Sensitive quantification of clonal evolution in Acute Myeloid Leukemia

Daniel Richter

Dissertation der Fakultät für Biologie der

Ludwig-Maximilians-Universität München

München

2022

1. Gutachter: Prof. Dr. Wolfgang Enard

2. Gutachter: Prof. Dr. Heinrich Leonhardt

Tag der Abgabe: 30.05.2022

Tag der mündlichen Prüfung: 01.12.2022

Eidesstattliche Versicherung und Erklärung

Eidesstattliche Versicherung

Ich versichere hiermit an Eides statt, dass die vorgelegte Dissertation mit dem Titel

„*Sensitive quantification of clonal evolution in Acute Myeloid Leukemia*“

von mir selbständig und ohne unerlaubte Hilfe angefertigt ist.

München, den 27.05.2022

Daniel Richter

Erklärung

Hiermit erkläre ich, dass die Dissertation nicht ganz oder in wesentlichen Teilen einer anderen Prüfungskommission vorgelegt worden ist.

Ich erkläre weiter, dass ich mich anderweitig einer Doktorprüfung ohne Erfolg nicht unterzogen habe.

München, den 27.05.2022

Daniel Richter

Table of contents

1 Abstract

Cancer cells arise by acquiring genetic and epigenetic alterations. The transformation of healthy cells towards a cancerous state is a multi-step process in which several driver mutations that increase cellular fitness are selected for over time. Due to this evolutionary process cancer cells are genomically and phenotypically heterogeneous. This clonal evolution and the resulting intratumoral heterogeneity (ITH) have been recognized as major contributors to the stagnating success in the "war" against cancer. In order to better explore clonal evolution and ITH, two complementary approaches were established that allow to track and quantify the subclonal composition within hematopoietic neoplasms.

Establishment of an optimized, highly sensitive and cost-efficient single-molecule Molecular Inversion Probes (smMIPs) panel allowed targeted sequencing of hundreds of human gDNA samples within a first cohort study. This panel allows to detect residual leukemic cells as well as expanded hematopoietic clones. The latter are indicative of clonal hematopoiesis which represents the onset of clonal evolution and often precedes the acquisition of hematopoietic neoplasms such as Acute Myeloid Leukemia (AML). The established assay showed robust performance and high concordance in comparison to a commercial targeted sequencing panel, while drastically reducing the costs for library preparations and represents a useful tool for future studies of clonal evolution in the context of clonal hematopoiesis and AML.

As a second approach, cellular barcoding assays were established to enable direct investigation of ITH within a mouse model for AML patient-derived xenografts (Pdx). Advanced cloning strategies enabled the construction of millions of different barcodes that are used to uniquely tag leukemic cells by lentiviral transduction, allowing to directly identify changes in cell compositions of leukemic populations. Within the first experimental setups the established barcoding constructs allowed to quantify the bottleneck upon engraftment of cells within the mouse model, to determine the frequency of Leukemia Initiating Cells (LICs) with fewer mice than traditional limiting-dilution transplantation assays (LDTAs) and provided direct prove for the presence of subclones with increased resistance towards *in vivo* chemotherapeutic treatment. Additionally, cellular barcodes enabled identification of isolates derived from single-cells, thereby allowing to isolate and further characterize genetically and phenotypically distinct subclones from Pdx samples of a single AML patient.

Both established assays demonstrated their usefulness within first pilot experiments and represent powerful tools to further study clonal evolution and heterogeneity associated with AML.

2 Introduction

2.1 Cancer – An evolutionary disease of the genome

Cancer is one of the major causes for premature death with about 19 million new cases and almost 10 million caused deaths in 2020 [*1*]. Although more than 270 types of cancer can be distinguished based on the affected tissues and other characteristics [*2*], its general emergence can be simplified to acquisition of genomic mutations, which ultimately result in cells with limitless potential to replicate. Additionally, among other general features, cancers are able to evade cell death and responses of the immune system and to invade tissues leading to impairment of normal tissue functions [*3, 4*]. Somatic driver mutations that confer these phenotypes are randomly gained throughout cell divisions in parallel to neutral passenger mutations that do not cause any advantage for the cell [*5*].

The development of cancer can therefore be seen as an evolutionary process in which those randomly acquired somatic mutations which increase the cellular fitness relative to competing cells are selected for [*6, 7*]. Upon gain of a first driver mutation, different fractions of the transformed tumor population may acquire additional functionally advantageous mutations, leading to the presence of subclones which may differ genotypically as well as phenotypically among each other (Figure 1). Therefore acquired driver mutations not only differ between patients, but often also among different subclones within a single cancer, which is known as Intratumoral Heterogeneity (ITH) [*8-11*].

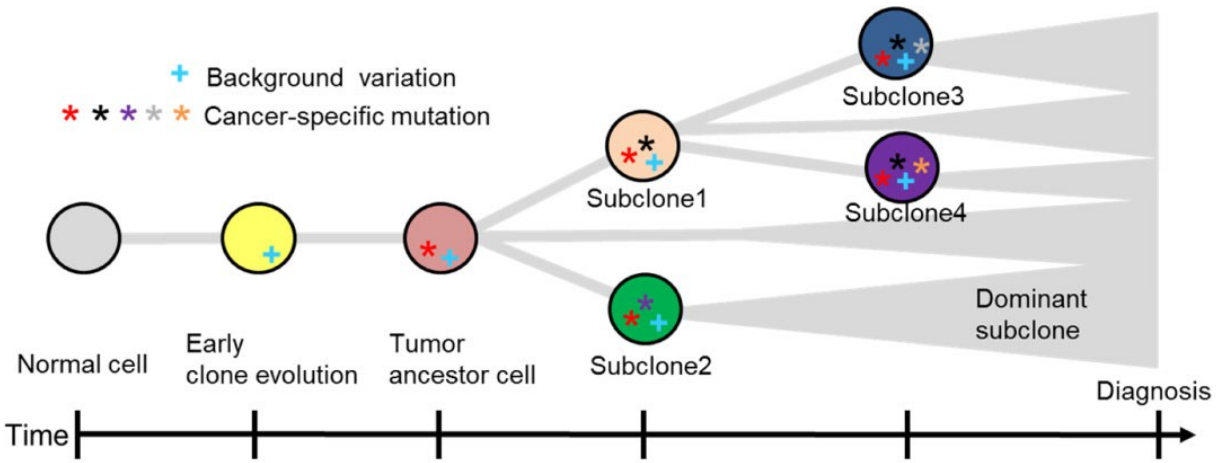


Figure 1: Concept of cancer evolution. Random mutations are acquired over time. Occurrence of a first driver mutation transforms the healthy clonal lineage into a tumor ancestor cell. More driver mutations may be acquired over time in fractions of the cancerous cells, leading to emergence of different subclonal lineages with different genotypic and phenotypic properties. Modified from Lin et al.[12] (CC BY 4.0)

ITH represents a major issue for the successful treatment of cancer as previously established subclonal lineages resistant to chemotherapy may be selected for by standard therapies, in turn leading to a relapse consisting of therapy resistant cells after seemingly successful initial treatment [*13-15*].

Therefore, investigating the emergence and characterization of adverse subclones within cancerous populations has become an important part in order to elucidate more details about cancer evolution and to further improve existing therapeutic approaches as well as to develop new ones [*15*].

2.2 Clonal hematopoiesis – a "not-yet-disease" state preceding AML

The progression from a normal, healthy cell towards a cancerous state usually happens gradually by acquisition of multiple mutations that progressively disequilibrate the normal cellular state. In the case of blood cancers hematopoietic stem or progenitor cells may first acquire somatic mutations that only confer a slight proliferative advantage over the rest of the cell population. Subsequently the fraction of blood cells derived from the mutated stem cell slowly expands over time making its acquired mutations detectable via sequencing of gDNA derived from peripheral blood [*16, 17*]. This condition was termed clonal hematopoiesis, derived from the observed clonally expanded hematopoietic lineage.

Clonal hematopoiesis itself does not represent a disease, but has been shown to be associated with a tenfold increase in the risk of developing blood cancer [*18*] as well as a two to four times increased risk for atherosclerotic cardiovascular diseases [*19*]. Due to these associated risks without the imperative advancement to diseased states, the presence of mutations associated with hematological neoplasia at a variant allele frequency of at least 2% in absence of any evidence of a hematological neoplasm has been termed Clonal Hematopoiesis of Indeterminate Potential (CHIP) [*20*].

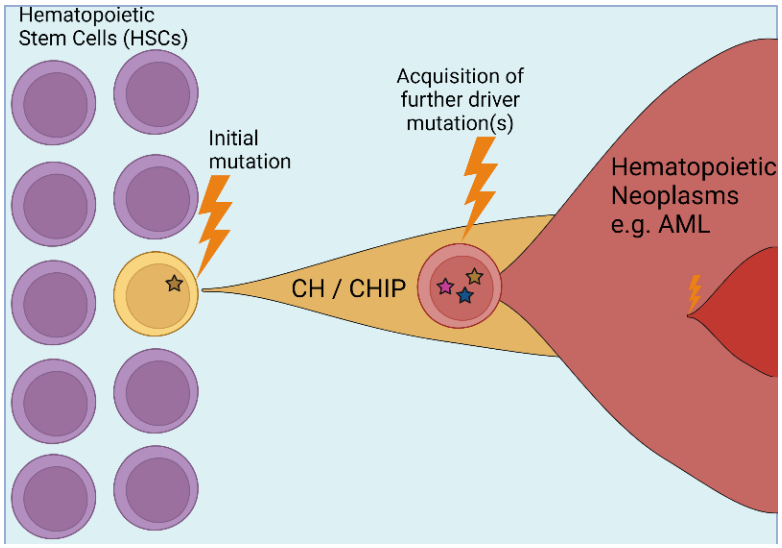


Figure 2: Clonal hematopoiesis as a precursor of hematopoietic neoplasms such as Acute Myeloid Leukemia (AML). Acquisition of a somatic mutation that confers slight proliferative advantage within a hematopoietic stem cell (HSC) leads to a clonal expansion of cells derived from this HSC – known as clonal hematopoiesis. Acquisition of further driver mutations can subsequently cause development of hematopoietic neoplasms such as Acute Myeloid Leukemia (AML). In many cases of AML clonal hematopoiesis precedes the actual disease and can therefore be seen as the start of the clonal evolution ultimately leading to leukemia.

While many individuals with detectable CH do not progress to further disease patterns it has been shown that leukemic mutations can precede the onset of acute myeloid leukemia (AML) years before diagnosis (Figure 2). Additionally over 80% of mutations detected in CHIP are related to leukemia- and lymphoma-associated genes [*21*]. Hence, these expanded HSCs can eventually gain further driver mutations that may ultimately lead to development of AML or other hematopoietic disorders, such as myelodysplastic syndrome [*22-24*]. Clonal hematopoiesis can consequently be seen as part of the clonal evolution within the hematopoietic system that precedes the initial diagnosis of AML. Therefore tools are necessary to enable investigation of these early stage processes before manifestation and diagnosis of AML in order to fully characterize clonal evolution from healthy towards leukemic cells.

2.3 Acute myeloid leukemia (AML) as framework to study intratumoral heterogeneity and clonal evolution

Acute myeloid leukemia (AML) is a subtype of hematopoietic cancers and represents the most common acute leukemia in adults. Over 20,000 new cases of AML are estimated within the USA for 2021 [*25*] with a median age of 68 years at diagnosis [*26*].

AML specifically affects the myeloid cell lineages within the hematopoietic system. Acquired mutations in either hematopoietic stem cells (HSCs) or myeloid progenitor cells result in the inability to further differentiate into functional cell types (Figure 3). Ultimately, the accumulation of immature myeloid blasts within the bone marrow leads to leukocytosis and bone marrow failure as well as anemia, thrombocytopenia and leukocytosis due to the decrease in matured functional cells like erythrocytes and thrombocytes [*27*].

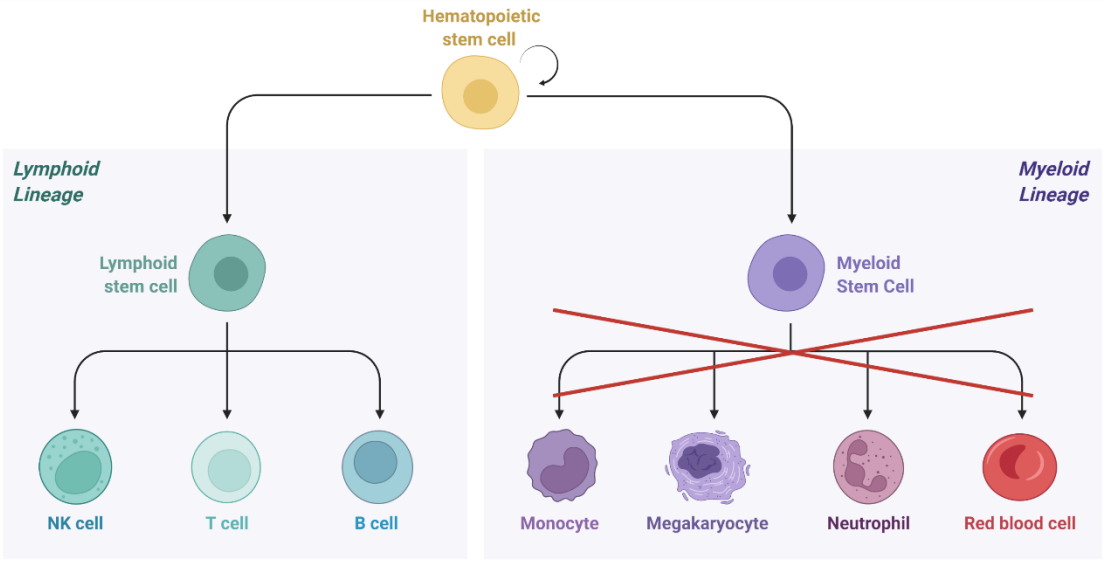


Figure 3: Basic hierarchy of hematopoietic cell types. Hematopoietic cells can differentiate into lymphoid or myeloid stem / progenitor cells, which furhter differentiate into functional cell subtypes within their lineage. In case of myeloid leukemias, acquired mutations lead to an inability to further differentiate within the myeloid stem cell compartment, resulting in accumulation of immature myeloid blasts and a decrease in terminally differentiated functional myeloid cells.

Although the median number of detected driver mutations within clinically reported AML cases is only at 4 with a maximum of ten driver mutations [*28*] about half of all AML cases show intratumoral heterogeneity, i.e. the presence of different leukemic subclones. These subclones have been shown to not only differ genetically but also phenotypically and functionally, e.g. in their differentiation potential [*29*]. As some of subclones, like those carrying an internal tandem duplication in the *FLT3* tyrosine kinase (FLT3-ITD), are associated with adverse outcome [*30, 31*], targeted therapies have been developed to directly target these adverse subclones [*32*] in addition to the standard chemotherapy that usually involves treatment using Cytarabine (Ara-C). Despite increased efforts to enhance AML therapies only one third of patients will survive five or more years after diagnosis due to relapses after initial treatment [*33*]. Therefore further research is needed to deepen the understanding of clonal evolution leading to AML in order to exploit more possibilities for its treatment and prevention.

2.4 Tracking of clonal evolution within patient samples by targeted re-genotyping of driver mutations

Identification of subclones within cancer samples is mostly based on the detection of genetic alterations that are present at different frequencies within the sample. Detected somatic mutations can subsequently be used to distinguish between different subclones. Many studies investigating subclonal architectures rely on whole exome sequencing (WES) or whole genome sequencing (WGS) to maximize the number of variants that can be used to distinguish subclones from one another [*34-36*]. Besides the associated high sequencing costs, these approaches are limited in their sensitivity and therefore often complemented by additional targeted deep-sequencing to reliably detect low frequency variants [*37-40*]. Combining *a priori* knowledge about known driver genes allows to restrict the sequencing to recurrently mutated genes, thereby decreasing associated sequencing costs and enabling higher coverage of disease-specific genomic variants known to functionally influence the cells' properties [*41*].

For the purpose of investigating the clonal evolution already before diagnosis of AML, a cost-efficient and sensitive method for targeted sequencing of human gDNA samples will be established. This method should allow the cost-efficient detection of variants down to low variant allele frequencies of 1-2% within recurrently mutated CHIP and AML driver genes to enable sequencing of larger cohorts, thereby allowing further insights into ITH and clonal evolution before and after diagnosis of AML.

2.4.1 Error-corrected sequencing enables reliable detection of subclonal, low-frequency genetic variants

Variants at low variant allele frequencies need to be reliably detected in order to identify minor subclones within cancer samples. However due to the reported typical error rate of around 1% [*42*] precise detection of variants below or close to 1% allele frequency remains challenging even at very high sequencing depths.

In the past years, several methods have been developed to increase the precision of next generation sequencing [*43-46*]. Nearly all of these approaches share the same basic idea of molecularly tagging target molecules. These molecular tags, usually called Unique Molecular Identifiers (UMIs) consist of random nucleotide sequences that are attached to the library fragments before amplification. Afterwards, the library fragments are amplified by PCR and sequenced. Hence all library fragments derived from the same original molecule, e.g. genomic DNA fragment, will carry the same UMI. Likewise, those

fragments derived from different template molecules will also differ in their UMI sequence. Upon analysis, the UMIs allow to determine which sequencing reads were derived from which initial template molecule. Consequently, this approach allows to remove any frequency biases introduced by PCR amplification of the sequencing library.

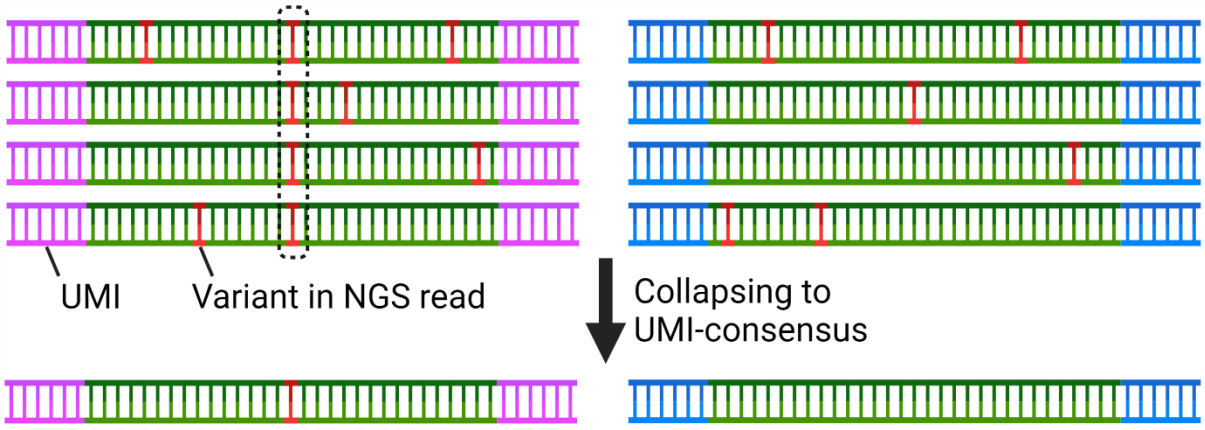


Figure 4: Utilization of Unique Molecular Identifiers (UMIs) can significantly decrease the number of false-positive variant calls in sequencing data. Library fragments derived from the same template molecule are tagged with the same UMI sequence (depicted in cyan and blue). During PCR amplification and sequencing stochastic errors can give rise to different variants present within the sequencing reads that cannot be distinguished from real variants (depicted in red). By generation of a consensus sequence per UMI, stochastic errors introduced by technical noise can be efficiently eliminated, thereby increasing precision of variant calling especially at low variant allele frequencies.

Additionally, when sequenced deep enough, every UMI will be sequenced with multiple sequencing reads enabling additional error-correction. The reads obtained per UMI can be subsequently used to assemble a UMI-consensus read (Figure 4) for example using a 'majority vote' method. If variants are only present in one or few reads carrying the same UMI, they are likely derived from technical artefacts such as PCR errors and will not be included in the consensus sequence. Only variants observed within the majority of sequencing reads belonging to the respective UMI will be kept. Hence, precision of sequencing is significantly increased resulting in approximately 20-fold lower error rates [*46*].

2.4.2 Target enrichment strategies for targeted sequencing

As high sequencing depth is necessary to robustly detect low frequency variants, a strategy for target enrichment is needed in order to reduce the sequencing to the target areas of interest. Here, two general approaches can be distinguished.

The first approach is direct PCR amplification of the genomic regions of interest. PCR reactions generally allow for highly specific amplification of genomic target areas, but is limited in terms of

multiplexing primers targeting different loci. Due to interactions between primer pairs off-target amplification as well as increased formation of primer-dimers often reduce the overall efficiency in multiplex PCRs [*47*], thereby limiting the number of loci that can be efficiently targeted within one reaction. As sequencing length is restricted to about 300 nucleotides on most Illumina sequencers [*48*], amplicons need to be small in order to be completely sequence. Consequently, multiple PCR reactions would be necessary in order to efficiently larger sets of genomic loci. This, in turn, increases the amount of template gDNA needed as well as costs for setting up the enzymatic reactions. Although modified approaches that allow to amplify hundreds of amplicons at once are commercially available, they are not easily customizable and require specialized equipment and primer libraries [*49*].

The second basic approach is enrichment of target fragments by hybridization-capture reactions. Here, most commonly, the gDNA is physically or enzymatically fragmented. The resulting DNA fragments are subsequently hybridized to oligonucleotide probes complementary to the genomic regions of interest. As the probes are biotinylated they can be used for affinity purification of hybridized DNA fragments using Streptavidin beads. The retained fragments are subsequently enriched for the genomic regions of interest. Although enrichment is less strong compared to targeted PCRs, the size of genomic regions that can be targeted within one reaction is usually not a limiting factor [*50, 51*].

For the planned targeted sequencing approach an intermediate size for target regions of several kilobases will be necessary in order to sequence the most important, highly recurrent known driver genes involved in clonal hematopoiesis as well as AML. Additionally, a strong target enrichment is desirable in order to maximize cost-efficiency for sequencing.

2.4.3 Single molecule Molecular Inversion Probes provide a highly sensitive and flexible approach for targeted re-genotyping

An attractive method for targeted re-genotyping that combines the advantages of both general enrichment strategies described above are Molecular Inversion Probes (MIPs). Molecular Inversion Probes are similar to Padlock Probes that were already described about 30 years ago [*52*] and were further refined to make use of massively parallel sequencing technologies as well as UMIs for enhanced precision in genotyping, since called single-molecule Molecular Inversion Probes (smMIPs) [*53*].

The principle of smMIPs represents a mixture of hybridization and PCR approaches for target enrichment. Molecular Inversion Probes are phosphorylated DNA oligonucleotides with a length of about 80 nucleotides. The ends of the probes are used to hybridize the probe to the target region of the

template DNA, hence called hybridization arms, and are connected via a backbone sequence that is constant for all probes (Figure 5). After successful hybridization, a polymerase can fill the gap between the hybridization sites by extending the double-stranded hybridization site of the extension-arm, thereby copying the sequence from the DNA template. A ligase subsequently joins the ends of the gap-fill sequence and the phosphorylated ligation-arm to produce a circularized DNA fragment. Afterwards, exonucleases are used to digest the linear gDNA template as well as probes that were not hybridized and circularized. Due to the absence of complex template DNA the remaining circular probes can be utilized within a highly specific PCR. Here, primers which introduce Illumina sequencing adapters anneal within the constant probe-backbone sequence in order to create the final sequencing library.

