **361,** 1058–66 (2009).

45. Ley, T. J. *et al.* DNMT3A Mutations in Acute Myeloid Leukemia. *N. Engl. J. Med.* **363,** 2424–2433 (2010).

46. Esteller, M. Epigenetics in Cancer. *N. Engl. J. Med.* **358,** 1148–1159 (2008).

47. Paschka, P. *et al.* IDH1 and IDH2 Mutations Are Frequent Genetic Alterations in Acute Myeloid Leukemia and Confer Adverse Prognosis in Cytogenetically Normal Acute Myeloid Leukemia With NPM1 Mutation Without FLT3 Internal Tandem Duplication. *J. Clin. Oncol.* **28,** 3636–3643 (2010).

48. Shah, S. P. *et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461,** 809–13 (2009).

49. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463,** 191–196 (2010).

50. Pleasance, E. D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463,** 184–190 (2010).

51. Ellis, M. J. *et al.* Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* **486,** 353–60 (2012).

52. Garraway, L. a & Lander, E. S. Lessons from the cancer genome. *Cell* **153,** 17–37 (2013).

53. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339,** 1546–58 (2013).

54. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502,** 333–339 (2013).

55. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505,** 495–501 (2014).

56. Cancer Genome Research Atlas Network. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *N. Engl. J. Med.* **368,** 2059–2074 (2013).

57. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–8 (2013).

58. Yates, L. R. & Campbell, P. J. Evolution of the cancer genome. *Nat. Rev. Genet.* **13,** 795–806 (2012).

59. Höglund, M., Gisselsson, D., Säll, T. & Mitelman, F. Coping with complexity. *Cancer Genet. Cytogenet.* **135,** 103–109 (2002).

60. Yachida, S. *et al.* Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467,** 1114–1117 (2010).

61. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366,** 883–92 (2012).

62. Welch, J. S. *et al.* The Origin and Evolution of Mutations in Acute Myeloid Leukemia. *Cell* **150,** 264–278 (2012).

63. Walter, M. J. *et al.* Clonal architecture of secondary acute myeloid leukemia. *N. Engl. J. Med.* **366,** 1090–8 (2012).

64. Ding, L. *et al.* Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481,** 506–10 (2012).

65. Kronke, J. *et al.* Clonal evolution in relapsed NPM1-mutated acute myeloid leukemia. *Blood* **122,** 100–108 (2013).

66. Takahashi, K. *et al.* Clonal evolution of acute myeloid leukemia relapsed after 19 years of remission. *Am. J. Hematol.* **90,** E134–E135 (2015).

67. Landau, D. A. *et al.* Evolution and Impact of Subclonal Mutations in Chronic

Lymphocytic Leukemia. *Cell* **152,** 714–726 (2013).

68. Keats, J. J. *et al.* Clonal competition with alternating dominance in multiple myeloma. *Blood* **120,** 1067–1076 (2012).
69. Mullighan, C. G. *et al.* Genomic Analysis of the Clonal Origins of Relapsed Acute Lymphoblastic Leukemia. *Science (80-.).* **322,** 1377–1380 (2008).
70. Schuh, A. *et al.* Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood* **120,** 4191–4196 (2012).
71. Roche-Lestienne, C. Several types of mutations of the Abl gene can be found in chronic myeloid leukemia patients resistant to STI571, and they can pre-exist to the onset of treatment. *Blood* **100,** 1014–1018 (2002).
72. Shah, N. P. *et al.* Multiple BCR-ABL kinase domain mutations confer polyclonal resistance to the tyrosine kinase inhibitor imatinib (STI571) in chronic phase and blast crisis chronic myeloid leukemia. *Cancer Cell* **2,** 117–125 (2002).
73. Oshima, K. *et al.* Mutational landscape, clonal evolution patterns, and role of RAS mutations in relapsed acute lymphoblastic leukemia. *Proc. Natl. Acad. Sci. U. S. A.* (2016). doi:10.1073/pnas.1608420113
74. Ma, X. *et al.* Rise and fall of subclones from diagnosis to relapse in pediatric B-acute lymphoblastic leukaemia. *Nat. Commun.* **6,** 6604 (2015).
75. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472,** 90–4 (2011).
76. Xu, X. *et al.* Single-Cell Exome Sequencing Reveals Single-Nucleotide Mutation Characteristics of a Kidney Tumor. *Cell* **148,** 886–895 (2012).
77. Hou, Y. *et al.* Single-Cell Exome Sequencing and Monoclonal Evolution of a JAK2-Negative Myeloproliferative Neoplasm. *Cell* **148,** 873–885 (2012).
78. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* **17,** 175–188 (2016).
79. Gawad, C., Koh, W. & Quake, S. R. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. *Proc. Natl. Acad. Sci.* **111,** 201420822 (2014).
80. Hughes, A. E. O. *et al.* Clonal Architecture of Secondary Acute Myeloid Leukemia Defined by Single-Cell Sequencing. *PLoS Genet.* **10,** e1004462 (2014).
81. Klco, J. M. *et al.* Functional heterogeneity of genetically defined subclones in acute myeloid leukemia. *Cancer Cell* **25,** 379–92 (2014).
82. Griffith, M. *et al.* Optimizing Cancer Genome Sequencing and Analysis. *Cell Syst.* **1,** 210–223 (2015).
83. Miyamoto, T., Weissman, I. L. & Akashi, K. AML1/ETO-expressing nonleukemic stem cells in acute myelogenous leukemia with 8;21 chromosomal translocation. *Proc. Natl. Acad. Sci.* **97,** 7521–7526 (2000).
84. Hong, D. *et al.* Initiating and Cancer-Propagating Cells in TEL-AML1-Associated Childhood Leukemia. *Science (80-.).* **319,** 336–339 (2008).
85. Jan, M. *et al.* Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia. *Sci. Transl. Med.* **4,** 149ra118 (2012).
86. Corces-Zimmerman, M. R., Hong, W.-J., Weissman, I. L., Medeiros, B. C. & Majeti, R. Preleukemic mutations in human acute myeloid leukemia affect epigenetic regulators and persist in remission. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 2548–53 (2014).

87. Shlush, L. I. *et al.* Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature* **506,** 328–333 (2014).

88. Pløen, G. G. *et al.* Persistence of DNMT3A mutations at long-term remission in adult patients with AML. *Br. J. Haematol.* **167,** 478–486 (2014).

89. Wiseman, D. H. *et al.* Frequent reconstitution of IDH2R140Q mutant clonal multilineage hematopoiesis following chemotherapy for acute myeloid leukemia. *Leukemia* **30,** 1946–1950 (2016).

90. Wood, B. L. Principles of minimal residual disease detection for hematopoietic neoplasms by flow cytometry. *Cytom. Part B Clin. Cytom.* **90,** 47–53 (2016).

91. Orfao, A., Ortuño, F., de Santiago, M., Lopez, A. & San Miguel, J. Immunophenotyping of acute leukemias and myelodysplastic syndromes. *Cytom. Part A* **58A,** 62–71 (2004).

92. Loken, M. R. *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* **120,** 1581–1588 (2012).

93. Rubnitz, J. E. *et al.* Minimal residual disease-directed therapy for childhood acute myeloid leukaemia: results of the AML02 multicentre trial. *Lancet Oncol.* **11,** 543–552 (2010).

94. Terwijn, M. *et al.* High Prognostic Impact of Flow Cytometric Minimal Residual Disease Detection in Acute Myeloid Leukemia: Data From the HOVON/SAKK AML 42A Study. *J. Clin. Oncol.* **31,** 3889–3897 (2013).

95. Freeman, S. D. *et al.* Prognostic Relevance of Treatment Response Measured by Flow Cytometric Residual Disease Detection in Older Patients With Acute Myeloid Leukemia. *J. Clin. Oncol.* **31,** 4123–4131 (2013).