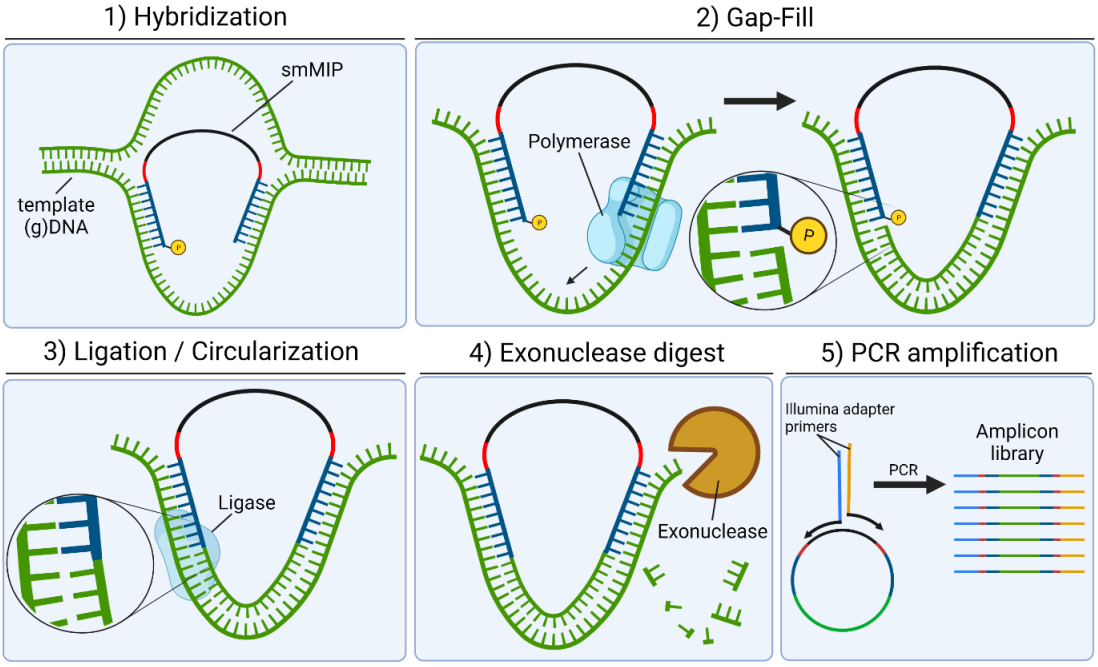


Figure 5: Overview over smMIP (single-molecule Molecular Inversion Probes) reactions. 1) Target-specific hybridization arms (blue), which are connected by constant backbone sequence (black), of the phosphorylated oligonucleotide-probe anneal to the complementary target region on the denatured template DNA. For error-correction the probes carry UMI sequences (red). 2) A polymerase is used to copy the sequence in the gap between the hybridization arms from the template DNA. 3) A thermostable Ligase circularizes the probe, including the captured sequence, by ligation. 4) Exonucleases are used to digest any remaining linear DNA, i.e. probes without successful target-capture and template DNA. 4) PCR amplification of probes carrying the captured target sequences using Illumina adapter primers creates the final amplicon sequencing library.

The smMIP hybridization arms have similar function as primers in targeted PCRs. However, in contrast to PCR both arms anneal to the same DNA strand, thereby enabling strand-specific targeting of DNA loci. Additionally, the hybridization arms are connected via the probe-backbone which decreases the likelihood of off-target hybridizations, as both annealing sequences need to be in close proximity within the target DNA. Previous studies have shown multiplexing of thousands of probes within one smMIP capture reaction, enabling highly flexible target sizes up to whole exomes [*54-56*].

Furthermore, published cost calculations demonstrate the high cost-efficiency of the approach, resulting in total costs for library preparation and sequencing below 15 USD [*57*]. Moreover, hybridization, gap-fill reaction, circularization of probes and digestion of linear DNA can be carried out in a single reaction, which minimizes hands-on times for library preparation and thus facilitates processing of up to tens of thousands of samples for large cohort studies [*55, 58, 59*].

Therefore single-molecule Molecular Inversion Probes represent an optimal approach for highly sensitive and cost-efficient high-throughput targeted re-genotyping. smMIPs will consequently be used to establish a new assay allowing the detection of low-frequency variants associated with clonal hematopoiesis and AML for further investigation of clonal evolution.

2.5 A mouse model system allows for genetic-engineering of patient-derived xenograft samples

Although analyses of subclonal compositions within patient samples can provide deeper insights into clonal evolution further experiments to phenotypically characterize subclones, e.g. their response towards newly developed drugs, are difficult to perform due to ethical restrictions. Several AML cell lines exist that enable *in vitro* experiments and allow for further characterization of the disease [*60, 61*]. However, many cancer cell lines fail to reflect the genomic and phenotypic heterogeneity observed within patients, as some subclones may be eradicated due to failure to survive within the artificial cell culture system whereas others may adapt to the altered environment, e.g. by changes in their gene expressions [*62*]. Although the use of patient-derived cells allows to closer recapitulate the situation within patients, most cells are hardly proliferating in *in vitro* cultures and can be cultured only for limited time [*63*]. Mouse models have proven to be valuable tool in order to allow for cultivation of patient-derived cancer cells while avoiding irreversible phenotypic changes and a loss of heterogeneity [*64, 65*]. Specifically, the use of NSG (NOD.Cg-*Prkdc*^{scid} *IL2rg*^{tm1Wjl}/SzJ) mice allowed to establish a mouse model system for patient-derived AML cells at the Helmholtz Zentrum München [*66*].

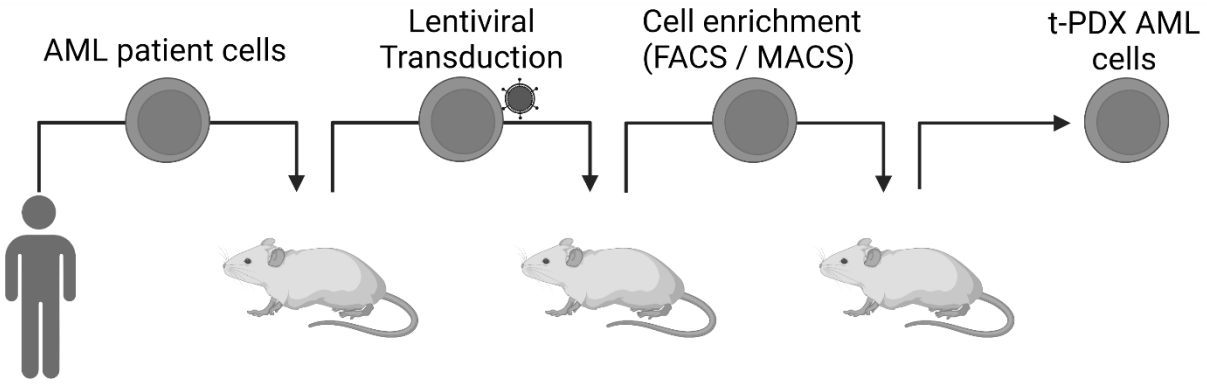


Figure 6: Mouse model for the generation of AML Pdx lines. Using the established model system leukemic cells from AML patients can be transplanted into immunodeficient mice. After initial expansion, the leukemic cells are lentivirally transduced to integrate marker genes which enable cell enrichment by FACS or MACS as well as in vivo bioluminescence measurements to non-invasively determine the leukemic burden within mice. The resulting transgenic patient-derived xenograft AML cells (t-PDX AML cells) preserve heterogeneity observed within the patient sample while allowing for efficient cultivation and further in vivo characterization.

After injection of leukemic cells into the blood-stream of these immunocompromised mice, the cells can home within the bone marrow, which provides a more natural environment and enables the cells to proliferate. After expansion of cells, the leukemic population can be re-isolated from bone marrow of the animal and subsequent cell sorting based on human cell surface markers (Figure 6). The isolated leukemic cells can be genetically engineered by lentiviral transduction, e.g. enabling expression of fluorescent proteins or additional surface markers. After an additional *in vivo* expansion, the transgenic patient derived xenograft (t-PDX) AML cells can be easily re-isolated by FACS or MACS enrichment and used for further experiments. Additionally, transgenic expression of luciferase within AML Pdx cells allows to determine the leukemic burden within experimental mice via *in vivo* bioluminescence imaging. Bleeding the mice in order to determine the fraction of human leukemic cells within peripheral blood can consequently be avoided.

This mouse model system enables long-term cultivation of patient-derived AML samples, which closely resemble the original disease within patients by retaining their subclonal heterogeneity [*66*]. Moreover, the possibility to genetically engineer Pdx samples via lentiviral transductions enables a multitude of experimental approaches to investigate the intratumoral heterogeneity which is usually lost when utilizing *in vitro* cultivations.

In order to expand the mouse model's ability to analyze the behavior of different subclones within AML Pdx cell populations a suitable method will be established that allows to distinguish between cells and enables to detect phenotypic differences among them.

2.6 Cellular barcoding as a tool to enable direct tracking of cell populations on a single cell level

One key methodology to analyze the composition of heterogeneous cell populations is called 'cellular barcoding', also known as "genetic barcoding'. The general principle is based on integration of artificial genetic markers into the cells' genomes, which will consequently also be inherited by their offsprings. Every cell receives a unique artificial gDNA tag that allows to easily distinguish between cells and their descendants within cell populations. Hence, this barcode significantly simplifies differentiation between similar cells that otherwise could not be easily distinguished genotypically and phenotypically (Figure 7). As the cellular barcode is inherited by all offspring cells, barcodes of cells that show increased fitness under the given environmental conditions will increase in their relative frequency within the barcoded cell population over time. Consequently, cellular barcodes allow for direct identification and observation of clonal heterogeneity within cell populations, making it a powerful tool for the study of heterogeneous cell populations.

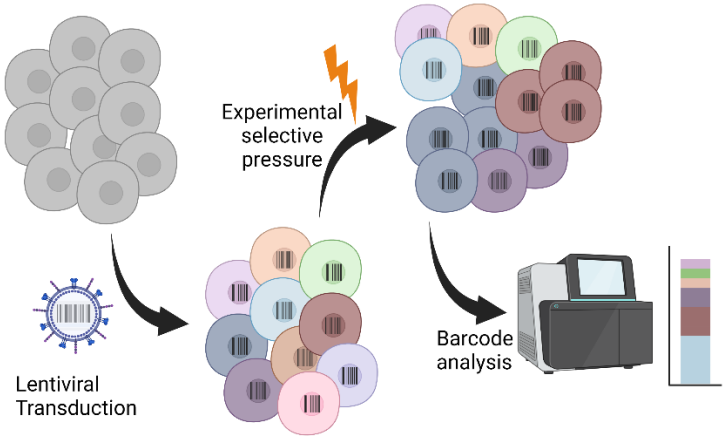


Figure 7: General approach of the cellular barcoding technique. A cell population is labelled with cellular barcodes via lentiviral transduction, integrating a unique artificial sequence tag into their genomes. Hence, all cells as well as their offspring can be distinguished based on their cellular barcodes. Cells can be exposed to selective pressures, e.g. toxins, to elucidate whether cells within the population show differences in cellular fitness. The cellular composition of the population can be determined by amplification of barcodes from isolated gDNA via PCR and subsequent sequencing.

The first application of cellular barcoding 30 years ago [*67*] used a library of 100 different DNA fragments to create a retroviral pool used to transduce cerebral cortical progenitor cells within early developing rat brains in order to track localization patterns of their progeny neurons. PCR amplification of barcodes from genomic DNA (gDNA) and restriction digest of the obtained amplicons was used to identify each of the barcodes representing the clonal compositions within different regions of the cortex. The observation that some barcodes could be observed in different functional areas of the cerebral cortex

demonstrated that specification of cortical areas occurs after neurogenesis without restrictions of cell intermixing during embryonic development.

With the emergence of next generation sequencing, allowing to easily sequence millions of different DNA sequences at once the genetic tags used for labelling the cells evolved into shorter barcode-like sequences by utilizing degenerate nucleotide positions for oligonucleotide synthesis [*68*]. Since then, cellular barcoding has been successfully used for research in many different contexts, especially for investigating differentiation patterns and heterogeneity within hematopoiesis [*69, 70*] as well as clonal dynamics in various cancers [*71-74*], including leukemias [*75-79*].

To allow direct investigation of intratumoral heterogeneity within Pdx AML samples the cellular barcoding technique will be established for use within the mouse model system at the Helmholtz Zentrum München.

3 Materials and methods

3.1 Materials

3.1.1 Enzymes

Table 1: Enzymes used in experimental procedures

Enzyme	Manufacturer	Cat.No.
Actinase E	Sigma-Aldrich	P5147-100MG
Ampligase	Biozym	114100
AvrII	New England BioLabs	R0174L
DreamTaq Polymerase	Thermo Fisher Scientific	EP702
Exonuclease I	New England BioLabs	M0293L
Exonuclease III	New England BioLabs	M0206L
Hemo KlenTaq Polymerase	New England BioLabs	M0332L
KpnI-HF	New England BioLabs	R3142L
NheI-HF	New England BioLabs	R3131L
Phusion II HotStart Polymerase	Thermo Fisher Scientific	F549L
Q5 Hot Start Polymerase	New England BioLabs	M0493L
rSAP	New England BioLabs	M0371S
SpeI-HF	New England BioLabs	R3133L
T4 DNA Ligase	New England BioLabs	M0202L
T4 DNA Ligase	New England BioLabs	M0202M
T4 Polynucleotide Kinase	New England Biolabs	M0201S

3.1.2 Buffers, Chemicals & Media

Table 2: Buffers and chemicals used in experimental procedures

Buffer/Reagent	Manufacturer	Cat.No.
Ampligase buffer	Biozym	115005
ATP	New England Biolabs	P0756S
Buffer EB	Qiagen	19086
CutSmart	New England Biolabs	B7204S
dNTPs	New England Biolabs	N0446S
DreamTaq Green Buffer	Thermo Fisher Scientific	B71
EDTA, 0.5M	Sigma-Aldrich	E7889-100ML
GeneRuler 100 bp DNA ladder	Thermo Fisher Scientific	SM0242
GeneRuler 100 bp Plus DNA ladder	Thermo Fisher Scientific	SM0322
Igepal CA630	Sigma-Aldrich	I8896-50ML
NEBuffer 1	New England Biolabs	B7001S
NEBuffer 3.1	New England Biolabs	B7003S
NEBstable/NEB10beta Outgrowth Medium	New England Biolabs	B9035S

(Table 2 continued)

Buffer/Reagent	Manufacturer	Cat.No.
PEG-8000	Sigma-Aldrich	89510-1KG-F
Phusion High-Fidelity buffer	Thermo Fisher Scientific	F518L
Q5 HotStart buffer	New England Biolabs	B9027S
Sera-Mag SpeedBeads	Sigma-Aldrich	GE65152105050250
Sodium Azide (NaN_3)	Sigma-Aldrich	S2002-100G
Sodium Chloride (NaCl), 5M	Sigma-Aldrich	S5150-1L
Tris-HCl, pH 8.0, 1M	Sigma-Aldrich	T2694-100ML
UltraPure Water	Thermo Fisher Scientific	10977035

3.1.3 Commercial Kits

Table 3: Commercial kits utilized within experimental procedures

Kit	Application	Manufacturer	Cat.No.
DNeasy Blood& Tissue Kit	gDNA extraction	Qiagen	69504
QIAamp DNA Micro Kit	gDNA extraction	Qiagen	56304
Quant-iT PicoGreen dsDNA Assay	Quantification of DNA	ThermoFisher Scientific	P7589
PureYield Plasmid Midiprep System	Isolation of plasmid DNA	Promega	A2495

3.1.4 Oligonucleotides

Table 4: Sequences of utilized oligonucleotide in 5'-3' orientations. '/Phos-/' indicates a 5'-phosphorylation of the respective oligonucleotides. Illumina adapter primers ('[*]') carry different eight nucleotides long index sequences denoted as 'XXXXXXXX'. The respective index sequences are listed in Supplemental Tables 2 & 3. Sequences of smMIP oligonucleotide probes are listed separately in Supplemental Table 1.

Name	Sequence	Manufacturer	Purity	Usage
DNABC_NT_screen_fwd	AGTGAACGGATCTCGACGGT	Integrated DNA Technologies	Desalted	Screening plasmid for presence of DNA barcode insert
DNABC_NT_screen_rev	CCTTCTCTAGGCACCCGTTC	Integrated DNA Technologies	Desalted	Screening plasmid for presence of DNA barcode insert
ExBC_AmpSeq_fwd	CTGGTACCTTTAAGACCAATGACT	Integrated DNA Technologies	Desalted	Screening plasmid for presence of expressed barcode insert
ExBC_AmpSeq_HP_fwd	GGACACTCTTTCCCTACACGACGC TCTTCCGATCTNNNNNNNNNNNN ATGGGAAAGAGTGTCCCTGGTACC TTTAAGACCAATGACT	Sigma Aldrich	HPLC	amplification of expressed barcodes from plasmid or lentiviral inserts

(Table 4 continued)

Name	Sequence	Manufacturer	Purity	Usage
ExBC_AmpSeq_HP_rev	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCTTAAGCAGTGGGTTCCCT	Sigma Aldrich	HPLC	amplification of expressed barcodes from plasmid or lentiviral inserts
ExBC_AmpSeq_rev	GCTTAAGCAGTGGGTTCCCT	Integrated DNA Technologies	Desalted	Screening plasmid for presence of expressed barcode insert
ExprBC_1st	/Phos-CTTTAAGACCAATGACTTACAAGGCNNNNTTNNNAANNNTTAGCTGTAGATG	Sigma Aldrich	HPLC	1st strand of expressed barcode insert
ExprBC_2nd	/Phos-CTAGCATCTACAGCTAANNNTTNNNAANNNNGCCTTGTAAGTCATTGGTCTTAAAGGTAC	Sigma Aldrich	HPLC	2nd strand of expressed barcode insert
HP_N7 adapters [*]	CAAGCAGAAGACGGCATACGAGAT[XXXXXXXX]GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Sigma Aldrich	HPLC	Illumina P7 adapters for barcode libraries
NT+RS P7 xGen adapters [*]	CAAGCAGAAGACGGCATACGAGAT[XXXXXXXX]GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG	Sigma Aldrich	Cartridge	Illumina P7 adapters for smMIP libraries
TruSeq P5 xGen adapters [*]	AATGATACGGCGACCACCGAGATCTACAC[XXXXXXXX]ACACTCTTTCCCTACACGACGCTCTTCCGATCT	Sigma Aldrich	Cartridge	Illumina P5 adapters for smMIP & barcode libraries

3.1.5 Bacterial strains

Table 5: Bacterial E.coli strains used for transformation and propagation of plasmids during experimental procedures.

Strain	Genotype	Manufacturer	Cat.No.
NEB 10-beta Electrocompetent E. coli	*Δ(ara-leu) 7697 araD139 fhuA ΔlacX74 galK16 galE15 e14- Φ80dlacZΔM15 recA1 relA1 endA1 nupG rpsL* (Str^R) *rph spoT1 Δ(mrr-hsdRMS-mcrBC)*	New England Biolabs	C3020K
NEB Stable Competent *E. coli* (High Efficiency)	F' *proA+B+ lacI^q Δ(lacZ)M15 zzf::Tn10* (Tet^R)/ *Δ(ara-leu) 7697 araD139 fhuA ΔlacX74 galK16 galE15 e14- Φ80dlacZΔM15 recA1 relA1 endA1 nupG rpsL* (Str^R) *rph spoT1 Δ(mrr-hsdRMS-mcrBC)*	New England Biolabs	C3040I

3.1.6 Scientific instruments

Table 6: Scientific instruments utilized throughout experimental procedures.

Type	Instrument name	Manufacturer
Centrifuge	Centrifuge 4K15	Sigma Aldrich
Centrifuge	Centrifuge 5424	VWR Peqlab
Capillary gel-electrophoresis	Bioanalyzer 2100	Agilent Scientific Instruments
Electroporator	ECM 600	Genetronics, BTX
Incubator	Innova 42	New Brunswick
Mini centrifuge	SPROUT	Biozym
PCR workstation	PCR Workstation Pro	VWR Peqlab
Plate-Centrifuge	PerfectSpin P	VWR Peqlab
Platereader	POLARstar OPTIMA	BMG Labtech
Platereader	Infinite 200 PRO	Tecan
Spectrophotometer	ND-1000	NanoDrop Technologies
Thermocycler	SimpliAmp Thermal Cycler	Eppendorf
Thermocycler	peqSTAR 96 Universal Gradient	VWR Peqlab
Vortexer	Vortex-Genie 2	Scientific Industries SI
Waterbath	MB-5	Julabo

3.2 Methods

3.2.1 Preparation of homemade SPRI beads for purification of DNA

For cleanup of DNA, e.g. PCR-fragments, homemade SPRI beads were prepared according to Rohland & Reich, 2012 [*80*].

Table 7: Composition of the PEG solution used for preparation of homemade SPRI beads containing 22% PEG.

Ingredient	Quantity
5 M NaCl	10 ml
1 M Tris-HCl, pH 8.0	500 µl
0.5 M EDTA	100µl
PEG 8000	11 g
10 % Igepal CA630	50 µl
10 % Sodium Azide	250 µl
UltraPure Water	ad 49 ml

A PEG solution according to Table 7 was prepared and incubated at 40°C until all PEG was dissolved. 1 ml of Sera-Mag Speed Beads (Thermo Fisher Scientific) was transferred to a 1.5 ml Eppendorf cup and put onto a magnet rack. After separation of beads from the solution, supernatant was removed. The pellet was washed two times by resuspension in 1x TE buffer (10 mM Tris-HCl, pH 8.0, 1 mM EDTA). The pellet was eluted using 0.9 ml 1x TE buffer, resulting in a final volume of 1ml eluate. The resulting bead supsension was added to the PEG solution and vigourously mixed by vortexing. The final bead solution was stored at 4°C until usage.

3.2.2 DNA purification using homemade SPRI beads

Before usage of the homemade SPRI beads, the suspension was equilibrated at room temperature and vortexed until magnetic particles were fully resuspended. For purification of DNA derived from PCR or enzymatic reactions the respective samples were mixed with the SPRI beads suspension in a volume ratio of 0.8:1 to 1.8:1 beads to reaction volume, depending on the desired size-cutoff for short DNA fragments. The solution was mixed by pipetting and incubated for 5 minutes at room temperature. The sample vessel was transferred to a magnetic stand and incubated for 2 to 5 minutes until the solution was clear. The supernatant was discarded and the pellet of SPRI beads was washed by addition of 200 µl 80% ethanol. If the initial volume exceeded 200 µl ethanol volume was increased to the initial sample

volume in order to wash the whole pellet. The ethanol was discarded and the wash step was repeated. SPRI bead pellets were allowed to dry for 2 to 5 minutes until the surface of the pellet was dried. The sample was put off of the magnetic stand, resuspended in 10 to 25 µl of ddH_2O or buffer EB for elution and incubated for 5 to 10 minutes for elution. After SPRI beads had been completely separated, the sample was put back on a magnetic stand and the supernatant containing the clean DNA was transferred to a new vessel.

3.2.3 Design of a single-molecule molecular inversion probes panel

For the design of smMIPs the software MIPgen [*81*] was used. Initial parameters used (Table 8) comprised a target capture size of 120 nucleotides in order to produce only one amplicon size after PCR amplification at library preparations, hence allowing to easily identify off-target hybridizations with differing capture lengths.

Table 8: Parameters used for the MIPgen software to design the initial set of potential smMIP probe candidates.

Parameter	**Value**	**Description**
min_capture_size	120	Minimum length of target sequence per probe
max_capture_size	120	Maximum length of target sequence per probe
tag_sizes	5,5	Include 5 random nucleotides on each hybridization arm
double_tile_strands_separately	on	Create overlapping probes on (+)-strand and (-)-strand
score_method	svr	Predict probe performance based on support vector regression

As some target regions could not be covered using this parameter set, MIPgen was run again using a relaxed target size of 110 – 130 nucleotides. The resulting pool of potential probes covered both strands of the target regions. In order to reduce the initial costs for the panel, the probe pool was reduced to cover only one strand per position, with neighboring probes being located on alternating strands, thereby avoiding sterical hindrance of probes at hybridization sites (Figure 22 A). Probes for the final panel were picked based on MIPgen's predicted performance score and the prerequisite to have a small overlap

with their neighboring probes. For smMIPs having known SNPs within their hybridization arms, two probes were included in the final set – with and without the known polymorphism – in order to provide robust performance.