96. Al-Mawali, A., Gillis, D. & Lewis, I. The use of receiver operating characteristic analysis for detection of minimal residual disease using five-color multiparameter flow cytometry in acute myeloid leukemia identifies patients with high risk of relapse. *Cytom. Part B Clin. Cytom.* **76B,** 91–101 (2009).

97. Buccisano, F. *et al.* Prognostic and therapeutic implications of minimal residual disease detection in acute myeloid leukemia. *Blood* **119,** 332–41 (2012).

98. Grimwade, D. & Freeman, S. D. Defining minimal residual disease in acute myeloid leukemia: which platforms are ready for 'prime time'? *Blood* **124,** 3345–3355 (2014).

99. Paietta, E. Minimal residual disease in acute myeloid leukemia: coming of age. *Hematology Am. Soc. Hematol. Educ. Program* **2012,** 35–42 (2012).

100. Kern, W., Bacher, U., Haferlach, C., Schnittger, S. & Haferlach, T. The role of multiparameter flow cytometry for disease monitoring in AML. *Best Pract. Res. Clin. Haematol.* **23,** 379–390 (2010).

101. Duncavage, E. J. & Tandon, B. The utility of next-generation sequencing in diagnosis and monitoring of acute myeloid leukemia and myelodysplastic syndromes. *Int. J. Lab. Hematol.* **37,** 115–121 (2015).

102. Baer, M. R. *et al.* High frequency of immunophenotype changes in acute myeloid leukemia at relapse: implications for residual disease detection (Cancer and Leukemia Group B Study 8361). *Blood* **97,** 3574–80 (2001).

103. Zeijlemaker, W., Gratama, J. W. & Schuurhuis, G. J. Tumor heterogeneity makes AML a 'moving target' for detection of residual disease. *Cytom. Part B Clin. Cytom.* **86,** 3–14 (2014).

104. Grupp, S. a. *et al.* Chimeric Antigen Receptor–Modified T Cells for Acute Lymphoid Leukemia. *N. Engl. J. Med.* **368,** 1509–1518 (2013).

105. Spencer, D. H. *et al.* Detection of FLT3 Internal Tandem Duplication in Targeted, Short-Read-Length, Next-Generation Sequencing Data. *J. Mol. Diagnostics* **15,** 81–93 (2013).
106. Wu, D. *et al.* High-throughput sequencing detects minimal residual disease in acute T lymphoblastic leukemia. *Sci. Transl. Med.* **4,** 134ra63 (2012).
107. Wu, D. *et al.* Detection of Minimal Residual Disease in B Lymphoblastic Leukemia by High-Throughput Sequencing of IGH. *Clin. Cancer Res.* **20,** 4540–8 (2014).
108. Salipante, S. J., Fromm, J. R., Shendure, J., Wood, B. L. & Wu, D. Detection of minimal residual disease in NPM1-mutated acute myeloid leukemia by next-generation sequencing. *Mod. Pathol.* 1–9 (2014). doi:10.1038/modpathol.2014.57
109. Kohlmann, A. *et al.* Monitoring of residual disease by next-generation deep-sequencing of RUNX1 mutations can identify acute myeloid leukemia patients with resistant disease. *Leukemia* **28,** 129–137 (2014).
110. Thol, F. *et al.* Next-generation sequencing for minimal residual disease monitoring in acute myeloid leukemia patients with FLT3-ITD or NPM1 mutations. *Genes, Chromosom. Cancer* **51,** 689–695 (2012).
111. Klco, J. M. *et al.* Association Between Mutation Clearance After Induction Therapy and Outcomes in Acute Myeloid Leukemia. *JAMA* **314,** 811 (2015).
112. Wong, T. N. *et al.* Rapid expansion of preexisting nonleukemic hematopoietic clones frequently follows induction therapy for de novo AML. *Blood* **127,** 893–897 (2016).
113. Spencer, D. H. *et al.* Performance of Common Analysis Methods for Detecting Low-Frequency Single Nucleotide Variants in Targeted Next-Generation Sequence Data. *J. Mol. Diagnostics* **16,** 75–88 (2014).
114. Perea, G. *et al.* Prognostic value of minimal residual disease (MRD) in acute myeloid leukemia (AML) with favorable cytogenetics [t(8;21) and inv(16)]. *Leukemia* **20,** 87–94 (2006).
115. Leung, W. *et al.* Detectable minimal residual disease before hematopoietic cell transplantation is prognostic but does not preclude cure for children with very-high-risk leukemia. *Blood* **120,** 468–72 (2012).
116. Hourigan, C. S. & Karp, J. E. Minimal residual disease in acute myeloid leukaemia. *Nat. Rev. Clin. Oncol.* **10,** 460–71 (2013).
117. Schmitt, M. W. *et al.* Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl. Acad. Sci.* **109,** 14508–14513 (2012).
118. Schmitt, M. W. *et al.* Sequencing small genomic targets with high efficiency and extreme accuracy. *Nat. Methods* **12,** 423–426 (2015).
119. Kennedy, S. R. *et al.* Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat Protoc* **9,** 2586–2606 (2014).
120. Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K. W. & Vogelstein, B. Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl. Acad. Sci.* **108,** 9530–9535 (2011).
121. Kinde, I., Papadopoulos, N., Kinzler, K. W. & Vogelstein, B. FAST-SeqS: A Simple and Efficient Method for the Detection of Aneuploidy by Massively Parallel Sequencing. *PLoS One* **7,** e41162 (2012).
122. Lou, D. I. *et al.* High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proc. Natl. Acad. Sci.* **110,** 19872–19877 (2013).
123. Kirsch, S. & Klein, C. a. Sequence error storms and the landscape of mutations in cancer. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 14289–14290 (2012).

124. Jabara, C. B., Jones, C. D., Roach, J., Anderson, J. A. & Swanstrom, R. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proc. Natl. Acad. Sci.* **108,** 20166–20171 (2011).
125. Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9,** 72–74 (2011).
126. Keys, J. R. *et al.* Primer ID Informs Next-Generation Sequencing Platforms and Reveals Preexisting Drug Resistance Mutations in the HIV-1 Reverse Transcriptase Coding Domain. *AIDS Res. Hum. Retroviruses* **31,** 658–668 (2015).
127. Kinde, I. *et al.* Evaluation of DNA from the Papanicolaou test to detect ovarian and endometrial cancers. *Sci. Transl. Med.* **5,** 167ra4 (2013).
128. Kennedy, S. R., Salk, J. J., Schmitt, M. W. & Loeb, L. A. Ultra-Sensitive Sequencing Reveals an Age-Related Increase in Somatic Mitochondrial Mutations That Are Inconsistent with Oxidative Damage. *PLoS Genet.* **9,** e1003794 (2013).
129. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11,** 163–166 (2013).
130. Hiatt, J. B., Pritchard, C. C., Salipante, S. J., O'Roak, B. J. & Shendure, J. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res.* **23,** 843–854 (2013).
131. Kumar, A. *et al.* Deep sequencing of multiple regions of glial tumors reveals spatial heterogeneity for mutations in clinically relevant genes. *Genome Biol.* **15,** 530 (2014).
132. Godley, L. a & Larson, R. a. Therapy-related myeloid leukemia. *Semin. Oncol.* **35,** 418–29 (2008).
133. Larson, R. a. Etiology and Management of Therapy-Related Myeloid Leukemia. *Hematology* **2007,** 453–459 (2007).
134. Wong, T. N. *et al.* Role of TP53 mutations in the origin and evolution of therapy-related acute myeloid leukaemia. *Nature* **518,** 552–555 (2014).
135. Young, A. L. *et al.* Quantifying ultra-rare pre-leukemic clones via targeted error-corrected sequencing. *Leukemia* **29,** 1608–1611 (2015).
136. Busque, L. *et al.* Nonrandom X-inactivation patterns in normal females: lyonization ratios vary with age. *Blood* **88,** 59–65 (1996).
137. Busque, L. & Gilliland, D. G. X-inactivation analysis in the 1990s: promise and potential problems. *Leukemia* **12,** 128–135 (1998).
138. Gale, R. E., Fielding, A. K., Harrison, C. N. & Linch, D. C. Acquired skewing of X-chromosome inactivation patterns in myeloid cells of the elderly suggests stochastic clonal loss with age. *Br. J. Haematol.* **98,** 512–519 (1997).
139. Busque, L. *et al.* Recurrent somatic TET2 mutations in normal elderly individuals with clonal hematopoiesis. *Nat. Genet.* **44,** 1179–1181 (2012).
140. Laurie, C. C. *et al.* Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nat. Genet.* **44,** 642–650 (2012).
141. Jacobs, K. B. *et al.* Detectable clonal mosaicism and its relationship to aging and cancer. *Nat. Genet.* **44,** (2012).
142. Kyle, R. A. *et al.* A Long-Term Study of Prognosis in Monoclonal Gammopathy of Undetermined Significance. *N. Engl. J. Med.* **346,** 564–569 (2002).
143. Rawstron, A. C. *et al.* Monoclonal B-Cell Lymphocytosis and Chronic Lymphocytic Leukemia. *N. Engl. J. Med.* **359,** 575–583 (2008).
144. Xie, M. *et al.* Age-related mutations associated with clonal hematopoietic expansion and