Additionally to targeting-arms and UMIs the MIP-probes comprise a fixed backbone sequence used for annealing the primers in subsequent PCR amplifications after target capturing. However, the backbone sequence does not contain any of Illumina's standard read-start priming sequences, making it necessary to use custom read-start primers when sequencing smMIP amplicons. Therefore, the original smMIP-backbone sequence was re-designed and replaced by 'AGATCGGAAGAGCGTGTGTATAAGAGACAG' to allow for Illumina sequencing utilizing standard read-start primers. The initial pool of 303 oligonucleotide probes was ordered at SigmaAldrich at a synthesis scale of 25 nmole with subsequent reverse-phase chromatography cleaning ('cartridge purification'). A list of all probes contained within the final panel after rebalancing (see 4.2.2 and 4.2.3) is included in Supplemental Table 1.

3.2.4 Pooling and phosphorylation of smMIP probes

The initial pool of smMIP probes was pooled at equimolar ratios. Phosphorylation was carried out in 50 µl reactions of 1x T4 DNA Ligase Buffer, 12 units T4 Polynucleotide Kinase and 25 picomole of the pooled smMIP probes. After incubation at 37°C for 2.5 hours in a thermocycler the reaction was heat-inactivated at 65°C for 25 minutes. The probe pool was subsequently diluted to a 10x working-concentration of 503 pM per smMIP using buffer EB and split up into single-use aliquots in order to avoid freeze-thaw cycles that potentially could impair the phosphorylation of the probes.

For refinement of the smMIP panel via re-balancing the probes were subset according to the planned relative concentrations within the final pool, phosphorylated in independent reactions and pooled together afterwards.

3.2.5 Preparation of smMIP sequencing libraries

In order to minimize the risk of cross-contaminations all reactions were set up within a PCR workstation, which was decontaminated using DNA-Away and UV-light before and after each usage.

gDNA samples were first diluted to a concentration of 8 ng/µl in order to normalize the input for hybridizations to 100 ng genomic DNA within 12.5 µl. Samples that already had concentrations slightly below a concentration 8 ng/µl were used undiluted.

Hybridizations and target captures were carried out in 25 µl reactions with 1x Ampligase buffer, 1 unit Ampligase, 0.32 µl Hemo KlenTaq Polymerase, 320 nM dNTPs, 100 ng gDNA template and 5 µl of the 100 pM smMIP pool, corresponding to a ratio of 1000 molecules of each smMIP probe per haploid genome within the reaction. Reactions were subsequently cycled within a thermocycler with an initial denaturation of 95 °C for 10 minutes, followed by 60 °C for 18 hours for hybridization and gap-filling and subsequent storing at 8°C.

After Hybridization and Gap-Fill reactions finished, an exonuclease treatment was performed in order to digest non-circularized probes as well as the gDNA template. 2 µl of an exonuclease mix, consisting of 10 units Exonuclease I and 50 units Exonuclease III in 1x NEBuffer 1 were added to each reaction, keeping the reactions on the cooled block of the thermocycler. After mixing, the reaction were incubated for 1 hour at 37 °C, followed by heat-inactivation at 80 °C for 25 minutes and subsequent storage at 8°C.

In order to utilize as much hybridization products as possible, two PCR amplification reactions per hybridization reaction were set up for library construction. Each PCR reaction consisted of 1x Q5 reaction buffer, 1 unit Q5 HotStart High-Fidelity DNA Polymerase, 200 µM dNTPs, 1 % DMSO, 250 nM of both P5- and P7-adapter primers (‘NT+RS P7 xGen adapter’ & ‘TruSeq P5 xGen adapter’, see Table 4) and 12.5 µl of the exonuclease-treated hybridization reaction in a total volume of 50 µl. PCR reaction were cycled using 95 °C for 2 minutes for initial denaturation and 25 amplification cycles consisting of 95 °C for 15 sec, 47.5 °C for 30 sec and 72 °C for 30 sec. Following a final elongation for 2 minutes at 72 °C, samples were stored at 8 °C.

PCR products derived from the same hybridization reactions were pooled and purified twice using homemade SPRI beads at a volume ratio of samples to beads of 0.8:1. The second clean-up was necessary to further minimize unwanted by-products derived from amplification of smMIPs that were circularized without capturing a target, thereby producing amplicons of similar size.

Samples were quantified using Quant-iT PicoGreen dsDNA Assay (ThermoFisher) picoGreen and subsequently pooled in equimolar ratios. The final library pools were additionally quantified using a High Sensitivity DNA Chip on an Agilent Bioanalyzer.

For each sample within the final cohort sequencing study smMIP libraries were prepared in two technical replicates thereby by utilizing a total of 200 ng gDNA per individual, in order to increase the unique coverage across target areas.

Initial smMIP libraries for balancing of the probe pool were either sequenced on an Illumina MiSeq (dual-indexed 150B paired-end) or on an Illumina Hiseq1500 (dual-indexed 100b paired-end).

The first five library batches of the cohort study were sequenced on an Illumina HiSeq 1500 at the GeneCenter (Ludwig-Maximilians University Munich) within a dual-indexed 100B paired-end sequencing run. Due to discontinued support of the HiSeq 1500, batch six as well as batch seven, which contained additional quality controls, were sequenced on a NextSeq 1000 using the same sequencing setup.

3.2.6 Balancing of smMIPs to improve uniformity of coverage across target areas

Initially, all probes were pooled at equimolar ratios in order to test their performance based on hybridization efficiencies. Library preparations were carried out using genomic DNA isolated from cultured B-Lymphocyte cells of the GM18505 cell line [*82*] via the DNeasy Blood&Tissue Kit (Qiagen) according to the manufacturer's protocol. NGS data was analyzed using the MIPgen analysis tools [*81*] to determine the number of UMIs detected per smMIP probe and their relative performance within the probe pool. All probes having more than 2.5-fold fewer detected UMIs than the median across the panel were defined as bad performers. Subsequently a new probe pool was created, including a 25-fold or 50-fold increase of relative concentration within the pool for these bad performers. Library preparations, sequencing and analysis were repeated using the rebalanced probe pool. A subset of probes did not benefit from higher concentrations and still showed bad hybridization efficiencies as indicated by a low number of detected UMIs. Therefore, 47 probes were ordered to replace these bad performing probes. Additionally, all probes targeting the *CEBPA* gene were excluded due to bad performance and increased tendency towards self-circularization.

The set of smMIPs included in the final panel, as well as their relative concentrations within the pool are shown in Supplemental Table 1.

3.2.7 Design of Illumina sequencing adapters for smMIP libraries

As only a few million reads per smMIP sequencing library are necessary to achieve sufficient coverage across the panel, many samples can be multiplexed on one lane. Therefore enough adapter-primers carrying different sequencing indices to discriminate the individual samples need to be available. For this purpose, new adapters compatible with the re-designed smMIP backbone were ordered. 192 P5-

adapters were utilized, using the standard TruSeq-P5-adapter sequence (AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCG ATCT) but with indices derived from xGen Dual Index UMI Adapters (Integrated DNA Technologies) [*83, 84*], as the original set of TruSeq indices comprised only 8 sequences. The index sequences were further filtered based on published recommendations for Illumina Index Design [*85*]. Index sequences starting with "AC" and indices containing homopolymers of three or more nucleotides were excluded. All remaining adapter sequences were checked for their Gibson free energy using Quickfold [*86*] (http://www.unafold.org/Dinamelt/applications/quickfold.php). Adapters showing additional secondary structures due to their index sequence were excluded from further processing. The remaining index sequences were filtered for a minimum hamming distances of 3 among all indices and the first 192 adapters remaining were used as final set of adapters.

Another set of 96 P7-adapters was ordered based on the sequence of Nextera-P7-adapters, additionally including the mosaic sequence that would normally introduced by the Transposase in Nextera workflow library preparations (*CAAGCAGAAGACGGCATACGAGAT*[i7]*GTCTCGTGGGCTCGGAGATGTGTA TAAGAGACAG*). As the mosaic sequence contains the priming site for Illumina's read start primers, this sequence is necessary to enable processing of prepared libraries on Illumina sequencers using standard workflows and primers. Again, indices were derived from xGen Dual Index UMI Adapters, as the original set of 29 indices provided by Illumina was too small to multiplex enough samples. Selection of most suitable index sequences was carried out as for P5 adapters, except for skipping the initial filtering for indices starting with 'AC'.

All oligonucleotides were ordered at Sigma Aldrich in batches of 96 oligonucleotides at 50 nmole synthesis scale, purified using reverse-phase cartridge purification.

3.2.8 Setup of a custom analysis pipeline for smMIPs sequencing data

In order to process hundreds of samples an automated analysis pipeline was set up. First, samples get demultiplexed based on their sample indices using deML [*87*] version 1.13 and output files are named in a standardized way.

Afterwards, hundreds of samples can be submitted for processing using a single Bash script that utilizes the SLURM workload manager in order to optimize the resource management of the server, thereby minimizing the hands-on time necessary for processing of the sequencing data.

In this workflow, it is assumed that for every biological sample two technical replicates of the smMIP libraries were prepared. First, UMIs are extracted using fastp (version 0.20) by cutting the first five nucleotides of all reads and moving them to the respective header in the fastq file. This is carried out for both technical library replicates (see 3.2.5) of the sample. The resulting fastq files are modified to extend the UMI by 'GC' or 'AT' using 'sed'. Afterwards fastq files of both technical replicates can be merged without the risk of UMI collisions. Additionally, artificial UMI elongation preserves the information from which technical replicate a specific read was derived and can potentially be utilized for downstream analysis.

Hybridization arm sequences are trimmed using using cutadapt [*88*] (version 3.6.8) in linked adapters trimming mode with the following additional parameters*: -e 0, -O 16 ,--no-indels, --discard-untrimmed, --pair-adapters –g [R1-smMIP_arms.txt] –G [R2- smMIP_arms.txt]*. Linked adapter trimming mode requires matches for both hybridization arms derived from one smMIP to be detected within a read-pair in order to trim the sequences. All reads having no valid pair of arm-sequences detected are discarded.

Reads are mapped via BWA-MEM (version 0.7.17) to the human genome version hg19 using standard settings. Resulting SAM files are converted to sorted and indexed BAM files using samtools (version 1.8). Sample names are added to the BAM's SM fields via *Picard AddorReplaceReadGroup*. Base recalibration is calculated using *GATK BaseRecalibrator* (GATK version 4.1.6.0), utilizing SNP information from dbSNP [*89*] (build 138) and ExAC non-TCGA sites (release r0.3.1), as well as InDel information from 'Mills and 1000Genomes Gold Standard InDels'. Additionally 'NotDuplicateReadFilter' is disabled in order to utilize and retain all reads. The calculated recalibration table is applied using *GATK ApplyBQSR*.

Variant calling is subsequently carried out using *Mutect2* in single-sample mode with the following parameters in order to increase sensitivity and obtain a list of all variants possibly present within the sample: "*--interval-padding 10, --max-reads-per-alignment-start 0, --disable-read-filter NotDuplicateReadFilter, --f1r2-max-depth 500000, --minimum-allele-fraction 0.002, --genotype-germline-sites true, --mitochondria-mode true, --force-active true*". Calling of variants is subset to the regions of interest by using a bed file containing the panel's target coordinates in order to reduce processing time. The resulting variants are filtered using *GATK FilterMutectCalls* with *"--disable-read-filter NotDuplicateReadFilter*" and "*--mitochondria-mode true*".

Variants are filtered for allele frequencies above 0.2% and rows containing multiallelic sites are split and, if necessary, re-normalized using bcftools (version 1.8). The resulting variants are annotated with ANNOVAR [*90*] using the following databases: refGene [*91*], exac03, avsnp147, dbnsfp30a, clinvar_20190305, gnomad211_genome, cosmic91_coding.

For the generation of a position-specific error rate the mapped and base-recalibrated bam file (see above) gets de-duplicated based on UMI information using gencore [*92*] with the following parameters: "*--supporting_reads 2, --ratio_threshold 0.51, --umi_diff_threshold 0*". Samtools (version 1.8) is used to sort the resulting bam file and subsequently produce pile-ups using the following parameters: *"-d 0 -B -Q 0*". Error rates per positions are finally calculated by counting the number of matches and mismatches at each position within all sequenced samples. Positions that showed more than 5% mismatches within an individual sample were excluded for the calculation of the overall error rate, as they likely represent true somatic variants rather than background errors.

In order to utilize the UMI information within the data, the final set of variants obtained from *Mutect2* was used as input for *umivariants*[*93*]. Here, UMI consensus sequences were determined in "*majority*" mode and UMIs observed with only one sequencing reads are ignored. The annotated *Mutect2* variant set was merged with *umivariants'* output table in order to create the final raw set of variants for each sample.

3.2.9 Filtering of raw variant call sets derived from smMIPs

In order to exclude SNPs, potential artifacts and irrelevant variants preliminary filtering criteria were established based on the data collected from the cohort sequencing study: Synonymous and intronic variants were excluded from further analysis. Remaining variants need to be supported by at least 3 UMIs, as determined using umivariants. To further reduce technical noise, all variants below 0.5% variant allele frequency were also excluded. Furthermore, all variants that are not associated to leukemia were filtered for a Benjamini-Hochberg adjusted p-value smaller than 0.005. To further reduce technical noise, variants that are covered on both DNA strands were excluded, if variant allele frequencies differed more than five-fold between the strands.

All variants above 80% were excluded as likely representing homozygous SNPs. Variants having no association to leukemia and more than 0.01 frequency within ExAC being detected with more than 40% VAF were filtered as potential SNPs. Similarly all variants with more than 30% variant allele frequency having a frequency of at least 0.01 within the gnomAD database were filtered as potential SNPs. Variants tagged with "*strand_bias*" by *Mutect2*, were also flagged as potential artefacts within the set of filtered variants.

A set of 7 loci comprising 13 nucleotides had to be excluded from variants calls (Table 9). These loci showed recurring artefacts that were not excluded by the previous filtering, resulting in repeated artefactual low-frequency variants.

Table 9: Genomic (hg19) positions of artefact loci that were excluded from the final analyses. High recurrence of variants at these positions indicated artefacts not removed by filtration using the position-specific error rates.

Chromosome	Position	#variants after filtering	median VAFs [%] (min - max)
chr2	25467481-25467484	22	0.63 (0.50 - 3.52)
chr4	106162600	10	0.58 (0.50 - 1.57)
chr4	106164949	21	1.42 (0.83 - 3.90)
chr4	106196299-106196302	210	0.97 (0.50 - 5.88)
chr17	7577042	22	0.80 (0.52 - 1.65)
chr20	31022677	7	1.06 (0.63 - 1.38)
chr20	31022794	8	0.95 (0.51 - 2.55)

For comparison of re-sequenced patients (see 4.2.5.3) a strict filtering was chosen in order to focus on high confidence variants. All variants were additionally filtered for variant allele frequencies being greater than 1% and having reported a STRANDQ value greater than 70 by *Mutect2*. In order to verify absence or presence of variants missing within the variant call sets Integrated Genomics Viewer [*94*] (IGV, version 2.3.46) was used to examine the BAM files of raw and de-duplicated reads.

3.2.10 Cloning of the expressed barcode plasmid pool

The expressed barcode plasmid library was based on a pCDH-derived vector backbone. In short, the GLuc coding sequence in pCDH-EF1α-GLuc-T2A-NGFR (Addgene #104832) [*95*] was replaced by a gBlock encoding for H2Kk using the restriction enzymes EcoRI and BamHI. The NheI restriction enzyme site was destroyed by site-directed mutagenesis. Additionally, a small PCR fragment was cloned into the vector by using KpnI and PciI restriction enzymes, in order to introduce an additional AvrII restriction site next to the KpnI site. The preparation of the resulting pCDH-EF1α-H2Kk-T2A-NGFR vector was carried out by Christina Zeller at the Helmholtz.

The vector was subsequently digested in multiple parallel reactions each containing 1.5 µg vector 5 units AvrII and 10 units KpnI in 1x CutSmart buffer. Reactions were incubated for four to six hours at 37°C. The linearized backbone was purified using a SPRI bead clean-up at a ratio of 1:1.

The expressed barcode insert was prepared by annealing two complementary HPLC purified, phosphorylated oligonucleotides (ExprBC_1st and ExprBC_2nd, Sigma Aldrich, Table 4) in a 20 µl reaction, containing 0.5x NEBuffer 3.1 and 10 µl of each 10 µM oligonucleotide, in a PCR cycler by heating the reactions to 90 °C and decreasing the temperature by 0.1 °C every 10 seconds for 700 times,

followed by cooling to 8 °C. The annealed double-stranded insert includes sticky-ends to allow direct usage in downstream ligations without the need of an additional enzymatic digest.

For cloning the barcode insert into the vector a cut-ligation was performed overnight in order to maximize yields and minimize the amount of re-circularized vector not carrying a barcode insert. Seven 20 µl reactions containing 1.5 mM ATP, 250 fmol of the double-stranded barcode insert, 120 ng of pre-digested vector backbone, 2.5 units AvrII, 10 units NheI-HF and 400 units T4 Ligase (400 U/µl) in 1x CutSmart buffer were set up. The reactions were incubated in a PCR cycler using the following program: 55 cycles consisting of 5 minutes at 37 °C and 5 minutes at 20 °C, followed by 30 minutes at 37 °C, 20 minutes at 80 °C and a final storage at 8 °C.

The reactions were pooled and 5 µl were used to transform each of 16 transformation reactions using chemically competent NEB stable cells according to the manufacturer's protocol. After 1 hour of outgrowth 1% of four of the 16 transformations were used for plating as a proxy to determine overall transformation efficiencies by colony counts. Four Midi cultures (90 ml LB medium with 100 mg/ml Ampicillin) were inoculated with 4 transformation reactions each and incubated overnight at 37°C and 220rpm. Liquid cultures were pelleted by centrifugation at 4000 xg for 10 minutes at 8 °C. Plasmid isolations were carried out using the PureYield Midi Prep Kit (Promega) according to manufacturer's protocol. Finally, the four plasmid preparations were pooled in equimolar ratios to create the final plasmid pool for the expressed barcode.

3.2.11 Cloning of the high-complexity DNA barcode plasmid pool

The initial preparation of the pCDH-EF1α-GLuc-T2A-mtagBFP was carried out by Christina Zeller from the Helmhotz Center Munich by PCR-amplifying the coding sequence of mtagBFP and introducing EcoRI and SalI restriction sites via primer-overhangs. Subsequently, the PCR product was cloned into the pCDH-EF1α-GLuc-T2A-copGFP plasmid [*96*] using EcoRI and SalI, thereby replacing the copGFP with mtagBFP. Lastly, the NheI restriction site was destroyed by site-directed mutagenesis. The resulting vector was used for cloning the DNA Barcode plasmid pool.

The vector pCDH-EF1α-GLuc-T2A-mtagBFP vector was digested and dephosphorylated in multiple 50 µl reactions each containing 20 units SpeI-HF, 2 units rSAP and 1.1 µg of the prepared vector in 1x CutSmart buffer. Reactions were incubated at 37 °C for three hours and subsequently heat-inactivated at 80 °C for 20 minutes. Reactions were pooled and purified using the SV Gel and PCR Purification Kit (Promega) according to the manufacturer's protocol.

Insertion of the barcode insert into the vector backbone was carried out using a cycling restriction and ligation ('Cut-Ligation') approach in order to maximize the yield of correct ligation products. For this, reactions were set up containing 1x CutSmart buffer, 1.5 mM ATP, 10 units SpeI-HF, 10 units NheI-HF, 2,000 units T4 DNA Ligase, 120 ng of pre-digested vector backbone and 2 µl of the 1:10 diluted barcode annealing reaction, corresponding to 250 fmol insert, in a total volume of 20 µl per reaction. Reactions were transferred to a thermocycler and cycled 55 times with 5 minutes at 37 °C and 5 minutes at 20 °C, followed by a final heat-inactivation at 80 °C for 20 minutes and subsequent sample storage at 8°C. In order to reduce unwanted ligation products, containing either no or multiple barcode inserts, 10µl of an 'After-Cut' mixture consisting of 10 units NheI-HF and 10 units SpeI-HF in 1x CutSmart buffer were added per reaction and samples were further incubated for 2 hours at 37 °C with subsequent heat-inactivation for 20 minutes at 80 °C.

Reactions were pooled and purified using homemade SPRI beads at a sample to beads ratio of 1:1 and the clean product was quantified via picoGreen according to the manufacturer's protocol. The resulting vector was transformed via electroporation using NEB10beta cells taking special care to optimize conditions in order to achieve high transformation efficiencies. Hence, electroporation cuvettes as well as 0.5 ml Eppendorf cups and the cloned vector were cooled on ice. Additionally, 1 ml aliquots of NEB10stable/NEB10beta Outgrowth Medium in 2 ml low-bind Eppendorf cups were pre-warmed in a ThermoMixer at 37 °C. NEB10beta cells were thawed on ice and aliquoted á 25 µl into the pre-chilled 0.5 ml cups. 2 µl of the pre-cooled vector, corresponding to about 150 ng of plasmid, were added to each aliquot of competent cells and gently stirred using the pipet tip. The mixture was quickly transferred to the electroporation cuvette, taking care not to introduce bubbles. Immediately after electroporation at 2.1 kV with 48 Ohm, resulting in an effective pulse of 2.16 kV for 2.5 ms, 1 ml of the pre-warmed outgrowth medium was added and the cuvette was inverted six times. As much as possible of the suspension was transferred back into the pre-warmed 2 ml cup and immediately put into an incubator at 37 °C and 250 rpm. Cups were incubated horizontally in order to increase the air-medium boundary. After 1 hour of incubation, 0.025% of the transformation reactions were plated on LB-Agar plates containing 100 ng/ml Carbenicillin. Midi cultures containing 100 ml LB supplemented with 100 ng/ml Carbenicillin were each inoculated with the rest of two transformation reactions. Midi cultures were incubated overnight for 16 hours and 250 rpm at room temperature. Cultures were harvested by centrifugation at 4000 rcf for 30 min at 8 °C. The resulting cell pellets were used for the purification of the plasmids utilizing the PureYield Midi Prep Kit (Promega) according to manufacturer's protocol. The final Plasmid pool was created by pooling the plasmid preparations considering the complexities of the included transformation reactions as estimated by their colony counts in order to achieve a uniform frequency distribution of the barcodes within the pool.

3.2.12 Colony PCR to screen for the presence of the expressed barcode

To determine the fraction of plasmid not carrying the barcode insert, colony PCRs were performed. 20µl PCR reactions were set up utilizing 0.5 units DreamTaq polymerase, 200µM dNTPs, 300nM of the forward and reverse primers ExBC_screen2_fwd and ExBC_screen2_rev (Table 4) in 1x DreamTaq Green buffer. Additionally 1.5 mM magnesium chloride were added to enhance lysis of bacteria. Colonies were picked from agar plates using 10 µl pipette tips and added to the reaction by stirring the tip in the reaction and pipetting up and down five times. Reactions containing 400 pg of the pCDH-EF1α-H2Kk-T2A-NGFR vector either with or without a barcode insert were used as positive and negative controls for later size comparisons of amplicon bands.

The reactions were cycled in a thermocyler with a prolonged initial denaturation of 3 minutes at 95°C to increase lysis of bacteria. Subsequently, 35 cycles with 94°C for 30 s, 53°C for 30 sec and 72°C for 1 minute were carried out, followed by a final elongation at 72°C for 1 minute.

5 µl of the samples were loaded onto a 2.5% agarose gel and electrophoresis was carried out for 50 minutes at 90V.