malignancies. *Nat. Med.* **20,** 1472–1478 (2014).

145. Jaiswal, S. *et al.* Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371,** 2488–98 (2014).

146. Genovese, G. *et al.* Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. *N. Engl. J. Med.* **371,** 2477–2487 (2014).

147. McKerrell, T. *et al.* Leukemia-Associated Somatic Mutations Drive Distinct Patterns of Age-Related Clonal Hemopoiesis. *Cell Rep.* **10,** 1239–1245 (2015).

148. Holstege, H. *et al.* Somatic mutations found in the healthy blood compartment of a 115-yr-old woman demonstrate oligoclonal hematopoiesis. *Genome Res.* **24,** 733–42 (2014).

149. Steensma, D. P. *et al.* Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. *Blood* **126,** 9–16 (2015).

150. Young, A. L., Challen, G. A., Birmann, B. M. & Druley, T. E. Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. *Nat. Commun.* **7,** 12484 (2016).

151. Shibutani, S., Takeshita, M. & Grollman, A. P. Insertion of specific bases during DNA synthesis past the oxidation-damaged base 8-oxodG. *Nature* **349,** 431–434 (1991).

152. Stiller, M. *et al.* Patterns of nucleotide misincorporations during enzymatic amplification and direct large-scale sequencing of ancient DNA. *Proc. Natl. Acad. Sci.* **103,** 13578–13584 (2006).

153. Teer, J. K. & Mullikin, J. C. Exome sequencing: the sweet spot before whole genomes. *Hum. Mol. Genet.* **19,** R145–R151 (2010).

154. Hindson, B. J. *et al.* High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. *Anal. Chem.* **83,** 8604–8610 (2011).

155. Cha, R. S. & Thilly, W. G. Specificity, efficiency, and fidelity of PCR. *PCR Methods Appl.* **3,** S18-29 (1993).

156. Forshew, T. *et al.* Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci. Transl. Med.* **4,** 136ra68 (2012).

157. Hoang, M. L. *et al.* Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. *Proc. Natl. Acad. Sci.* 201607794 (2016). doi:10.1073/pnas.1607794113

158. Untergasser, A. *et al.* Primer3--new capabilities and interfaces. *Nucleic Acids Res.* **40,** e115 (2012).

159. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).

160. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–9 (2009).

161. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14,** 178–92 (2013).

162. Wickham, H. *ggplot2*. (Springer New York, 2009). doi:10.1007/978-0-387-98141-3

163. Fu, G. K. *et al.* Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparations. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 1891–6 (2014).

164. Catlin, S. N., Busque, L., Gale, R. E., Guttorp, P. & Abkowitz, J. L. The replication rate of human hematopoietic stem cells in vivo. *Blood* **117,** 4460–4466 (2011).

165. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous

variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4,** 1073–81 (2009).

166. Loman, N. J. *et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* **30,** 434–9 (2012).

167. Landgren, O. *et al.* Monoclonal gammopathy of undetermined significance (MGUS) consistently precedes multiple myeloma: a prospective study. *Blood* **113,** 5412–5417 (2009).

168. Colditz, G. a & Hankinson, S. E. The Nurses' Health Study: lifestyle and health among women. *Nat. Rev. Cancer* **5,** 388–396 (2005).

169. Hankinson, S. E. *et al.* Alcohol, Height, and Adiposity in Relation to Estrogen and Prolactin Levels in Postmenopausal Women. *JNCI J. Natl. Cancer Inst.* **87,** 1297–1302 (1995).

170. Zhang, X., Tworoger, S. S., Eliassen, A. H. & Hankinson, S. E. Postmenopausal plasma sex hormone levels and breast cancer risk over 20 years of follow-up. *Breast Cancer Res. Treat.* **137,** 883–892 (2013).

171. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).

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

173. McVean, G. A. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).

174. Harrow, J. *et al.* GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* **22,** 1760–1774 (2012).

175. Karolchik, D. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32,** 493D–496 (2004).

176. Koboldt, D. C. *et al.* VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22,** 568–576 (2012).

177. Huggett, J. F. *et al.* The digital MIQE guidelines: Minimum information for publication of quantitative digital PCR experiments. *Clin. Chem.* **59,** 892–902 (2013).

178. Weissmann, S. *et al.* Landscape of TET2 mutations in acute myeloid leukemia. *Leukemia* **26,** 934–942 (2012).

179. Pang, W. W. *et al.* Human bone marrow hematopoietic stem cells are increased in frequency and myeloid-biased with age. *Proc. Natl. Acad. Sci.* **108,** 20012–20017 (2011).

180. Kuranda, K. *et al.* Age-related changes in human hematopoietic stem/progenitor cells. *Aging Cell* **10,** 542–546 (2011).

181. Challen, G. a *et al.* Dnmt3a is essential for hematopoietic stem cell differentiation. *Nat. Genet.* **44,** 23–31 (2011).

182. Moran-Crusio, K. *et al.* Tet2 Loss Leads to Increased Hematopoietic Stem Cell Self-Renewal and Myeloid Transformation. *Cancer Cell* **20,** 11–24 (2011).

183. Collins, F. S. & Varmus, H. A New Initiative on Precision Medicine. *N. Engl. J. Med.* **372,** 793–795 (2015).

184. Diaz, L. A. & Bardelli, A. Liquid Biopsies: Genotyping Circulating Tumor DNA. *J. Clin. Oncol.* **32,** 579–586 (2014).