3.2.13 Colony PCR to screen for the presence of the high-complexity DNA barcode

As for the expressed barcode cloning the fraction of plasmid not carrying the barcode insert were determined by colony PCRs. 20µl PCR reactions were set up utilizing 0.5 units DreamTaq polymerase, 200 µM dNTPs, 300 nM of the forward and reverse primers DNABC_NT_screen_fwd and DNABC_NT_screen_rev in 1x DreamTaq Green buffer. Additionally 1.5mM magnesium chloride were added to enhance lysis of bacteria. Colonies were picked from agar plates using 10µl pipette tips and added to the reaction by stirring the tip in the reaction and pipetting up and down five times. Reactions containing 200 pg pCDH-EF1a-mtagBFP without a barcode insert were used as controls for later size comparisons of amplicon bands.

Reaction were incubated in a thermocycler with an initial denaturation of 3 minutes at 95 °C and 30 cycles with 94 °C for 30 sec, 53 °C for 30 sec and 72 °C for 1 minute, followed by a final elongation at 72 °C for 1 minute. 10 µl of the reactions were loaded on a 1.5% agarose gel and bands were separated for 90 minutes at 90V.

3.2.14 Amplicon library preparation for high-complexity DNA barcodes

In order to determine barcode frequencies as accurate as possible, a PCR strategy including the use of UMIs was utilized for the amplification of barcodes from plasmid or genomic DNA of barcoded samples. Therefore, the SiMSen-Seq [*97*] workflow was adapted for the purpose of amplifying barcodes from genomic DNA, even when very few barcodes are present within the reaction.

For every sample, SiMSen-Seq libraries were prepared in technical triplicates, unless stated otherwise. Generally 5 µl of isolated genomic DNA were used as template for the first PCR, including 0.1 units Phusion HotStart II Polymerase, 200 µM dNTPs and 40 nM of both primers, HP-DNABC_amp_fwd and HP-DNABC_amp_rev2, in 1x Phusion HF buffer in a volume of 10 µl. Reactions are incubated in a thermocycler at 98 °C for 30 sec, followed by 3-6 cycles with 98 °C for 10 sec, 62 °C for 6 min, 72 °C for 45 sec. The number of cycles necessary for sufficient amplification was adjusted based on the on-target input amount within the reactions. The reactions are inactivated by addition of 20 µl Actinase E in 1x TE buffer (45 ng/µl). For efficient protease treatment samples are further incubated at 65°C for 15 min, followed by heat-inactivation at 95 °C for 15 min and subsequent storage at 4 °C.

For each reaction of the first PCR two downstream PCR reactions were set up in order to maximize the utilized amount of products. Each second PCR consisted of 200 µM dNTPS, 400 nM of *HP_N7 adapter* and *TruSeq P5 xGen adapter* primers and 10µl of the previous PCR reaction in 1x Q5 HotStart buffer in a final volume of 40µl. Samples were incubated in a thermocycler at 98 °C for 3 minutes for initial denaturation and 3 – 6 cycles consisting of denaturation at 98 °C for 10 sec and annealing and elongation at 80 °C for 1 sex, 72 °C for 30 sec, 76 °C for 30 sec. Ramping rates between annealing and elongation temperatures were set to 0.2 °C/sec to resolve hairpin-structures and allow for efficient amplification. After amplification the samples were purified using homemade SPRI beads with a ratio of 0.8:1. PCR reactions derived from the same technical replicate were pooled at the elution step using 20µl buffer EB.

DNA-Barcode sequencing libraries were sequenced at LAFUGA with 150 cycles single-end and dual-index reads utilizing an Illumina HiSeq 1500.

3.2.15 Amplicon library preparation for expressed barcodes

Similarly as for high-complexity DNA barcodes, expressed barcodes were amplified using an adapted SiMSen-Seq protocol as described in 3.2.14 with the following alterations. For the first PCR primers

ExBC_AmpSeq_HP_fwd and *ExBC_AmpSeq_HP_rev* were utilized with an annealing temperature of 58 °C for 3 – 9 cycles. The number of cycles was dependent on the gDNA input amount to ensure sufficient amplification for all samples.

Expressed barcodes libraries were sequenced using a 100-nt single-end dual-indexed layout.

3.2.16 Analysis of sequencing data from high-complexity DNA barcode libraries

Sequencing data was demultiplexed using deML [*87*], allowing for no mismatch within the index reads. Barcodes were extracted from the demultiplexed samples using *bartender_extractor* [*98*], by defining the barcode pattern *('-p GATGG[4]ACT[2]CGA[2]CTT[2]CGA[2]CTT[2]GGA[2]CTA[2]ACT[2]CGA[3]CCACA'*). UMIs were extracted as the first 12 sequenced nucleotides within the reads ('-u 0,12') and barcodes had to match the non-variable nucleotides without mismatches ('*-m 0*') with an average Q-Score equal to or greater than 30 ('*-q ?*'). Clustering of barcodes was performed using bartender_single [*98*], binning barcodes within a hamming distance of 4 or 5 ('*-d 4*' or '*-d 5*'), depending on the number of barcodes expected within the sample, without utilizing a read-ratio threshold ('*-z -1*'). The resultant lists of barcode clusters and their respective UMI counts were further analyzed using R [*99*] and the packages tidyverse [*100*], ggplot2 [*101*], upsetR [*102*] and fishplot [*103*].

3.2.16.1 Analysis of barcoding data for determination of the bottleneck

To provide a conservative estimation of engrafted cell numbers a hamming distance of 5 was used for clustering of detected barcodes. Additionally, barcode clusters detected with only one UMI were discarded as they are likely representing technical noise. Furthermore, all barcodes that were only detected in one of the three technical replicates per sample were also discarded to further reduce potential false-positive barcode calls. As only one orientation of the DNA barcode was detectable using the established amplification setup, the resulting number of detected barcodes was multiplied by two in order to provide an estimate for the total number of engrafted cells per mouse.

Gini indices providing a quantification of the inequality of detected barcode frequencies were calculated using the ineq R package.

3.2.16.2 Analysis of barcoding data for the in vivo treatment experiment

Barcodes were extracted from raw sequencing data as described in 3.2.15. In order to exclude cross-contaminations between samples, detected UMIs per barcode within each technical replicate were filtered for detection with at least two UMIs before clustering. The resulting list of barcodes was used for clustering as described above (3.2.15) utilizing a hamming distance threshold of 4. Relative frequencies of barcodes per sample were calculated based on the average fraction of UMIs detected within each replicate.

For illustration of barcode data within the pseudo-fishplot the fishplot package for R [*103*] was utilized. For each experimental group, frequencies of detected barcodes were averaged across all mice. Additionally, all barcodes present with less than 0.005% frequency were excluded from the initial input sample. Sporadic barcodes at very low frequencies that were not detected in intermediate stages (start and/or control groups) but at the final stage (therapy group) were manually set to fractions of 0.0001% to allow for visualization. Additionally, barcodes were ordered based on their frequency within the initial input sample before plotting.

DEBRA [*104*] was used to determine statistically significant differences of barcode frequencies between groups. As use of technical replicates is not directly supported, UMI counts of clustered barcodes per sample were averaged across technical triplicates and rounded to integers. Differential barcode representation analysis was carried out using the implemented DESeq2 Wald-test method ('*method = "DESeq2(Wald)"*') comparing either mice of the start group to mice of the control group or mice of the control group to mice of the therapy group. For plotting of significantly differentially represented barcodes a false-discovery rate threshold of 0.05 was used.

3.2.17 Analysis of sequencing data from expressed barcode libraries

Sequencing data was demultiplexed using deML [*87*], allowing for no mismatch within the index reads. Barcodes and UMIs were extracted from raw fastq files using the bartender [*98*] extractor, by utlizing the barcode pattern for the short barcode ('*-p AAGGC[4]TT[3]AA[3]TTAGC*') and defining the first 12 nucleotides of the reads as UMI ('*-u 0,12*'). Barcodes had to match all fixed bases within the barcode pattern ('*-m 0*') and needed an average Q-Score of 30 ('*-q ?*').

Due to the high number of barcodes present within the plasmid pool and Nalm6 control samples, no barcode clustering was carried out in order to avoid overclustering, whereby real barcode would end up in one cluster.

3.2.17.1 Analysis of expressed barcodes from barcoded Pdx samples

Barcodes were extracted from raw sequencing data and clustered using bartender as described above. Additionally, barcodes were clustered using starcode [*105*] (version 1.1) with different hamming distance thresholds of 1 ('*--dist 1*') up to 3 ('*--dist 3*') and a fixed ratio-threshold of 5 ('*-r 5*'). Multiple cases in which barcodes present at very high frequency seemingly contaminated one or more other samples at very low frequencies were observed. Subsequently, barcode cross-contaminations were identified based on the normalized UMI counts of the respective barcodes within all samples and discarded in all samples in which the normalized UMI counts were less than 20% compared to the highest count observed in all samples. All data were manually analyzed for cluster compositions to exclude possible binning of true barcodes.

3.2.18 Transduction of Pdx cells using the lentiviral barcode pools

Production of lentiviral particles, as well as lentiviral transduction of Pdx and related work with NSG mice cells were carried out by Christina Zeller (Helmholtz Zentrum München) in the laboratory of Prof. Dr. Irmela Jeremias. All workflows described hereafter are derived from Christina Zeller's doctoral thesis [*106*] in order to provide a comprehensive overview of the methods used to apply the genetic barcoding system generated and used within this work.

3.2.18.1 Production of lentiviral particles

Lentiviral particles were generated transfecting HEK-293T cells at 50 – 80 % confluency using pMD2.G pMD2.G (1.25 µg/ml final concentration), pMDLg/pRRE (5 µg/ml final concentration), pRSV-Rev (2.5 µg/ml final concentration) and transfer vector (250 ng/ml final concentration). Plasmid DNA was mixed in DMEM with 2.4% turbofect and incubated for 20 min at RT. The DNA-turbofect mix was added dropwise to the HEK-293T cell after changing their medium. After three days the supernatant was withdrawn, centrifuged (400 xg, 5 min, RT) and filtered (0.45 µm). The virus was concentrated by

ultrafiltrating the supernatant using an Amicon-Ultra 15 ml centrifugal filter unit and centrifugation (2,000 xg, 30-40 min, RT). Concentrated virus was used directly for determination of virus titer or lentiviral transduction. Alternatively, virus was frozen as aliquots at −80◦C.

3.2.18.2 Lentiviral transduction

Between 2 x 10^6 and 10^7 PDX AML cells in 1 ml of the appropriate medium were incubated with third generation lentivirus(es) together with 8 µg/ml polybrene. After one day cells were washed three times with PBS (400 xg, 5 min, RT) and either resuspended in phosphate buffered saline (PBS) for the injection into mice or cultured for 4 - 6 days in PDX AML cell medium for subsequent fluorescence-activated cell sorting (FACS) enrichment.

3.2.18.3 Engraftment and expansion of primary patients' and Pdx cells

To engraft leukemic cells from AML patients, up to 10^7 peripheral blood (pB) or bone marrow (BM) cells in 100 µl sterile filtered PBS were injected intravenously into 6 - 15 weeks old NSG mice. For expansion freshly isolated or thawed PDX AML cells were injected. After transplantation of cells Baytril (2.5%) was added to the drinking water of animals for 7 days to prevent infections.

Engraftment was monitored every 2 - 3 weeks by flow cytometry measurement of human leukemic cells in murine pB or bioluminescence in vivo imaging. Mice were sacrificed (i) at defined time points, (ii) at signs of advanced leukemia (more than 50% leukemic cells within murine pB), or (iii) at first clinical signs of disease (rough fur, hunchback, and/or reduced motility). If leukemia became not apparent, mice were killed 52 weeks after transplantation by latest.

3.2.18.4 Isolation of Pdx cells from the murine bone marrow

To isolate PDX AML cells from the murine BM, femur, tibiae, hips, spine and sternum were extracted and crushed using mortar and pestle. Cells were resuspended in PBS, filtered (70 µm cells strainer) and washed in PBS (400 xg, 5 min, RT). Cells were re-suspended in PBS or the required buffer, stained 1:10 with trypan blue and 10 µl were used for determination of cell numbers using a Neubauer counting chamber.

3.2.18.5 In vivo treatment of mice engrafted with Pdx AML cells

To assess in vivo response of PDX AML samples to treatment, NSG mice were transplanted with samples expressing eFFly luciferase treated systemically with 50 mg/kg cytarabine (Ara-C), dissolved in 50% Sodium-(S)-lactate solution, intraperitoneally 4 days per week (day 2 - 5) for 3 consecutive weeks starting at a leukemic burden of a total flux of $2.1 * 10^9 - 1.1 * 10^{10}$ photons/sec. Animal were monitored every one to two weeks using BLI and sacrificed 3 days after stop of therapy.

3.2.18.6 Limiting Dilution Transplantation Assay

To determine the stem cell, or leukemia initiating cell (LIC), frequency in PDX AML samples, limiting dilution transplantation assays (LDTAs) were performed. Here fresh or frozen cells were counted with trypan blue and suspended in PBS. Cells were diluted and injected into groups of mice. Engraftment and disease progression was monitored by flow cytometric analysis of murine peripheral blood or BLI. Mice were sacrificed and PDX AML cells isolated from the BM or spleen of engrafted animals. Stem cell frequencies were determined by Poisson distribution using the ELDA software [*107*] (http://bioinf.wehi.edu.au/software/elda/).

3.2.18.7 FACS staining

To analyze expression of huCD33 or transgenes such as H2Kk or NGFR in FACS, cell lines or PDX AML cells, fresh or thawed, were stained with an appropriate antibody. 5 x 10^5 cells were pelleted (400 xg, 5 min, RT) and resuspended in 40 - 100 µl PBE buffer. 5 µl of huCD33-PE antibody, 2 µl H2Kk-APC antibody or 2 µl of CD271(NGFR)-FITC/PerCP-Cy5.5 antibody was added and incubated for 30 min at RT, 10 min at 4°C or 20 min at 4°C, respectively. Cells were washed with PBE (400 xg, 5 min, RT) and resuspended in an appropriate amount of PBE or PBS for FACS analysis sorting.

3.2.18.8 Enrichment of Pdx Cells by magnetic cell separation

To enrich human PDX AML cells from murine BM cells negative selection by magnetic cell separation (MACS) was performed using a cocktail of monoclonal antibodies against murine epitopes bound to magnetic beads. After isolation from murine BM or thawing, cells were suspended in 3 ml PBS + 0.5% BSA and incubated with 100 µ l - 400 µ l mouse cell depletion cocktail for 15 min at 4°C. 10 ml PBS + 0.5% BSA was added and the solution loaded to a LS column in a magnet prepared by rinsing with PBS + 0.5% BSA. After washing the column twice with PBS + 0.5% BSA, the flow-through was collected, centrifuged (400 xg, 5 min, RT) and resuspended in a required buffer or medium for further applications.

3.2.18.9 Enrichment of Pdx cells and cell lines by Fluorescence-Activated Cell Sorting

In order to enrich PDX AML cells or AL cell lines carrying one or more transgenes such as H2Kk, NGFR or a fluorochrome (mtagBFP, eGFP, mCherry and/or iRFP720), FACS was performed using a cell sorter BD FACS AriaIII (BD Bioscience, Heidelberg, Germany). When H2Kk or NGFR was sorted, cells were antibody stained. PDX AML cells suspended in PBS at a concentration of around 10^7 cells/ml were sorted, gating on leukocytes and subsequently on transgene carrying cells, into a FACS tube containing appropriate medium.

3.2.19 Preparation of illustrations

Figures 2, 4-9, 16, 18, 22, 25 and 31A were created with BioRender.com.

Figure 3 was adapted from the "Blood Cancers" template, from BioRender.com (2022). Retrieved from https://app.biorender.com/biorender-templates

4 Results

4.1 Establishment of cellular barcoding pools to enable tracking of individual cells within cell populations

The AML Pdx mouse model allows to engraft AML patient samples into immunocompromised mice while conserving their subclonal heterogeneity, thereby closer resembling the actual disease observed in patients compared to established cell lines. AML Pdx cells can be genetically modified via lentiviral transductions thereby making new methodological approaches accessible to patient derived samples.

In order to enable distinction of single cells within a population of leukemic cells, the cellular barcoding technique will be established for use within this Pdx model. Cellular barcodes are unique artificial DNA sequences, lentivirally integrated into the cells' genomes in order to easily distinguish cells and their offspring from one another. This approach enables to directly measure the heterogeneity of leukemic populations over time and under selective pressure, e.g. *in vivo* chemotherapeutic treatment, thereby enabling direct investigation of subclonal heterogeneity.

4.1.1 Design and cloning of barcode libraries and a UMI-tagged PCR read-out

To enable cellular barcoding of AML cells two different constructs were chosen to be prepared. For clonal tracking of bigger cell populations the first construct was planned to contain at least 5 million different barcodes in order to reduce the chance of barcode re-usage, i.e. two cells receiving the same barcode. A high-complexity barcode design derived from Thielecke et al. [*108*] was placed upstream of the marker's promoter within the lentiviral insert vector (Figure 8B). This design is composed of two variable nucleotide positions alternating with three fixed nucleotides, thereby preventing longer homopolymer stretches as well as extreme GC-contents, which both can interfere with efficient PCR amplification and sequencing.

The cloned barcode insert additionally comprised primer annealing sites next to the barcode sequence that were originally designed for amplification of a 1.5 kb fragment of the Phytochrom Interacting Factor 3 (PIF3) from Arabidopsis thaliana. These primer annealing sequences, corresponding to the At1g09530.1f349r20 primer pair of the AtRTPrimer database [*109*] were chosen to ensure specific amplification within PCR reactions containing mammal genomic background.

These constant sequences at the fragment ends enable correct positioning of the complementary single-stranded DNA-oligonucleotides when annealing the single-stranded oligonucleotides, which is necessary for the generation of sticky-ends.

Additionally, a second barcode construct, which contains a shorter barcode within the 3'-UTR of the lentiviral marker gene was designed and cloned. Due to its positioning the barcode is transcribed together with the marker gene, thereby enabling read-out within bulk and single-cell 3'-RNAseq libraries in addition to targeted PCR amplification from gDNA (Figure 8A). In order to avoid side-effects like accelerated degradation of mRNAs this barcode was kept as short as possible.

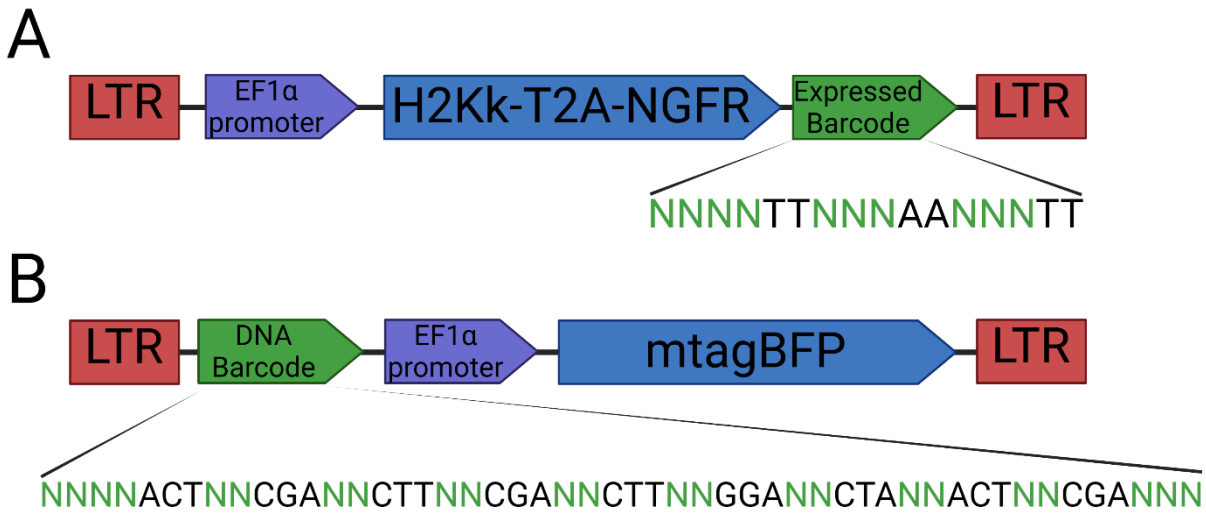


Figure 8: Schematic structure of lentiviral inserts and barcode sequences. (A) The lentiviral insert of the short, expressed barcode encodes for the two surface marker genes H2Kk and NGFR, which are linked by a T2A signal and under the control of EF1alpha promoter. The expressed barcode is placed within the 3'-UTR of the marker transcript, thereby potentially enabling direct read-out using 3'-(sc)RNAseq and consists of ten variable nucleotide positions (B) The lentiviral insert for the high complexity DNA barcode. The EF1alpha promoter controls expression of mtagBFP which used as marker to identify successfully barcoded cells. The high-complexity barcode is located upstream of the promoter and comprises 23 variable nucleotide positions.

For the insertion of the barcodes into the prepared vector backbones, a cycling restriction and ligation approach was chosen. This 'Cut-Ligation' comprised 55 cycles of restriction and ligation for 5 minutes each at 37°C and 20°C, respectively, followed by a final digestion step at 37°C for 60 minutes after addition of fresh enzymes. Compatible sticky-ends on insert and vector fragments, allow to selectively digest unwanted ligation products between two vector fragments or two insert fragments, while upon correct ligation of insert and vector fragments restriction sites are destroyed (Figure 9). Compared to classical restriction and ligation reactions carried out separately this setup maximizes the yield of correct ligation products while minimizing unwanted by-products like re-circularized vector-backbone

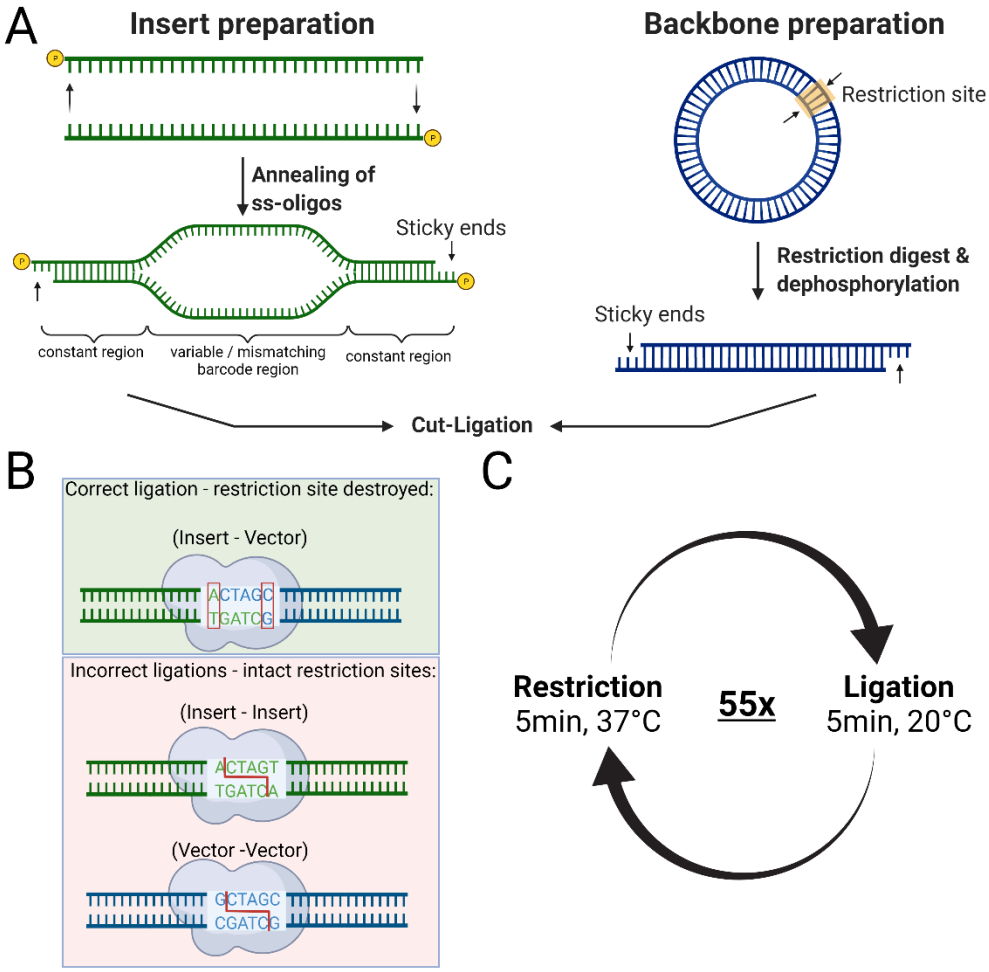


Figure 9: Overview of the cloning strategy for construction of barcode plasmid libraries. (A) The barcode insert is created by annealing of complementary oligonucleotides containing the degenerate barcode as well as constant regions on both ends that create sticky overhangs upon successful annealing. The vector backbone is digested using restriction enzymes that create single-stranded overhangs that are compatible to those created for the insert but differ in flanking nucleotides. (B) Correct ligation of insert and vector fragment destroys the restriction sites. Upon ligation of two vector or two insert fragments, intact restriction sites are produced which can subsequently be cut again. (C) Due to the use of compatible sticky-ends produced by restriction enzymes that differ between the insert and vector backbone, a cycling reaction of alternating restriction digests and ligations ('Cut-Ligation') can be used to maximize the efficiency of the cloning reaction.

In order to allow for efficient and precise read-out of barcode frequencies within barcoded cell populations via PCR, a molecularly-barcoded PCR strategy, SiMSen-seq [*97*], was established for both constructs. Here, a first PCR using primers that target the barcodes within the lentiviral inserts is performed with very few cycles. These primers additionally carry twelve degenerate bases, also called UMIs (Unique Molecular Identifiers), which uniquely tag every amplified barcode with a different unique sequence. To prevent mispriming, the UMIs are located within a hairpin loop that is formed by the primer at annealing temperatures and only opened at higher elongation temperatures, thereby increasing the specificity of the assay. A second round of PCR is subsequently used to generate sufficient amounts of amplicon and to introduce Illumina adapters necessary for sequencing. The use of UMIs allows to quantify the frequencies of distinct barcodes within the template more precisely, especially if only few barcodes are contained within a sample and thus many PCR cycles are needed to generate sufficient amounts of amplicon.

4.1.2 Quality control experiments for the validation of the barcode libraries' complexities

In order to assess the suitability of the created barcoding plasmid pools for use in real experiments different quality controls were performed. The most important property of the plasmids pools is their overall complexity given by the number of barcodes included in the pools, as it limits the number of cells that can be uniquely labelled and distinguished. In order to minimize the risk of labelling two distinct cells with the same barcode, the pools complexity should exceed the number of barcoded cells about 100-fold [*110*].

The cloned barcoding construct are 'undefined' libraries, as the exact number and sequences of barcodes within the final pools are unknown. To ensure unique labelling in experimental settings, the libraries' complexity was estimated using two approaches. First, the number of total colonies for each plasmid pool was estimated by counting the plated fractions of each transformation used to create the pool. All plasmid preparations were pooled based on the number of distinct barcodes expected to be contained within them in order to generate a plasmid library with evenly distributed barcodes. As a second readout the final barcode plasmid pools were used to create NGS libraries in order to directly measure their complexities.

After assuring sufficient complexity within the plasmid pools, lentiviral particle pools were created by Christina Zeller. Special care was taken to preserve a maximum complexity by scaling up production of lentiviral particles into many replicates produced in parallel. In order to assess the minimal complexity of these lentiviral pools, samples of Nalm6 cells were transduced in multiple replicates and harvested 2 days after transduction. gDNA was prepared from the resulting samples and used for the preparation of NGS libraries.

4.1.2.1 The DNA barcode plasmid pool contains about ten million unique barcodes

To estimate the overall complexity of the cloned DNA-Barcode plasmid pool, two complementary approaches were chosen. As a first estimate, the overall number of colony forming units achieved by the electroporations was estimated by counting the colonies on agar plates. In total, 16 electroporation reactions were carried out to transform the cloned plasmid into NEB10beta. For 12 of these reactions, 0.025% of the total volume were plated on agar plates in order to determine the overall number of transformants (Table 10).

Table 10: Estimation of the plasmid pool complexity by colony forming units (CFUs) obtained from transformation reactions. For 12 of 16 transformations 0.025% of the reactions were plated on agar plates in order to estimate the total number of transformants. To account for 1-2 potential cell divisions within the one hour of outgrowth before plating, total CFUs were divided by two or four. The resulting estimates represent the maximum and minimum number of total transformants expected to be present within the transformation reactions.

Transformation	Colony count	Total CFUs	Maximum total transformants	Minimum total transformants
Sample 1	520	2,080,000	1,040,000	520,000
Sample 2	227	908,000	454,000	227,000
Sample 3	1,404	5,616,000	2,808,000	1,404,000
Sample 4	2,155	8,620,000	4,310,000	2,155,000
Sample 5	1,022	4,088,000	2,044,000	1,022,000
Sample 6	2,120	8,480,000	4,240,000	2,120,000
Sample 7	686	2,744,000	1,372,000	686,000
Sample 8	814	3,256,000	1,628,000	814,000
Sample 9	400	1,600,000	800,000	400,000
Sample 10	1,771	7,084,000	3,542,000	1,771,000
Sample 11	1,578	6,312,000	3,156,000	1,578,000
Sample 12	1,425	5,700,000	2,850,000	1,425,000
Σ	**14,122**	**56,488,000**	**28,244,000**	**14,122,000**

Platings were carried out following a one hour incubation after the electroporation to allow for recovery of bacteria and expression of antibiotic resistance markers. As the generation time of *E. coli* is expected to be about 20 minutes in exponential growth phase [*111*] and the transformation procedure is expected to result in a prolonged lag-phase one to two cell divisions are expected within the initial outgrowth incubation. In order to estimate the number of different barcodes within the transformation reactions, the calculated number of total colony-forming units was divided by two or four to estimate the expected maximum and minimum number of total transformants.

Overall, more than 14,000 colonies were counted on the agar plates of the 12 tested transformation reactions and hence over 56 million colony forming units (CFUs) are expected to be contained within the respective transformations. When considering two cell divisions within the outgrowth phase, the minimum number of total transformants was estimated to be above 14 million for the 12 tested transformation reactions. Assuming the same mean transformation efficiency for all of the 16 transformations carried out, the total number of barcodes is expected to contain a minimum of 18.8 and a maximum of 37.7 million barcodes.

To ensure that most of the plasmids correctly carry the barcode insert, a screening PCR was performed using the *E.coli* colonies obtained from plating of the transformation reactions (Supplemental Figure 1). 98% of the screened colonies (89/91 successful PCRs) showed correct integration of the barcode insert into the vector, indicating that the fraction of plasmids without barcodes was successfully minimized by the cloning strategy.

In order to balance the frequencies of barcodes within the final plasmid pool, the prepared plasmid isolations were pooled based on their estimated complexity obtained by colony counts. Complexity of transformations for which no colony counts were obtained were assumed to be equal to the average complexity observed over all counted platings.

As a second independent read-out to assess the minimum number of barcodes contained within the final DNA-barcode plasmid pool, NGS libraries were prepared in technical triplicates using 1 ng of plasmid DNA as template and sequenced on an Illumina HiSeq 1500 with 48 – 65 million reads per replicate.

Notably, the barcode was cloned undirectionally, enabling the barcode insert to integrate in one of both directions into the vector. As one of the PCR primers for amplification of the barcodes needed to anneal outside of the insert in order to generate an amplicon of sufficient size, only half of the barcodes within the pool can be amplified and sequenced using the given library preparation workflows.

All barcode sequences detected within the sequencing data were binned by a hamming distance of four in order to reduce artefactual barcode calls arising from PCR and sequencing errors. The distance threshold was empirically determined by binning the detected barcodes by hamming distances of 1 to 6. (Figure 10).

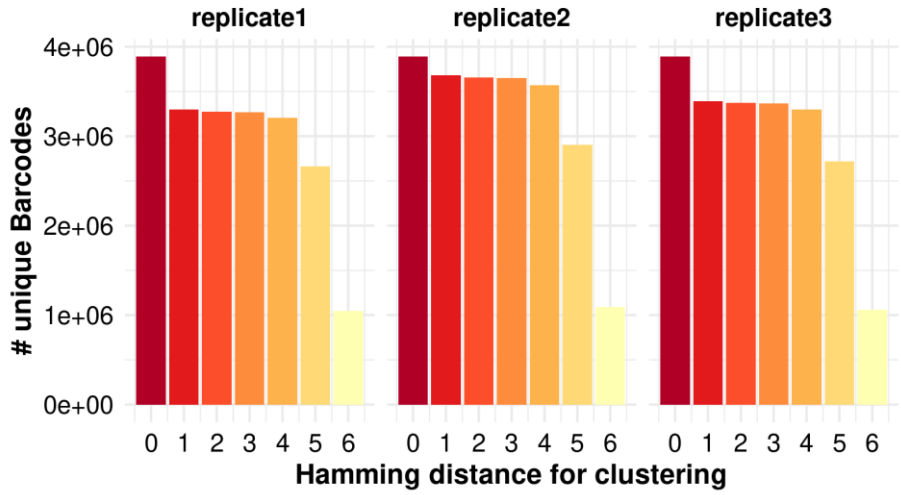


Figure 10: Number of barcodes detected within the technical triplicates of the sequenced high-complexity barcode plasmid pool after clustering of barcode sequences using different hamming distances. The number of barcodes is relatively stable between hamming distances 1 to 4, indicating that most amplicons only carried one sequencing error. With hamming distances of 5 or greater, barcode cluster counts significantly decrease due to real barcode sequences present in the library being binned together.

Most barcode artefacts were already removed by binning of barcodes using a hamming distance of one with only minor differences up to a distance of four. At hamming distances of five and six the number of total barcodes significantly dropped, suggesting that true barcodes not derived from sequencing and PCR errors are being binned together. Consequently, a hamming distance of four was used for clustering of barcodes within the sequencing data in order to minimize false-positive barcode calls.

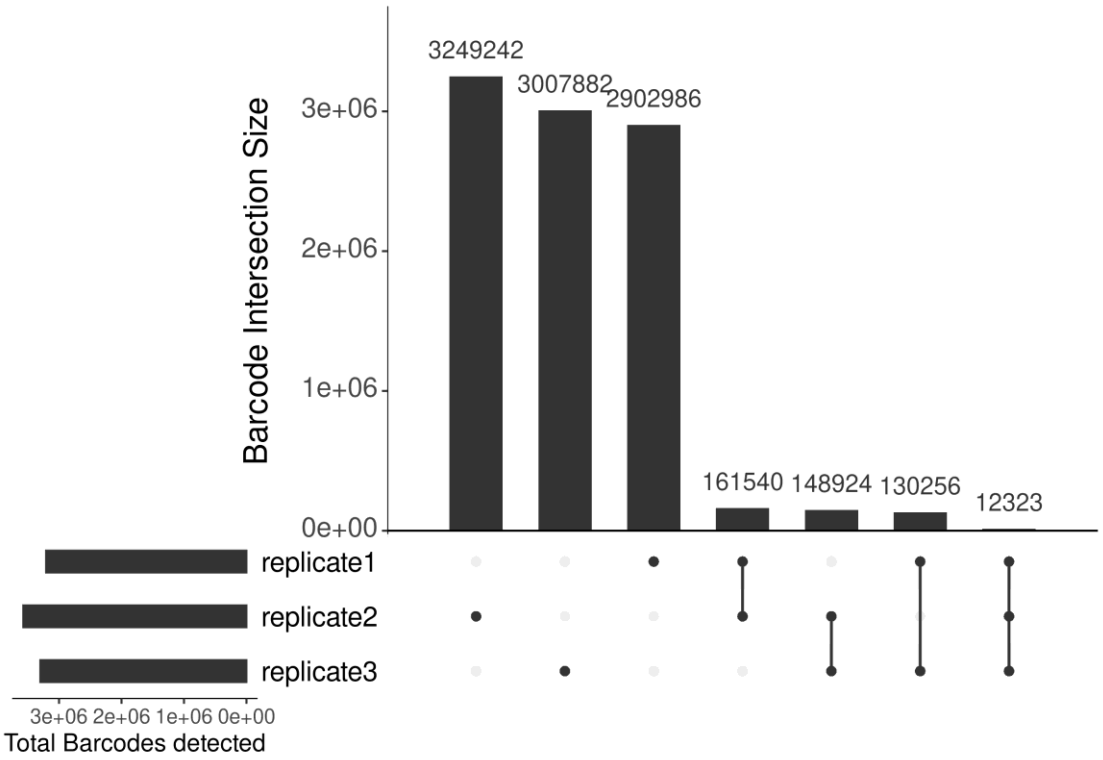


Figure 11: Number of unique and shared barcodes across technical replicates of the DNABC plasmid pool after binning of barcodes using a hamming distance of 4. About three million barcodes unique to each library replicate were detected. 130,000 – 160, 000 barcodes were shared between any two replicates.

After binning of barcode sequences using the established hamming distance threshold of 4, over 3 million barcode clusters were detected within each technical replicate, summing up to a total of over 9.6 million unique barcodes (Figure 11). Furthermore, for each technical replicate the fraction of barcodes detected within other replicates was consistently below 10% indicating that the plasmid pool has not been sequenced to saturation and is likely to contain even more barcodes.

Comparing the number of barcodes derived from plating of transformations and sequencing results of the plasmid pool shows that both estimates agree well. Estimations from colony counts suggested a complexity of about 19 to 39 million barcodes. As only half of the barcodes, having the correct orientation, can be read out using the established PCR workflow, 9.5 to 19.5 million barcodes were expected to be detectable. A total of 9.6 million unique barcodes detected via NGS agrees with this first estimation. As indicated by the relatively small overlaps between technical replicates, the plasmid pool was not sequenced to saturation and likely contains even more barcodes.

Taken together, the workflow for the generation of the barcode plasmid pool demonstrated very high efficiency, enabling the generation of about 10 million unique barcode plasmids.

4.1.2.2 The DNA barcode lentiviral pool exhibits sufficiently high complexity for use in AML Pdx experiments

A high complexity barcode plasmid pool containing about 10 million barcodes has been created. To apply these barcodes for labelling of cells, lentivirus particles need to be produced using this plasmid pool. The lentiviral pool can then be used to transduce cells and thereby label them with the barcodes. Production of lentiviral particles was carried out by Christina Zeller, whereby lentiviruses were produced in multiple parallel cultures in order to retain as much complexity within the barcodes as possible.

To estimate whether enough barcodes are left within the lentiviral barcode pool, Nalm6 cells were transduced in five replicates with the lentiviral high-complexity DNA barcode pool. Two days after transduction, cells were washed and used for preparation of gDNA. The resultant gDNA was used to create sequencing libraries in technical triplicates from all five transduction reactions.

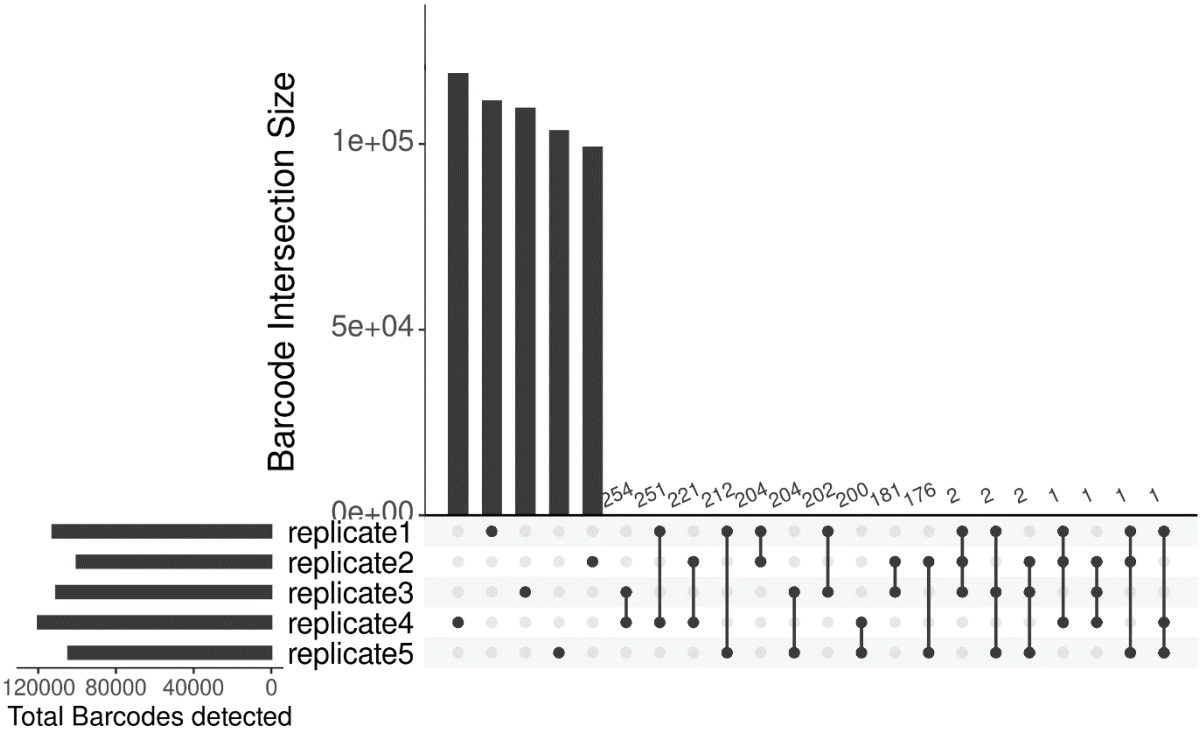


Figure 12: Number of detected barcodes unique to and shared between biological replicates of Nalm6 cells transduced with the lentiviral DNABC pool. More than 100,000 barcodes were detected within each biological replicate. Overlaps between the detected barcodes of two samples were generally very small, ranging from 176 to 251 barcodes. Only 1-2 barcodes were shared between any three replicates and none were shared between 4 or all replicates.

Library preparation from isolated gDNA of transduced cells was carried out in technical triplicates and worked as expected without detectable off-target amplifications. The sequencing results showed

100,000 to 120,000 barcodes being detected within each lentiviral transduction after binning of barcode sequences using a hamming distance of four (Figure 12). Only 176 to 254 barcodes were shared between any two samples. These minimal overlaps between the transduced replicates as well as the high number of over 100,000 barcodes being unique to every samples indicate that the virus production carried out by Christina Zeller could preserve the plasmid pools complexity.

As experiments involving the AML Pdx mouse model are limited to transplantation of a maximum of one to five million cells to prevent thrombosis and AML Pdx lines usually show LIC frequencies of about one in 300 to 5000 cells [*112*], the number of leukemic stem cells is expected to be well below 20,000 cells even in extreme experimental setups. Hence, the lentiviral barcode pool is sufficiently complex to minimize the risk of labelling two different cells with the same barcode and enables accurate quantification of cell numbers.

In summary, the pool of lentiviral particles created from the high-complexity DNA-Barcode plasmid-pool demonstrated sufficiently high complexity for accurate counting of barcoded cells within the proposed AML Pdx model system.

4.1.2.3 The expressed barcode plasmid pool contains well over 500,000 barcodes

As for the high-complexity barcode library, two independent measures – colony counts of transformation reactions as well as sequencing of the plasmid pool – were used to estimate the number of barcodes contained within the plasmid pool.

A total of 23 transformation reactions were used to transform the cloned plasmid pool into *E.coli*. In contrast to the high-complexity barcode pool, heat-shock transformations instead of electroporations were utilized. 1% from eight of the 23 transformation reactions were used for plating to estimate the number of total CFUs (Table 11). Over 900,000 CFUs were present within all eight analyzed transformation reactions. Considering one to two cell divisions within the one hour of outgrowth before plating of the cells, about 230,000 to 460,000 transformants are expected to be contained within these transformations. Assuming similar efficiencies for all transformation reactions, 670,000 to 1.3 million total transformants are expected to be present within the complete pool 23 transformations.

Table 11: Estimation of the complexity of the expressed barcode plasmid pool by colony forming units (CFUs). For 8 of 23 transformation reactions 1% of the total reaction volume was used for plating and subsequent colony counting in order to estimate the total number of transformants. To account for 1-2 potential cell divisions within the one hour of outgrowth before plating, total CFUs were divided by two or four in order to estimate the maximum and minimum number of total transformants expected to be present within the transformations.

Sample	Colony count	Total CFUs	Maximum total transformants	Minimum total transformants
Transformation 1	985	98,500	49,250	24,625
Transformation 2	572	57,200	28,600	14,300
Transformation 3	1,021	102,100	51,050	25,525
Transformation 4	1,295	129,500	64,750	32,375
Transformation 5	1,527	152,700	76,350	38,175
Transformation 6	1,608	160,800	80,400	40,200
Transformation 7	1,021	102,100	51,050	25,525
Transformation 8	1,295	129,500	64,750	32,375
Σ	9,324	932,400	466,200	233,100

Additionally, the fraction of plasmids not carrying a barcode insert was estimated by screening some of the counted colonies via colony PCR. This fraction should generally be smaller than 5% in order to allow for quantitative tracking of cells [*113*]. Here, only 2 of 62 (3%) successful PCRs indicated a missing barcode insert making the plasmid preparations suitable for precise barcoding experiments (Supplemental Figure 3). As the expected complexity was high enough and the fraction of plasmids not carrying barcodes was below 5%, all plasmid preparations were pooled in equimolar ratios to create the final plasmid pool for the low-complexity barcode.

For a second independent readout to estimate the overall number of barcodes contained within the expressed barcode plasmid pool, NGS libraries with 50 pg plasmid as template were prepared in five technical replicates.

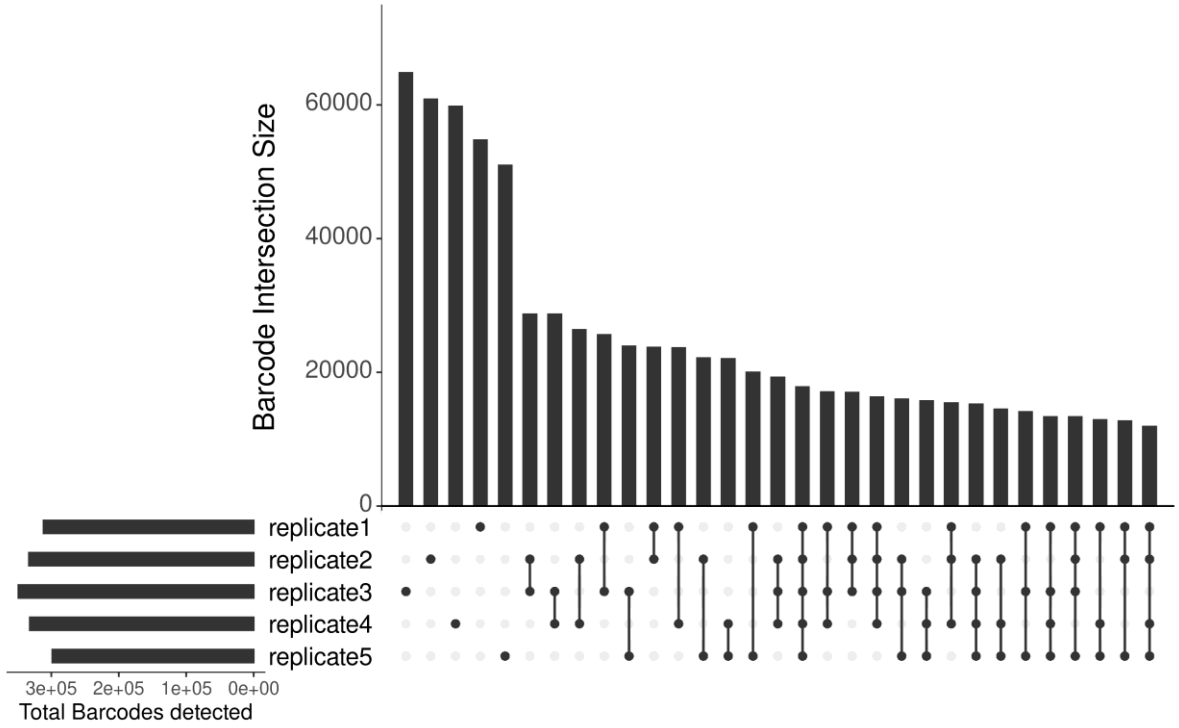


Figure 13: Number of detected barcodes unique to and shared between technical replicates of NGS libraries from the expressed barcode plasmid pool. About 300,000 to 350,000 barcodes were detected within library replicates with 50,000 to 60,000 barcodes being unique to each replicate. As the total number of detected barcodes is already close to the maximal theoretical complexity, the high fractions of shared barcodes between the samples are expected.