185. Check Hayden, E. Technology: The $1,000 genome. *Nature* **507,** 294–295 (2014).

186. Moore, G. E. Cramming More Components Onto Integrated Circuits. *Electronics* **38,** (1965).

187. Rahman, N. Realizing the promise of cancer predisposition genes. *Nature* **505,** 302–308

(2014).
188. Martincorena, I. *et al.* High burden and pervasive positive selection of somatic mutations in normal human skin. *Science (80-.).* **348,** 880–886 (2015).
189. Tomasetti, C. & Vogelstein, B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. (2014).
190. Eisenhauer, E. A. *et al.* New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *Eur. J. Cancer* **45,** 228–247 (2009).
191. Therasse, P. *et al.* New Guidelines to Evaluate the Response to Treatment in Solid Tumors. *JNCI J. Natl. Cancer Inst.* **92,** 205–216 (2000).
192. Cristofanilli, M. *et al.* Circulating Tumor Cells, Disease Progression, and Survival in Metastatic Breast Cancer. *N. Engl. J. Med.* **351,** 781–791 (2004).
193. Cohen, S. J. *et al.* Relationship of circulating tumor cells to tumor response, progression-free survival, and overall survival in patients with metastatic colorectal cancer. *J. Clin. Oncol.* **26,** 3213–21 (2008).
194. de Bono, J. S. *et al.* Circulating tumor cells predict survival benefit from treatment in metastatic castration-resistant prostate cancer. *Clin. Cancer Res.* **14,** 6302–9 (2008).
195. Friedlander, T. W. & Fong, L. The End of the Beginning: Circulating Tumor Cells As a Biomarker in Castration-Resistant Prostate Cancer. *J. Clin. Oncol.* **32,** 1104–1106 (2014).
196. Volik, S., Alcaide, M., Morin, R. D. & Collins, C. C. Cell-free DNA (cfDNA): clinical significance and utility in cancer shaped by emerging technologies. *Mol. Cancer Res.* molcanres.0044.2016 (2016). doi:10.1158/1541-7786.MCR-16-0044
197. Dawson, S.-J., Rosenfeld, N. & Caldas, C. Circulating tumor DNA to monitor metastatic breast cancer. *N. Engl. J. Med.* **369,** 93–4 (2013).
198. Murtaza, M. *et al.* Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature* **497,** 108–12 (2013).
199. Eid, J. *et al.* Real-Time DNA Sequencing from Single Polymerase Molecules. *Science (80-.).* **323,** 133–138 (2009).
200. Chin, C.-S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10,** 563–569 (2013).
201. Chaisson, M. J. P. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517,** 608–611 (2014).
202. Branton, D. *et al.* The potential and challenges of nanopore sequencing. *Nat. Biotechnol.* **26,** 1146–1153 (2008).
203. Goodwin, S. *et al.* Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Res.* **25,** 1750–1756 (2015).
204. Szalay, T. & Golovchenko, J. A. De novo sequencing and variant calling with nanopores using PoreSeq. *Nat. Biotechnol.* **33,** 1087–1091 (2015).
205. Loose, M., Malla, S. & Stout, M. Real-time selective sequencing using nanopore technology. *Nat. Methods* **13,** 751–754 (2016).
206. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161,** 1202–1214 (2015).
207. Zheng, G. X. Y. *et al.* Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* **34,** 303–311 (2016).
208. Braun, C. J. *et al.* Gene Therapy for Wiskott-Aldrich Syndrome--Long-Term Efficacy and Genotoxicity. *Sci. Transl. Med.* **6,** 227ra33-227ra33 (2014).
209. Lu, R., Neff, N. F., Quake, S. R. & Weissman, I. L. Tracking single hematopoietic stem

cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. *Nat. Biotechnol.* **29,** 928–933 (2011).

210. Sun, J. *et al.* Clonal dynamics of native haematopoiesis. *Nature* **514,** 322–327 (2014).

211. Howlader, N. *et al.* SEER Cancer Statistics Review, 1975-2013, National Cancer Institute. Bethesda, MD, http://seer.cancer.gov/csr/1975_2013/, based on November 2015 SEER data submission, posted to the SEER web site, April 2016.

212. Druker, B. J. *et al.* Efficacy and Safety of a Specific Inhibitor of the BCR-ABL Tyrosine Kinase in Chronic Myeloid Leukemia. *N. Engl. J. Med.* **344,** 1031–1037 (2001).

213. DeAngelo, D. J., Stein, E. M. & Ravandi, F. Evolving Therapies in Acute Myeloid Leukemia: Progress at Last? *Am. Soc. Clin. Oncol. Educ. Book* **35,** e302-12 (2016).