Every technical replicate showed 300,000 – 350,000 different barcodes with a total of 782,406 unique barcodes being detected across all five replicates (Figure 13). This is already relatively close to the theoretical maximum of 4^{10} or about 1.05 million barcodes. Therefore binning of barcodes by their hamming distance to correct for sequencing errors cannot be applied here, as the average distances across the detected barcode sequences are too small and many real barcodes would be binned together. However, the number of detected barcodes is in good agreement with the number of expected barcodes determined by CFUs of the transformation reaction (Table 11).

Additionally the expressed barcodes will only be used in small scale experiments, where a maximum of few hundreds barcoded cells are expected. Consequently, even if 50% of the barcodes detected within the plasmid pool are derived from technical errors, the pool's complexity would still be sufficiently high to enable unique labelling of small cell populations.

Although about 300,000 barcodes are detected in individual replicates, only 50,000 to 60,000 of these are not shared with any other replicate. These relatively big overlaps between the replicates are expected due to the limited maximal complexity of about one million barcodes and the high number of barcodes detected within each replicate, which already represent about one third of the maximum possible complexity of the barcode. Hence, the high fractions of barcodes shared between the biological

replicates are not derived from bottlenecks within the cloning procedures, but from the overall limited complexity of the utilized short expressed barcode sequence.

Taken together, the workflow for the creation of the expressed barcode plasmid pool yielded a high overall complexity that comprised over 70% of the maximum theoretic complexity.

4.1.2.4 The expressed barcode lentiviral pool exhibits sufficient complexity for use in small-scale AML Pdx experiments

Generation of the pool of lentiviral particles from the lentiviral vector poses another experimental bottleneck that may significantly reduce the total complexity of barcodes present. Therefore, as for the high-complexity barcode pool, generation of lentiviral particles was scaled up and carried out in many parallel reactions by Christina Zeller. The produced expressed barcode virus was used to transduce a population of Nalm6 cells in three biological replicates. Subsequently, cells were harvested and gDNA was prepared. For each of the three samples NGS libraries were prepared in technical triplicates.

Over 7,000 barcodes could be detected in samples 2 and 3, whereas only half were present within sample 1 (Figure 14). This decreased complexity in sample 1 indicates problems at the lentiviral transduction step. A lower multiplicity of infection likely resulted in fewer cells being barcoded and hence fewer barcodes being detected.

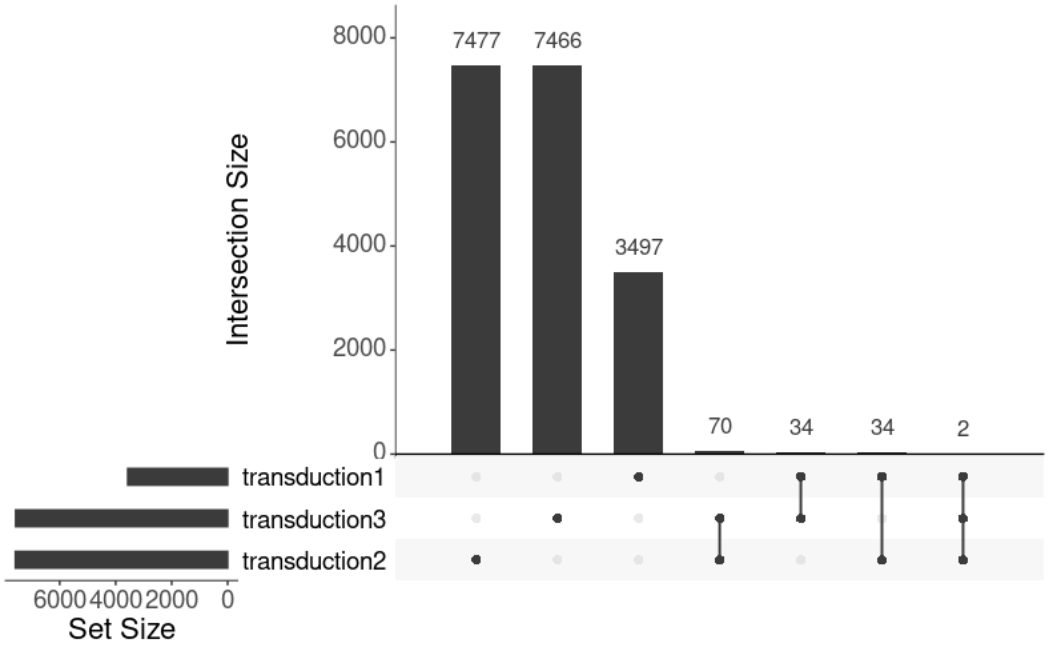


Figure 14: Number of detected barcodes unique to and shared between biological replicates of Nalm6 cells transduced with the lentiviral expressed barcode pool. The complexity of the lentiviral is sufficient for uniquely barcoding hundreds of cells. Minimal overlaps were detected between the different transductions. Significantly decreased barcode counts in 'transduction 1' might indicate problems at the lentiviral transduction step for this sample.

34 to 70 barcodes were shared among any two samples whereas two barcodes were detected in all three samples. However, these barcodes were prominent in one sample and were typically only detected in one technical replicate of the other samples. Therefore the observed overlaps may rather be caused by slight cross-contaminations that occurred during library preparations. Additionally, these overlaps are sufficiently small to enable unique labelling of a few hundred cells with no or only minimal barcode re-usage, thereby enabling highly accurate quantification of small barcoded populations.

Taken together, the final pool of lentiviral particles created from the expressed barcode plasmid pool showed sufficient complexity for the planned experimental setups in which not more than a few hundred barcoded cells are expected.

4.1.3 Barcoding enables quantification of the bottleneck in serial passaging of AML Pdx samples

An important consideration when using Pdx lines in animal models is whether the heterogeneity contained within the Pdx line is retained upon serial transplantation. If too few cells engraft upon re-transplantation, small subclones may be irreversibly lost in later passages. Therefore as a first use-case the new DNA barcode system was used to determine the number of cells engrafting when transplanting cell numbers routinely used in the AML Pdx model established at the Jeremias lab (Helmholtz Zentrum München).

To assess the number of leukemic cells engrafting and expanding within the utilized mouse model when transplanting cell numbers typically used for passaging of these Pdx lines, cells from the established AML491 Pdx line were transduced using the DNA-Barcode lentiviral pool. 550,000 cells were transplanted into each of six recipient mice. Mice were sacrificed after developing full-blown 57 days post injection. The gDNA of the re-isolated Pdx cells was used for preparation of barcode libraries in technical triplicates and sequenced on a HiSeq.

The obtained sequencing data indicated a strong skewing of frequencies among detected barcodes. 30 to 37% of the detected barcodes made up 99% of the barcoded cell populations. (Figure 15A, red lines) Although most of the technical noise from library preparations and sequencing should have been removed by clustering of barcodes based on their sequence similarity, it cannot be excluded that the barcodes showing very low frequencies are derived from technical artifacts [*108, 114*].

Therefore barcodes were additionally filtered for being detected in two of three technical sequencing library replicates, in order to provide a more conservative estimate for the number of cells engrafted per mouse.

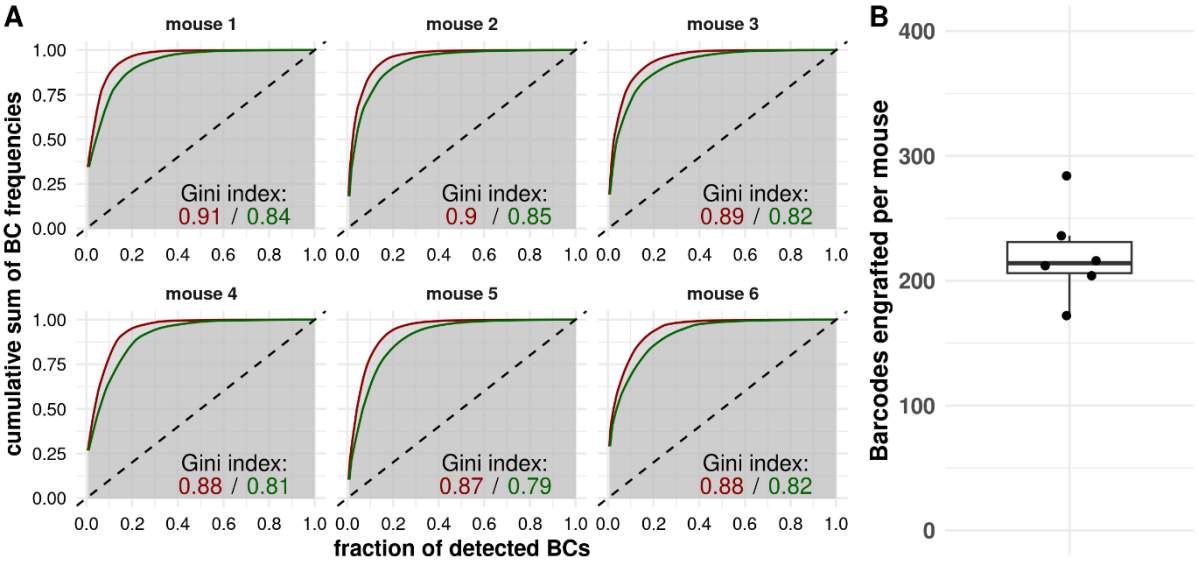


Figure 15: Skewing of barcode frequencies and number of engrafted barcodes for quantification of the passaging bottleneck. In the AML-491 Pdx line. (A) Cumulative fractions of barcodes sorted by decreasing abundance before (red) and after (green) filtering for barcodes present in at least two of three technical replicates per sample. Barcodes frequencies are significantly skewed, as indicated by Gini indices of 0.79 – 0.85 after filtering, suggesting subclonal heterogeneity in the utilized AML-491 Pdx line. (B) After application of additional filtering, a median of 214 barcodes is estimated to have engrafted per mouse.

By applying this additional filtering, 35 – 43% of the lowest frequency barcodes within each mouse were additionally discarded, resulting in 51 – 54% of barcodes making up of 99% of the barcoded cell population (Figure 15A, green lines). Nevertheless, the remaining barcodes still showed a high degree of skewing in their relative frequencies with Gini indices ranging from 0.79 to 0.85. Hence, this skewing likely reflects the sample's subclonal heterogeneity, with barcodes representing slowly growing subclones being significantly underrepresented after *in vivo* expansion, especially at the very late time point of full blown leukemia analyzed in this experiment.

The filtered barcode data was used to determine the number of cells engrafting in the recipient mice. Overall a median of 214 engrafted cells were present across analyzed mice (Figure 15B).

Hence, on average only one 0.04% of the 550,000 injected cells are still present at the final stage of full blown leukemia. This represents the overall bottleneck for serial transplantation and outgrowth to the final stage using this particular AML-491 Pdx line. Due to the additional filtering applied here, the determined bottleneck size is a conservative estimate and should rather be interpreted as a lower boundary. Assuming that all cells or subclones within the sample have the same likelihood to engraft, a specific clone would need to be present at 0.47% (i.e. a fraction of 1/213), to enable engraftment.

However, in order to estimate suitability of transplanted cell numbers for maintaining the Pdx line's heterogeneity, sampling effects need to be considered. According estimates based on the Poisson distribution, five cells of one subclone need to be sampled on average in order to reach a 99% probability to sample at least one cell of this subclone (Formula 1).

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$For\ \lambda = 5{:}\ P(X \geq 1) = 1 - P(X = 0) = 1 - \frac{5^0 e^{-5}}{0!} = 1 - e^{-5} = 1 - 0.0067 = 0.9933$$

Formula 1: Poisson likelihoods. In order to ensure that at least one cell of a specific clone is transplanted with 99% probability, the average number of cells sampled from this clone needs to be at about 5. Λ = mean of distribution, here: average number of cells sampled; k = number of cells sampled.

When assuming that all cells within this Pdx line have the same probability to engraft and proliferate equally fast, the tested transplantation setup allows to re-engraft subclones with as low as 2.3% frequency with over 99% probability. Consequently, no substantial loss of heterogeneity is expected to occur for this Pdx line.

In summary, cellular barcoding successfully allowed to comprehensively test the experimental scheme for serial passaging of the well-established AML-491 Pdx line. Using previously established techniques at the collaborators' laboratory this read-out would only have been possible indirectly, by serially passaging of cells across multiple generations in parallel to screening for the presence of known subclonal mutations via targeted next-generation sequencing.

4.1.4 Cellular barcoding allows to estimate LIC frequencies in leukemic Pdx samples with fewer mice than traditional limiting dilution transplantation assays

One important key characteristic of leukemic Pdx lines is the frequency of leukemic stem cells, also called leukemia initiating cells (LICs), present within the sample. These cells have self-renewal capacities theoretically allowing a single engrafting cell to cause a fully developing AML disease within recipient mice [*115*]. This key characteristic is usually determined by using limiting dilution transplantation assays (LDTAs) [*116-119*] in which different low cell numbers are each transplanted into multiple mice. Observations of un-/successful engraftments are then used to estimate the frequency of LICs.

As a proof-of-principle application for the low-complexity expressed barcode, an experimental setup to replace the classical limiting dilution transplantation assay for determination of LIC frequencies within Pdx lines was tested. Here, instead of the binary read-out of whether cells do or do not engraft at certain transplanted cell numbers the cellular barcodes enable quantification of how many different cells engrafted. Therefore, the total number of mice needed can be significantly reduced when using higher cell numbers for transplantations to ensure engraftment within every experimental mouse,

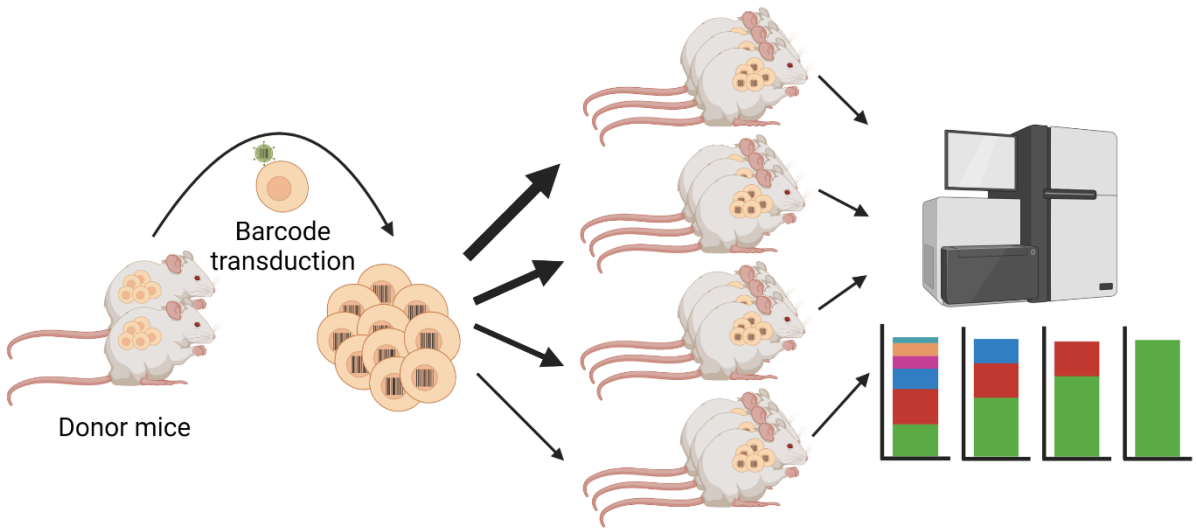


Figure 16: Experimental setup for estimation of LIC frequencies and isolation of single-cell derived subclones using the low-complexity barcode. Donor cells are transduced using the low-complexity barcode and different numbers of barcoded cells are transplanted into multiple recipient mice. Leukemic cells are re-isolated after outgrowth and a fraction of cells is used for barcode library preparation in order to determine the number of engrafted barcodes.

Hence, the lentiviral pool was used to barcode two Pdx samples, AML-491 and AML-661 in order to determine their LIC frequencies. The transduced Pdx cells were transplanted into multiple recipient mice at different cell numbers (Figure 16). Here, relatively low amounts of cells were chosen to enable determination of the LIC frequency using cellular barcodes as well as the routinely used read-out of un-/successful engraftment. Barcoded cells of successfully engrafted transplantations were re-isolated and used for preparation of barcode sequencing libraries in technical triplicates.

A total of 33 mice were utilized to determine the LIC frequency of the AML-491 line of which 26 mice showed successful engraftment (Table 12), whereas 16 out of 19 mice were positive for the AML-661 Pdx line.

In order to assess the LIC frequency using cellular barcodes, sequencing libraries were prepared in technical triplicates for 18 mice engrafted with AML-491 and 12 mice showing engraftment with the AML-661 line. The number of barcodes detected per sample were used for a linear regression through the origin in order to determine the samples' LIC frequencies (Figure 17), showing a good agreement between the number of cells injected and the number of barcodes detected within the sample.

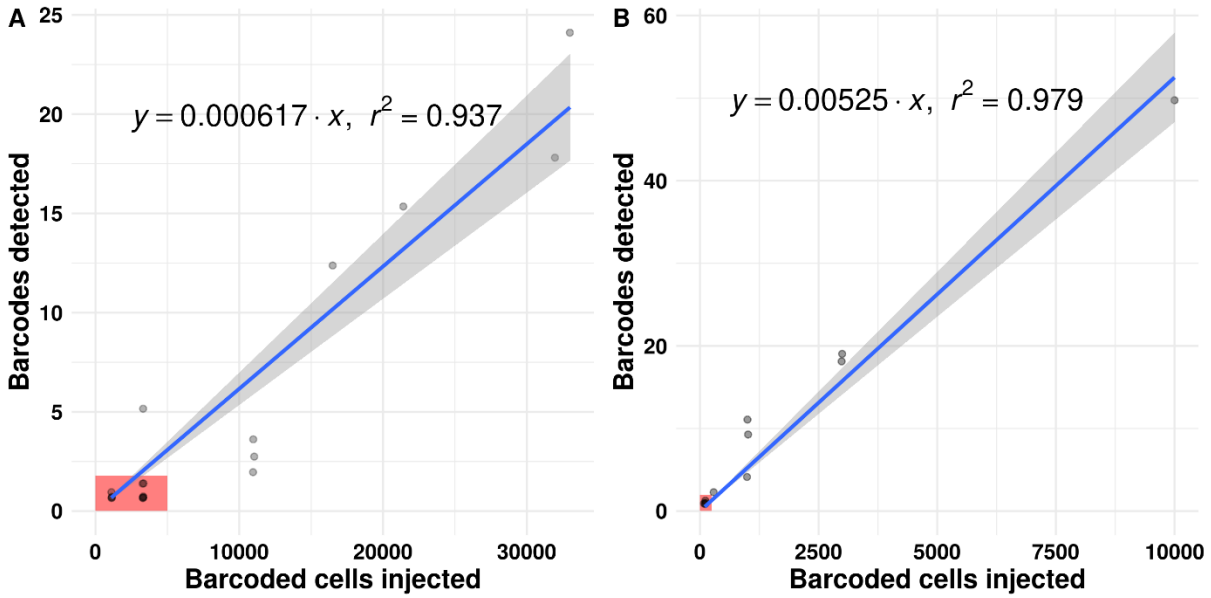


Figure 17: Relationship between injected cells and detected barcodes for the Pdx lines AML-491(A) and AML-661 (B). Linear regression through (0,0). Slopes of the regression lines represent the LIC frequency of the respective sample. Grey areas represent 95% confidence intervals. For better discrimination points were slightly jittered. Red boxes mark samples for which only one barcode was detected.

Linear regression indicates a LIC frequency of 1:1,620 for the AML-491 line (Figure 17A). Previously published data for the AML-491 line estimated a LIC frequency of 1:1,799 (95% CI: 1:945 – 1:3,426) [*112*]. Although samples within this study were derived from different passage numbers, the determined LIC frequencies agree very well.

The engraftment rates observed at the different cell doses of each Pdx line (Table 12) within this experiment were used by Christina Zeller (Helmholtz Zentrum München) to estimate the respective LIC frequencies using the ELDA software [*107*].

Table 12: Recipient mice of the two Pdx lines AML-491 and AML-661 transplanted with different cell numbers. Using the number of non-/engrafted mice at each cell number, LIC frequencies were determined by Christina Zeller using the ELDA software[107]

Pdx line	cells injected	engrafted / total mice	Mean LIC frequency (95% CI)
AML-491	33,000	1 / 3	1:5,810 (1:3,328 - 1:10,143)
	32,000	1 / 1	
	21,400	1 / 1	
	16,500	1 / 1	
	11,000	8 / 9	
	3,300	8 / 9	
	1,100	6 / 9	
AML-661	10,000	1 / 1	1:525 (1:237 - 1:1,161)
	3,000	4 / 5	
	1,000	5 / 5	
	300	2 / 4	
	100	4 / 4	

In contrast to the published data and the barcoding read-out from this experiment, the AML-491 line was estimated to have a LIC frequency of about 1:5810 (95% confidence interval: 1:3,328 – 1:10,143) by using the modelling approach based on observed engraftment rates.

This discrepancy is clearly caused by experimental outliers at the highest cell dose of 33,000 injected cells at which only one of three mice showed positive engraftment. As lower cell doses consistently showed significantly higher engraftment rates the observed drop-outs at the highest dose are clearly caused by additional technical noise, e.g. at injection of cells. When not considering the highest cell dose, the estimated LIC frequency increases more than 2-fold to a mean of 1:2,145 (95% CI: 1:1,182 – 1:3,895) and generally agrees with the quantification derived by the barcoding approach.

The LIC frequency within the AML-661 Pdx line is significantly higher, as it represents a sample of the same patient but derived from the second relapse, which was clinically more aggressive. The barcoding read-out for the AML-661 Pdx line determined a LIC frequency of 1:190 (Figure 17B), whereas the binary read-out of engraftment rates modeled an estimated LIC frequency of 1:525 (95% CI: 1:237 - 1:1,161). Although this discrepancy is smaller than previously observed for AML-491, the barcoding results again suggest a higher LIC frequency. A possible explanation is the reduced number of mice (n=19) utilized for the AML-661 line compared to the AML-491 line (n=33), which decreases the precision of the classical LIC frequency estimation by engraftment rates [*107*]. To allow for a more accurate estimation using this approach, additional mice receiving lower cell numbers would have been necessary in order to increase confidence of engraftment rates at low cell numbers.

However, as this experiment was designed to enable comparison of both readouts, engraftment rates and the number of barcoded cells within engrafted samples, a compromise had to be made in order to reduce the number of mice within the experiment. While transplantation of low cell numbers close to the samples' LIC frequencies are most informative when estimations rely on engraftment rates, the barcode approach profits from higher cell numbers due to the decreased technical noise, e.g. from serially diluting the cell population. Generally, the estimation of LIC frequencies by determining the number of engrafted barcoded cells should be rated as more precise due to the increased robustness of the quantitative data compared to the binary read-out of engraftment rates.

In summary, it could be shown that cellular barcoding allows to replace the classical limiting-dilution transplantation assay using a similar experimental setup. The barcode read-out enables acquisition of more detailed information per mouse, by providing a quantification of the number of cells engrafted. In combination with higher cell doses to minimize the fraction of mice showing no engraftment, cellular

barcoding can contribute to the reduction of experimental mice needed to assess LIC frequencies in Pdx models.

4.1.5 Cellular barcoding enables identification of isolates derived from a single leukemia initiating cell

Prior data from targeted sequencing indicated that the Pdx lines utilized for determination of LIC frequencies, AML-491 and AML-661, contain multiple genetically different subclones [*106, 112*] Furthermore, both Pdx lines are derived from the first and second relapse of the same patient and hence represent the same leukemia case at different stages of the disease. Therefore, samples from the barcoded LDTA experiment (4.1.4) that only had one barcode detected potentially represent cell populations of distinct subclones from different stages of the disease from a single individual. Consequently, these isoltes enable further phenotypic characterizations of these isolated subclones. Viable cells from all samples generated within the barcoded LDTA experiment were frozen. 13 samples that showed presence of a single barcode were further passaged in new recipient mice. Unfortunately, one of the samples failed to re-engraft within the mouse model. In order to verify that the remaining 12 isolates were indeed derived from single cells, all samples were analyzed for their barcode composition within later passages. No additional barcodes could be detected, indicating the successful isolation of single-cell derived subclones in eight cases for the AML-491 and four cases for the AML-661 Pdx line.

Christina Zeller (Helmholtz Zentrum München) further investigated the isolated subclones in order to analyze the intratumoral heterogeneity within this AML case. Based on known mutations in driver genes as well as exome sequencing, the isolated populations were identified to represent four genetically distinct subclones. For further experiments all samples were lentivirally marked with a unique combination of fluorphores, allowing to engraft multiple subclones within one mouse and analyze composition of mixed populations via flow-cytometry. In vivo experiments could identify significant differences in LIC frequencies as well as growth behavior and resistance towards chemotherapeutic treatment between subclones. Furthermore, information from phenotypic characterization of treatment resistance in combination with proteomics and transcriptomics data defined a score comprising 16 genes, which could be successfully used to predict outcome in data of a large independent AML patient cohort. This emphasizes that the characteristics observed within subclones of a single AML case can indeed be used to study general features of the disease.

Results of this study have been recently published as "*Adverse stem cell clones within a single patient's tumor predict clinical outcome in AML patients*" in the Journal of Hematology & Oncology [*120*].

4.1.6 *In vivo* treatment of barcoded AML Pdx samples

In many cases cellular genetic barcoding is used to analyze differential representations of cells within populations that underwent different experimental conditions, i.e. selective environmental pressures. In order to show, that the established high-complexity barcode system is capable of detecting such changes a new experimental setup was designed. As the established barcoding system will in the future mainly be used within patient-derived xenograft models of human AML cells this experiment relies on the well-established AML491 Pdx line.

The AML491 Pdx line is known to consist of multiple genetically different subclones and is sensitive towards *in vivo* therapy using Cytarabine (Ara-C) [*106, 112*]. Hence, it is well suited to analyze the effects of cancer treatment on the composition of a heterogeneous population of leukemic cells. In order to increase the chance of observing subclones that differentially respond towards chemotherapeutic treatment a low passage number of the AML491 Pdx line was chosen for barcoding.

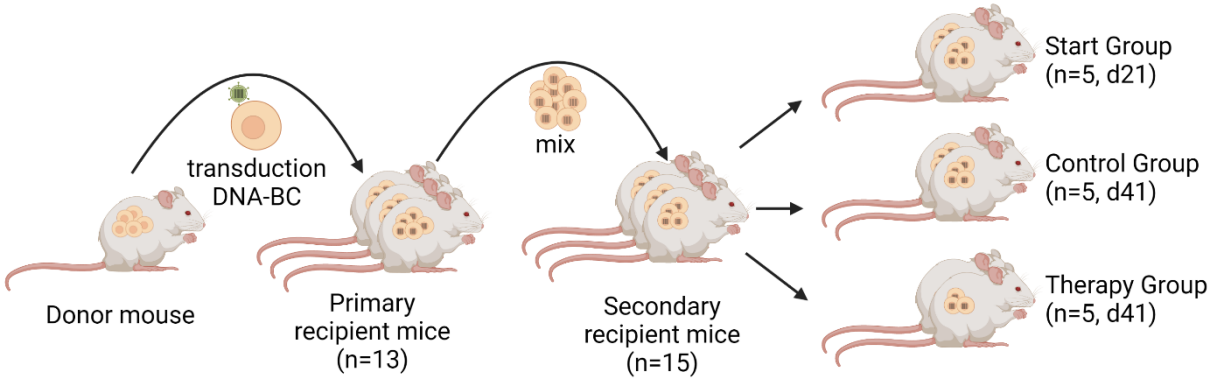


Figure 18: Experimental overview of the in vivo treatment experiment. Cells from an early passage of Pdx line AML491 were expanded within a donor mouse and transduced in vitro using the high-complexity barcoding virus. 500,000 cells were injected into each of 13 primary recipient mice in order to amplify the barcoded cells. After 47-49 days cells were harvested, subsequently mixed and 5 million cells were injected into each of 15 secondary recipient mice. The start group (n=5) was sacrificed and bone marrow cells isolated 21 days after transplantation of cells into secondary recipients. The remaining mice were treated for three weeks with four treatments with either PBS (Control Group) or 50 mg/kg Ara-C (Therapy Group).

To allow for a direct comparison between cells that receive chemotherapeutic treatment and the controls that grow without additional selective pressure within the mouse model, an additional round of cell amplification was included in the experiment (Figure 18). After lentiviral barcoding of the initial cell population, successfully transduced cells were transplanted into primary recipient mice to allow the barcoded cells to expand. This expansion enables the use of more biological replicates, i.e. experimental mice, in the subsequent experiment by increasing the number of available cells. Additionally, due to cell divisions every barcode will be represented by multiple cells, thereby potentially allowing the observation of the same barcodes within all secondary recipients and enabling direct comparison of barcoded cell populations within different mice.

Here, 82M cells were obtained from 13 primary recipient mice. Cells were mixed and 5 million cells were transplanted into each of 15 secondary recipient mice. Three groups of five secondary recipient mice were harvested at the time point of therapy start (d21, Start group) as well as after receiving three blocks of either Ara-C treatment (d41, Therapy group) or PBS as negative control (d42, Control group).

Due to severe weight loss one mouse of the therapy group had to be taken down before administration of the second therapeutic block. Data from this mouse were therefore excluded from analysis, reducing the number of biological replicates in the therapy group to four.

For each remaining mouse gDNA was isolated from bone marrow cells and barcode sequencing libraries were prepared in technical triplicates.

4.1.6.1 Amplification mice enable passaging of barcoded cells into multiple secondary recipient mice sharing the most common barcodes

The first important aspect of this experiment relates to the use of primary recipients to amplify barcoded cell populations in order to enable engraftment in multiple secondary recipients and thereby direct comparison of barcode frequencies between these mice.

To facilitate direct comparisons of cell populations between different conditions within an experiment barcodes need to be present in multiple mice in order to compare their relative frequencies and thereby the relative fitness of the underlying clonal cell population. Here, the initial barcoded cell population was further expanded in primary recipient mice in order to expand the cell population and allow for engraftment of the most common barcodes within all secondary recipients.

The generated barcode sequencing data indicated about 800 barcodes within the initial pool of cells after *in vivo* amplification and about 200 to 400 barcodes within each secondary recipient mouse (Figure 19, lower left panel). As expected, many barcodes that were only present at low frequencies within the initial pool of amplified barcoded cells (‘*Pooled input*’), did not engraft in any secondary recipient mouse (312 barcodes) or were exclusive to only one mouse (16-38 barcodes) (Figure 19, orange boxes).

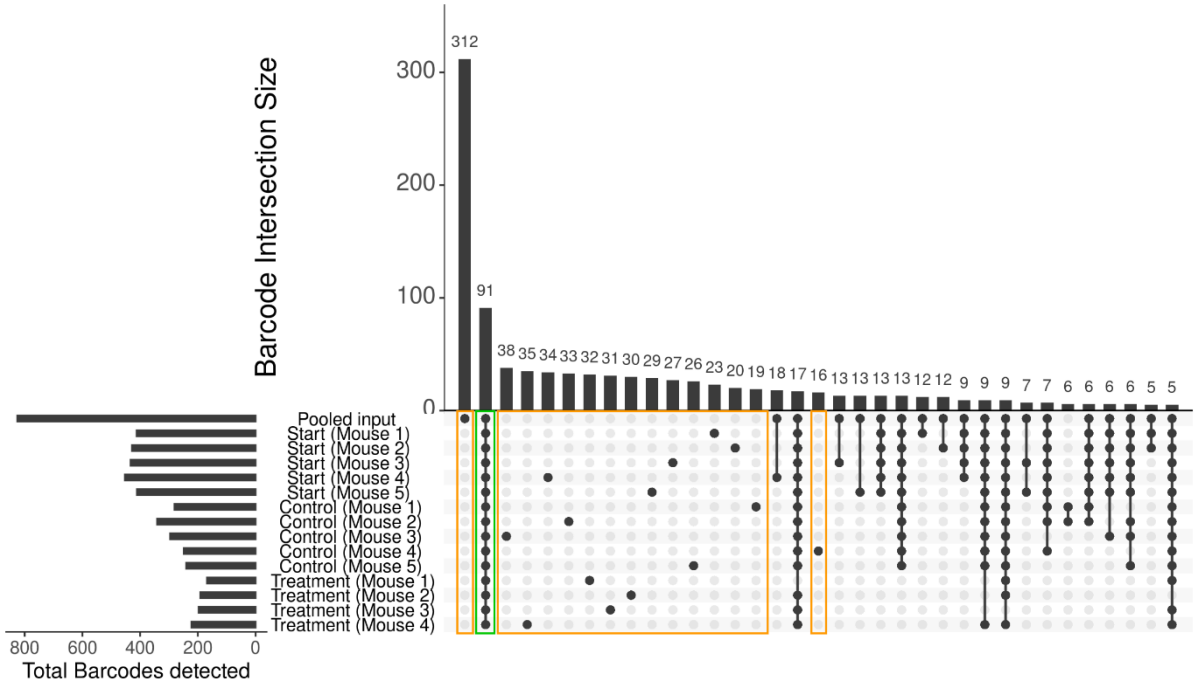


Figure 19: Number of detected barcodes within all biological samples. 312 barcodes were unique to the amplified cells from primary recipients (Pooled input) and 16-38 barcodes were unique to secondary recipient mice (orange boxes). 91 barcodes were detected within all samples (green box).

However, a subset of 91 barcodes was shared across all 14 secondary recipients (Figure 19, green box). Including barcodes only missing within one secondary recipient a total of 122 barcodes can be utilized to directly compare their frequencies among the experimental groups, which allows to also investigate minor subclones that may be present in the sample.

This emphasizes that the tested experimental setup is suitable to analyze barcoded cell populations in replicates within the Pdx model, thereby enabling robust and direct analysis of subclonal responses to stimuli like chemotherapeutic treatment.

4.1.6.2 Cellular barcoding allows to identify differential response of leukemic subclones towards in vivo chemotherapeutic treatment

During the course of the experiment all mice were monitored for their leukemic burden by bioluminescent imaging. Starting from the first treatment, a strong reduction in leukemic burden was observed in mice of the therapy-group compared (Figure 20A), as determined by *in vivo* bioluminescent imaging. In contrast, mice within the control group that did not receive chemotherapeutic treatment showed continuous expansion of the leukemic cells.

In agreement with the reduction in leukemic burden, a steady decrease in the number of detected barcodes within the samples from experimental start to end was visible (Figure 20B). 826 barcodes were detected within the pooled cells of primary recipient mice (‘*input*’). After transplantation of cells into secondary recipients, followed by an outgrowth period of 21 days about 430 barcodes, representing half of the initial complexity, were still detectable within mice of the start group (‘*start’*). Samples of the control group (‘*control’*), obtained 41 days after transplantation into secondary recipients, showed a further reduction in complexity with an average of 281 barcodes detected per mouse. Barcoded populations of the therapy group (‘*therapy*’) displayed a median of 197 barcodes per mouse, representing the overall lowest complexity seen within the experiment.

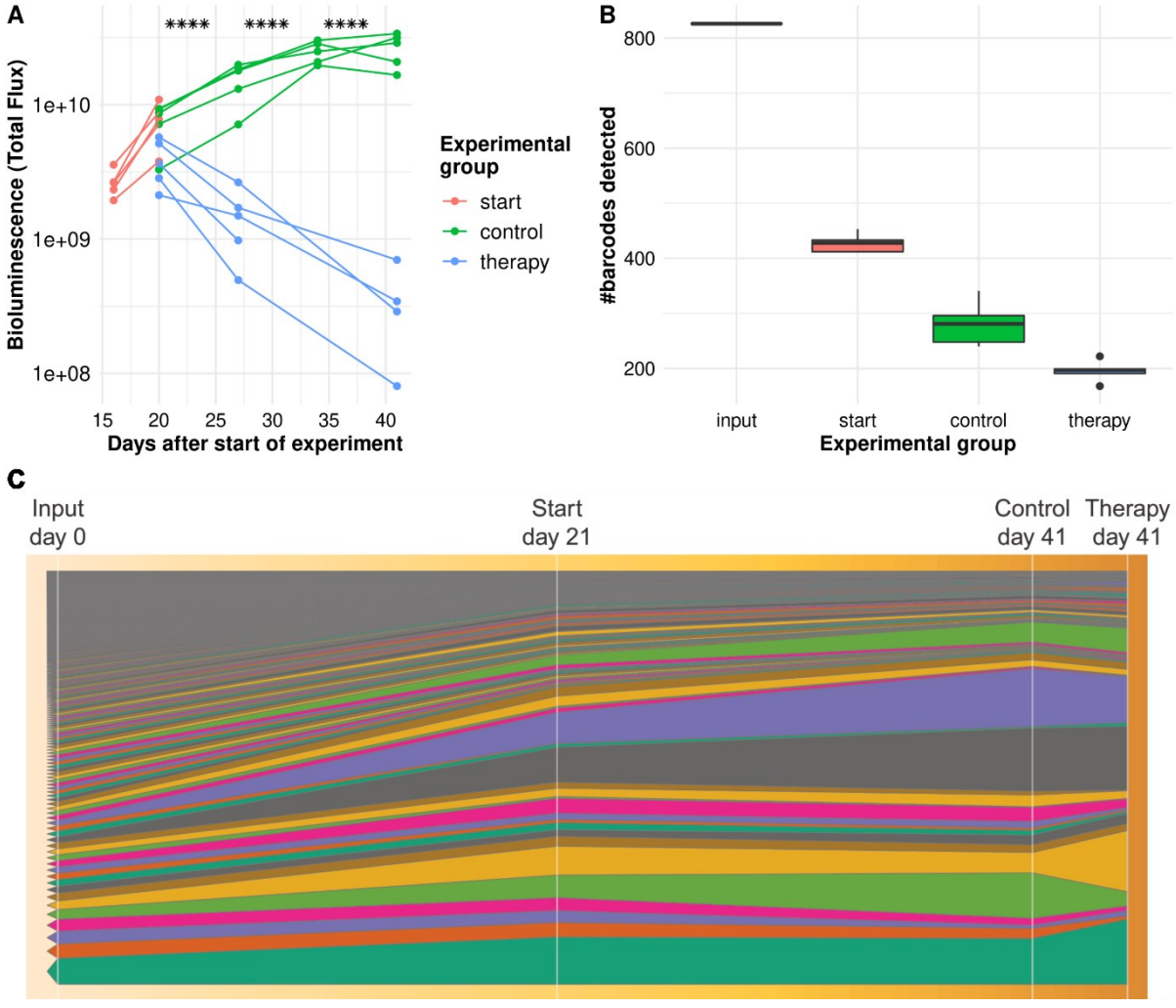


Figure 20: Tumor growth and complexity across experimental time points of the in vivo treatment experiment. (A) Data from bioluminescent imaging indicating the total leukemic burden within secondary recipient mice. Upon Ara-C in vivo treatment (indicated by asterisks) tumor burden is significantly decreased within the therapy group in contrast to the control group. (B) Number of barcodes detected within each biological sample, or mouse respectively. The number of detected barcodes step-wise decreases from the initial pool of barcoded cells (input), over the early time point (start group) to the control group and finally the treatment group. (C) Pseudo-Fishplot illustration of relative barcode frequencies over time. Each barcode is represented by one colored line, with its thickness representing the relative frequency. Barcodes are ordered by their initial frequency within the cell pool injected into secondary recipients. Relative barcode frequencies at each time point are averaged across biological replicates. Note that Control and Therapy groups were both sampled on the same day (41 dpi) and only separated for illustration purposes.

As all secondary recipient mice were transplanted using the same expanded population of barcoded cells, the most common barcodes are shared between them, enabling direct comparison of the composition of the leukemic cell population between the experimental groups (Figure 20C). Comparison of the amplified cell population used for transplantation into secondary recipient mice ('Input') to the data obtained from mice of the start group ('Start') shows that all barcodes initially present at high frequencies further expand until 21 days after injection. Hence, barcodes initially present at low frequencies are further attenuated, resulting in the observed reduction of overall detected barcodes. The same reduction in low frequency barcodes, though less drastically, can be observed when comparing the time points 21 ('Start') and 41 days ('Control') after injection of the cells. However, in contrast to the early growth phase not all barcodes present at high frequencies at day 21 display further expansion. While some barcoded cells continue to grow faster than the average, thereby making up higher fractions of the barcoded population (Figure 20C, barcodes in light green, gray and purple), other prominent barcodes show constant (Figure 20C, dark green) or slightly decreased relative frequencies (Figure 20C, yellow). At day 41 after injection mice already displayed full blown leukemia within the '*Start*' group, as indicated by the strong bioluminescent signal (Figure 20A). Hence, the slightly changed expansion pattern might be the result of elevated selectional pressure due to increased competition among leukemic cells. Finally, comparison of the '*Therapy*' group to the '*Control*' group, which did not receive chemotherapeutic treatment, indicates that some barcodes are differentially represented with some being strongly enriched or depleted within the '*Therapy*' group (Figure 20C, yellow & light green).

In order to verify the observed differences in the mean relative barcode frequencies between the experimental groups DEBRA (DESeq-based Barcode Representation Analysis) was used to test whether these changes are statistically significant when accounting for the variance of barcode frequencies observed within biological replicates of the same experimental groups.

Indeed, 39 barcodes were tested to be significantly different (FDR=0.05) between the '*Control*' and '*Therapy*' group (Figure 21). Thus, these barcodes represent at least two distinct subclones that display either increased resistance or sensitivity towards chemotherapeutic treatment compared to the average barcoded leukemic cell population and subsequently show significant enrichment or depletion when comparing barcode frequencies between these groups. These results demonstrate that the established barcoding construct allows for direct identification of chemotherapeutic resistance within AML Pdx lines.

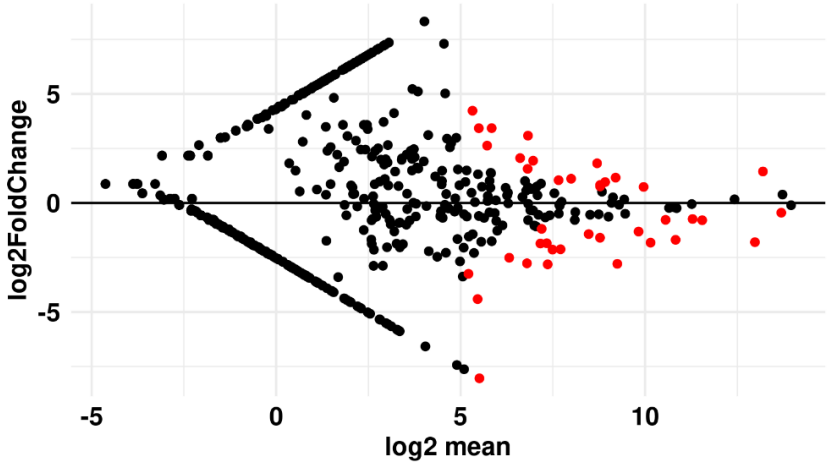


Figure 21: Differentially represented barcodes between mice of the Control and Therapy group. Frequencies of 39 barcodes showed statistically significant changes between conditions (highlighted in red, FDR 0.05. Barcodes enriched in the treatment group (positive log2-Fold-Changes, 17 BCs) represent at least one subclone that displays weaker response towards the chemotherapeutic treatment compared to the population mean. On the other hand barcodes significantly represented at lower frequencies in the Therapy group (22 BCs) are more sensitive towards treatment compared to the population average of barcoded leukemic cells. The observed differences prove the presence subclones that show increased resistance towards in vivo treatment in the AML-491 Pdx line.

Similarly, 73 barcodes were identified to be differentially represented (FDR 0.05) when comparing barcoded cell populations between the 'Start' and 'Control' groups (Supplemental Figure 5), indicating that subclones also show phenotypic differences in the absence of the selective pressure of *in vivo* chemotherapeutic treatment.

The observation of differences in the response towards chemotherapeutic treatment between barcoded cells, encouraged further characterizations of the subclones derived from this AML patient and led to isolation of single-cell derived subclones (as described in 4.1.5).

Taken together, these results proof that the established barcoding construct and experimental setup are capable to directly detect subclonal heterogeneity within AML Pdx lines by measuring differential responses of subclones to environmental factors like chemotherapeutic treatments.

Additionally, the established barcoding pools and workflows were also utilized in further projects at the Helmholtz Zentrum München to investigate other AML and ALL Pdx lines and will be continuously be used within further experiments in the future.

4.2 High sensitivity targeted sequencing for detection of subclonal mutations using single-molecule Molecular Inversion Probes (smMIPs)

For the tracing of subclonal evolution in samples of AML patients as well as elderly that may display clonal hematopoiesis a sensitive and cost-efficient targeted re-genotyping assay using single-molecule Molecular Inversion Probes (smMIPs) will be established. As a first application the assay will be used for a first cohort study in order to verify its suitability for sensitive re-genotyping within high-throughput studies.

4.2.1 Investigation of residual leukemia and clonal hematopoiesis of indeterminate potential in a cohort of AML patients in long-term remission

A cohort of 357 AML patients, who are in remission for more than five years after initial treatment with chemotherapy and/or hematopoietic stem cell transplantation (HSCT) was recruited by Dr. Klaus Metzeler (Klinikum der Universität München). The collected peripheral blood samples are being analyzed for the presence of residual leukemia and clonal hematopoiesis. Additionally, surveys were collected by Dr. Luise Hartmann (Klinikum der Universität München) in order to gather additional meta-data, thereby enabling e.g. to associate the presence of clonal hematopoiesis with the occurrence of cardiovascular diseases.

In order to compare the incidence of detected clonal hematopoiesis to individuals without prior hematopoietic diseases a control cohort of 154 age-matched individuals was included in the study. 77 of these samples are derived from hip surgeries, which were also utilized in another study aiming to elucidate the incidence of clonal hematopoiesis in elderly people using the commercial Haloplex panel for targeted sequencing. The remaining 82 blood samples are derived from the Covid-19 register of the LMU Hospital (“CORKUM”), which were obtained from patients infected with the Sars-CoV-2 virus.

To estimate the sensitivity of the newly established assay four gDNA dilution series were created using gDNA from two patients in order to artificially produce variants down to 0.7% VAF. The commercial Haloplex assay represents the current standard methodology for targeted sequencing of AML patient samples at the Klinikum der Universität München. Therefore variants calls from 36 samples that had been previously analyzed using this Haloplex panel were compared between both methods. As a last additional quality control, sequencing libraries for 16 patient samples were prepared twice from independent gDNA extractions in order to assess reproducibility of variant calls.

Downstream analyses involving the detected variants and incidence of clonal hematopoiesis within the AML patients in long-term remission and the control group are carried out by the group of Prof. Dr. Klaus Metzeler. Therefore data about detected variants are not further discussed within this work and are planned to be published in a separate research article.

4.2.2 Design of an enhanced cost-efficient smMIPs panel compatible with standard sequencing primers to target known AML and CHIP driver genes

One drawback of regular smMIP panels is the use of a backbone sequence that is not compatible with standard Illumina sequencing primers. This backbone provides a universal sequence to allow for PCR amplification of the captured sequences, thereby introducing sample indices and fragment ends needed for binding to the flow-cell when using Illumina sequencing. These sequences were designed to enable robust PCR amplification, but are not compatible with standard Illumina sequencing primers. Hence, Illumina sequencers need to utilize custom primers in order to sequence these libraries, which prohibits the sharing of flowcell lanes with unrelated sequencing libraries. As sequencing costs are higher when using smaller flowcells or sequencers, such as Illumina NextSeq 500, this increases costs for sequencing of larger library pools.

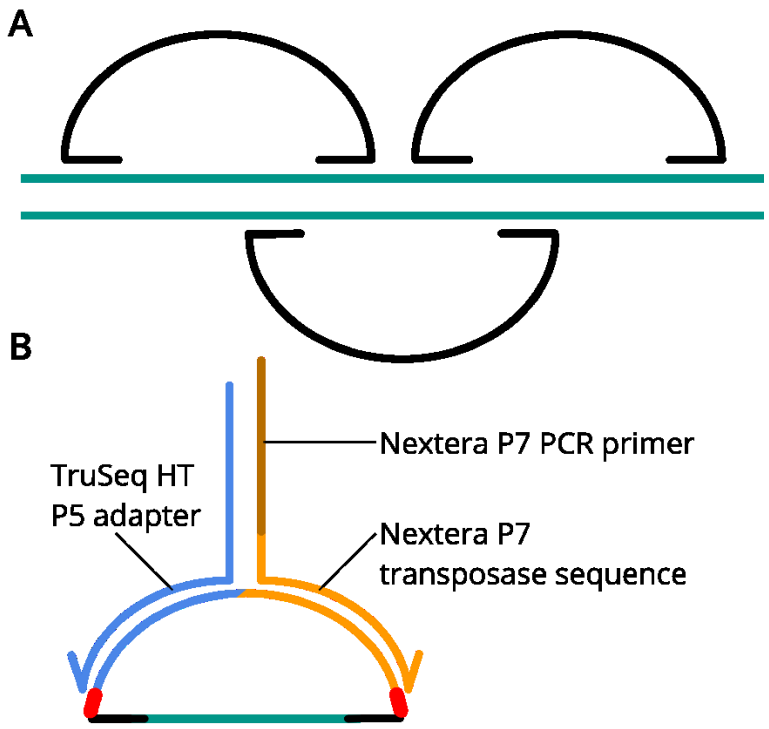


Figure 22: (A) Schematic of the tiling strategy used for selection of probes (black) within the smMIP panel. If suitable probes with high predicted performance were available, neighboring probes were chosen to hybridize to opposite strands of the gDNA template (green) in order to avoid sterical hindrance. (B) Schematic of the newly designed smMIP backbone. The standard backbone was replaced in order to facilitate sequencing without the neccessity for custom sequencing primers.

To circumvent this problem the backbone of the smMIPs was re-designed to incorporate sequences compatible to Illumina's standard sequencing primers while not altering the length of 30 nucleotides of the backbone (Figure 22B). Consequently, one half of the backbone consists of a sequence used to anneal a standard Illumina TruSeq HT P5 adapter. The other half represents the end of Illumina Nextera P7, normally introduced by a transposase. The mix of Nextera and TruSeq adapter sequences was chosen in order to avoid extensive interactions of primers during amplification.

TruSeq adapters are typically integrated into library fragments by ligation and tend to produce primer-dimers when both, P5 and P7, adapters are used in PCR reactions. Nextera adapters on the other hand are introduced in two steps into library fragments. The so called 'mosaic ends' are introduced by a transposase during the tagmentation step of library preparation and subsequently used as primer annealing sites for a downstream PCR which introduces the rest of the adapter. Here, the mosaic sequence 3'-ends of P5- and P7-sides are identical for the last 19 nucleotides [*121*], rendering them unusable as primer annealing sites in this case. Therefore, a combination of Nextera and TruSeq sequences was chosen to enable efficient amplification while keeping the original length of the backbone.

For the amplification of smMIPs after target capturing standard TruSeq P5 adapters can be utilized. These primers are unaltered from their original sequences and can be used in other library preparations, for example when preparing (sc)RNAseq libraries [*122, 123*]. The Nextera adapters, however, are specific for smMIP library preparations, as they need to include the mosaic-sequence which is not part of the standard adapter sequence, as it is usually introduced within a tagmentation reaction.

Ultimately, using this altered backbone sequencing of smMIP libraries can be carried out using standard sequencing primers, thereby enabling sharing of flowcell lanes with any other standard sequencing libraries on Illumina sequencers.

For the design of the smMIP panel, a single-tiling strategy (Figure 22A) was chosen for covering the genomic target areas in order to reduce the initial costs for synthesis of the oligonucleotide probes. Here, target sequences are generally only captured from one of both strands of gDNA template. Double-tiling, in contrast, would cover both strands of the DNA throughout the target regions. Hence, single-tiled panels reduce the initial costs by about 50%.

In order to obtain an optimized single-tiled smMIP panel, probes for target areas were first generated in a double-tiled approach and manually reduced to a single-tiled panel. This strategy in most cases allowed to choose neighboring probes to hybridize to different strands of the gDNA, thereby avoiding sterical hindrance and potentially allowing for higher hybridization efficiencies (Figure 22A). In cases where

known SNPs are located within the sequences targeted by the hybridization arms, two probes were ordered in order to allow for efficient hybridization in the presence of respective SNP on the template gDNA.

For the specific application of analyzing clonal hematopoiesis as well as residual leukemia a target list containing the most common driver genes was created by Dr. Klaus Metzeler (Klinikum der Universität München). This list comprises 24 genes known to contain recurrent driver mutations in either AML or clonal hematopoiesis (Table 13). Depending on contexts of known driver mutations within these genes, either whole exons or only short mutational hot spot regions were included in the targets list. The target areas were kept as short as possible in order maximize sequencing depths over the most recurrent driver genes.

*Table 13: Gene targets of the smMIP panel. The targets comprise about 16 kb in exons and hotspot regions of 24 genes known to be recurrently detected as driver mutations in either acute myeloid leukemia or clonal hematopoiesis. (ITD = Internal Tandem Duplication region, TKD = Tyrosine-Kinase Domain). *CEBPA was removed from the targets for the final probe pool due to overall poor performance of all probes within this region.*

Gene	Targeted Feature	Target length [bp]
ASXL1	Hotspot	817
BRAF	Hotspot	2
CALR	Exon 9	206
CBL	Hotspot	330
*CEBPA**	Exon 1	1,095
CSFR3	2 Hotspots	168
DNMT3A	Exons 10-26	2,408
FLT3	ITD & TKD region, N676 Hotspot	269
GNB1	Hotspot	2
IDH1	Hotspot	2
IDH2	HotSpot	96
JAK2	Hotspot	2
KIT	Exons 8, 17	274
KRAS	2 Hotspots	84
NPM1	Hotspot	15
NRAS	2 Hotspots	12
PPM1D	Exon 6	566
PTPN11	2 Hotspots	144
RUNX1	Exons 4B, 5-9	1,356
SF3B1	2 Hotspots	35
SRSF2	Hotspot	26
TET2	Exons 4A, 4C, 5-11	6,261
TP53	Exons 1-12	1,524
U2AF1	2 Hotspots	6
WT1	Exons 1, 3-9, 10B, 11	1,274
		16,974

MIPgen [*81*] was used to create a list of candidate probes targeting the defined genomic regions. This software includes an empirically derived model based on the performance of 12,000 probes to predict the performance of newly designed probes. Prediction is, among other features, based on the GC-content of the captured sequence and the probe's hybridization arms. Additionally, possible off-target hybridizations are estimated by determining the number of times the hybridization arms' sequences can be found within the target genome.

The candidate probes generated, representing a double-tiled panel, were manually reduced to an optimized single-tiling subset with minimized sterical hindrance between probes, as described above. The resultant smMIP panel comprised a total of 303 probes which were ordered as HPLC-purified oligonucleotides.

4.2.3 Balancing of the relative concentrations per probe within the smMIP pool improves uniformity of coverage across targets

A crucial factor for good performance of any smMIP panel is the balancing of capture efficiencies among the contained probes. If probes show strong differences in their efficiency to capture their target sequences, unique coverages and therefore the lower limit for detection of variants may greatly vary across the targets. In order to balance the performance across probes, the first hybridizations and subsequent library preparations were performed using genomic DNA derived from the GM18505 lymphocyte cell line [*82*] as a standardized template. As part of the 1000 Genomes Project [*124*] data from whole genome sequencing (WGS) and whole exome sequencing (WES), as well as SNP annotations derived thereof are available online.

After sequencing the probe performances were evaluated using MIPgen by looking at the UMIs detected for each smMIP. As each UMI corresponds to a unique capture event, the number of detected UMIs provides a direct measure for the hybridization efficiencies. Iteratively, based on the relative performance concentrations of probes with low capture efficiency were increased within the smMIP pool and library preparations were repeated.

Here, some probes showed either very low hybridization efficiencies or were not detected at all. Additionally, all probes targeting the *CEBPA* gene generally showed bad performance, that could not significantly improved by increasing the concentration of the respective probes within the panel. Interestingly, these probes also showed an increased tendency towards self-circularization, resulting in

library fragments not containing any sequences captured from the gDNA template. These library fragments subsequently produced sequencing reads not carrying any genotyping information and hence unnecessarily increase sequencing costs. The bad performance of all probes within this target area is most probably cause by its high GC content, which generally makes it hard to amplify via PCR [*125*]. Therefore all probes targeting *CEBPA* were completely removed from the probe pool, in order to improve the overall efficiency of the panel.

Furthermore, one probe used to target the NPM1 hotspot locus not only showed very high hybridization efficiencies but also contained unexpected homozygous variants within the captured sequence. These variants were not present within the SNP data provided by the 1000 Genomes project for this cell line. A BLAST search using the captured sequence indicated that this variant was derived from off-target hybridization to pseudogenes derived from NPM1. Hence, probes targeting the NPM1 mutational hotspot were redesigned to also target small intronic parts which are not present within the sequences of the pseudogenes, in order to avoid off-target hybridizations.

The altered probe pool was again used for preparation of smMIP libraries as before. Comparison of the sequencing results from the initial and final smMIP pool showed that hybridization efficiencies were significantly more homogenous after rebalancing (Figure 23)

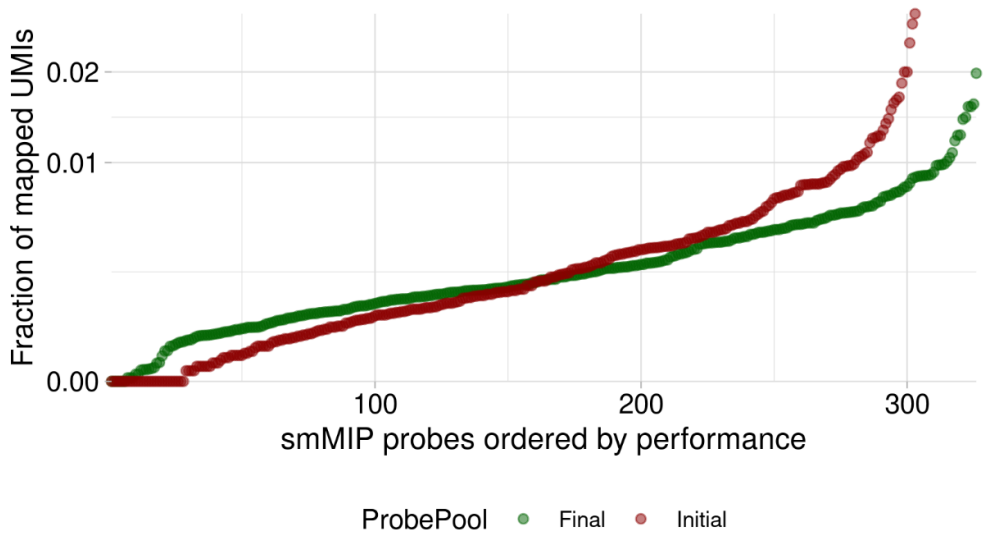


Figure 23: Performance of the initial and final pool of smMIP probes. In order to achieve a more even coverage probes that showed low hybridization efficiencies were either repooled or replaced by alternative smMIPs. The lower slope of the final pool indicates a more evenly distribution across UMIs detected per probe and hence a more even unique sequencing coverage. Most probes that had little to no sequencing coverage within the initial pool were targeting CEBPA and were excluded from the final pool as they did not benefit from higher probe concentrations.

Compared to the initial probe pool the final smMIP pool shows a more uniform coverage across target regions, as indicated by a lower slope in Figure 23. The unique coverage of bad performing probes could also be enhanced by rebalancing. Additionally, performance of over-represented probes showing

high target capture efficiencies was reduced due to their overall lower relative concentrations within the probe pool. The sequences of the resultant final smMIP pool of 326 smMIPs as well as the relative concentration of each probe within the pool are listed in Supplemental Table 1.

Overall, refinement of the smMIP pool during rebalancing experiments allowed to optimize uniformity of sequencing depth across the targets, while increasing the number of overall usable reads by excluding *CEBPA* as target, which showed extremely low capture efficiency and yielded artefactual reads without genotyping information.

4.2.4 Design of an optimized sequencing setup allowing for robust multiplexing of up to 192 libraries

One final goal of the established smMIP assay is the ability to sequence large cohorts in a very cost efficient way. Generally, sequencing costs decrease when using sequencers with higher output which produce more reads per sequencing run. In order to utilize these cost benefits an optimized sequencing setup was established aiming to optimize utilization of flowcell lanes.

As a first indicator for the number of reads necessary to detect most of the UMIs present within the smMIP libraries data acquired within the re-balancing experiments were used for downsampling. Raw sequencing reads were downsampled and the remaining number of detected UMIs was used to estimate the number of reads necessary to detect the majority of UMIs, representing unique capture events (Figure 24). After a steep increase of detected UMIs the curve's slope flattens at around one to two million reads, indicating that most UMIs contained within the library have been sequenced at this point.

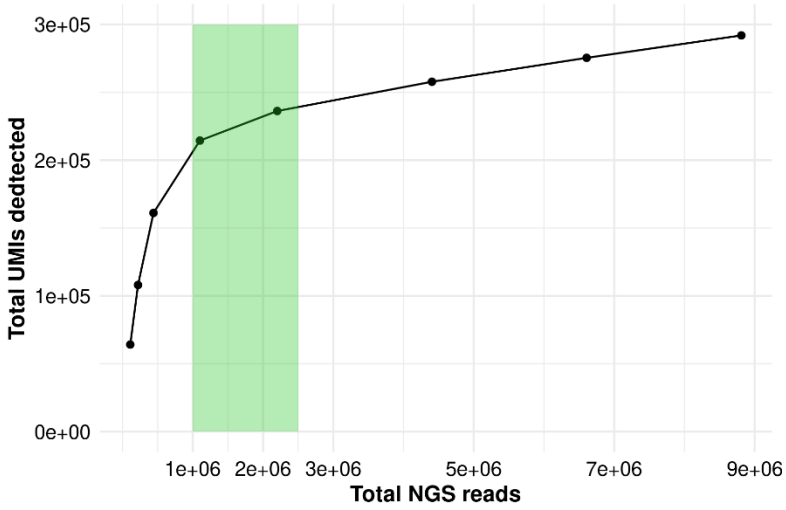


Figure 24: Downsampling of sequencing data from the rebalancing experiments indicates that the majority of UMIs, representing unique capture events, is already being detected with 1 – 2.5 million NGS reads per library (green box). Black dots represent data points obtained from downsampling of sequencing data.

The steady linear increase in total detected UMIs beyond about 2.5 million sequencing reads is most probably caused by sequencing errors which result in generation artefactual new UMIs. Hence, a sequencing depth of about one to two million read-pairs seems to be sufficient to detect most of the UMIs contained within the smMIP library.

In cooperation with the GeneCenter Munich, all prepared libraries were planned to be sequenced on a HiSeq 1500 system. The sequencer's output is equivalent to the newer HiSeq 2500 officially resulting in about 180 million reads per lane for high-output flowcells and 150 million reads per lane using rapid flowcells [*126*]. Due to optimized loading of flowcells resulting in increased cluster densities while retaining good sequencing data quality, sequencing runs at the GeneCenter typically yield over 200 million reads per lane for high-output flowcells.

Considering 1 – 2 million reads necessary per library and an effective output of about 200 million reads per lane, 96 – 192 libraries need to be multiplexed in order to fully utilize the flowcell lanes.

However, the use of single indexing or combinatorial dual-indexing for sequencing libraries bears an increased risk for sample cross-contamination. At library preparation level, a slight cross-contamination of one adapter with another may lead to incorporation of wrong indices into libraries and ultimately assignment of sequencing reads to the wrong sample. Additionally, amplification of mixed clusters on the flowcell during sequencing can also contribute to a wrong sample assignment [*127*]. This problem is further increased when utilizing newer sequencers, like the NovaSeq 6000 or NextSeq 1000 systems, due to the use of patterned flowcells and the new Exclusion-Amplification (ExAmp) chemistry for cluster generation. Here, slight amounts of free indexed adapter primer present in the sequencing libraries can cause incorporation of these primers into any cluster on the flowcell, as binding of library

fragments and cluster amplification can take place simultaneously [*128, 129*]. In the context of the smMIP panel, this could cause a variant present at higher allele frequencies in one individual to also be detected at low variant allele frequencies in another individual due to misassigned sequencing reads, which has already been reported for cancer exome sequencing studies [*130*].

In nearly all cases this 'index hopping' is typically only observed for one of two indices present in dual-indexed samples. Therefore, one way to efficiently avoid misassignment of sequencing reads is the use of a non-redundant dual-indexing setup, in which all libraries sequenced on the same lane of a flowcell are tagged with a unique i5 as well as a unique i7 index [*85*]. In order to avoid problems related to index hopping while also minimizing costs for indexed adapter primers, 192 indexed P5 and 96 indexed P7 adapters were used to enable sequencing of 96-192 libraries on one lane.

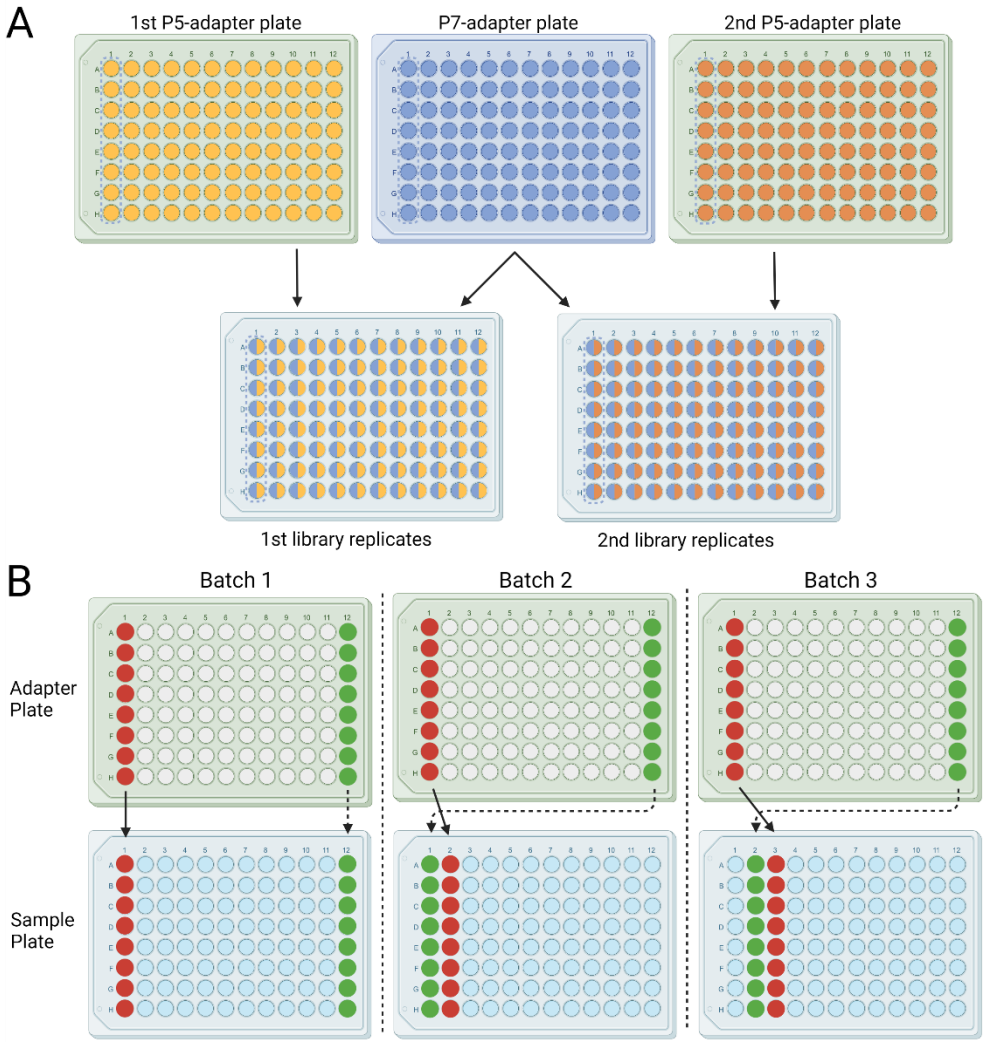


Figure 25: Indexing scheme for multiplexing of up to 192 libraries from 96 samples per lane with reduced risk for cross-contamination between samples. (A) Every sample is prepared in two technical replicates. Both replicates per sample get the same i7-index but different i5-indices. Different biological samples are hence differ in i5- as well as i7-indices making cross-contaminations unlikely. (B) Rotation scheme for indexing of different library batches. In order to prevent cross-contamination of library fragments between library preparation batches, the combinations of i5- and i7-indices are altered by shifting the pipetting scheme for one of both adapters.

To increase the overall sensitivity, sequencing libraries for every individual are planned to be prepared in technical duplicates. Hence, 192 libraries would consist of 96 unique individuals, each of which get assigned one P5 adapter index. Additionally, every prepared library is assigned one of 192 indexed P7 adapters in order to make all libraries distinguishable (Figure 25A).

Libraries of different individuals can consequently be distinguished by both sample indices, while technical replicates per individual have one index in common. Using this strategy 192 libraries can be multiplexed on one lane, while misassignment of sequencing reads is expected to be restricted to technical replicates of one individual.

Furthermore, smMIP sequencing libraries are planned to be prepared in batches of 96 samples for bigger sample sizes, resulting in 192 libraries per batch which make use of the full set of adapters. In order to minimize undetectable cross-contaminations of library fragments between batches the combinations of P5 and P7 adapters used are modified for each library preparation batch (Figure 25B).

In summary, the established sequencing multiplexing scheme allows for a high degree of multiplexing with up to 192 samples per lane, while minimizing the risk of cross-contaminations and misassignment of sequencing reads to allow for very high precision of variant calls.

4.2.5 Establishment of a custom smMIP analysis pipeline enables analysis of hundreds of samples with minimal hands-on time

To complement the high-throughput workflow for the generation of highly sensitive genotyping a computational analysis pipeline is needed to homogenously process the sequencing data for hundreds of samples. However, besides the analysis pipeline included within the MIPgen software package, no software solution existed that would allow to automatically process sequencing data derived from smMIP sequencing libraries. Although this would generally be possible using the MIPgen analysis pipeline in combination with a pile-up based variant caller, due to software bugs encountered during analysis of re-balancing experiments and insufficient flexibility a new processing pipeline was developed.

The new analysis pipeline was planned to be highly flexible, by using and preserving raw reads as well as UMI-deduplicated data for every detected variant call. Furthermore, the processing of hundreds of samples should be possible with low hands-on time.

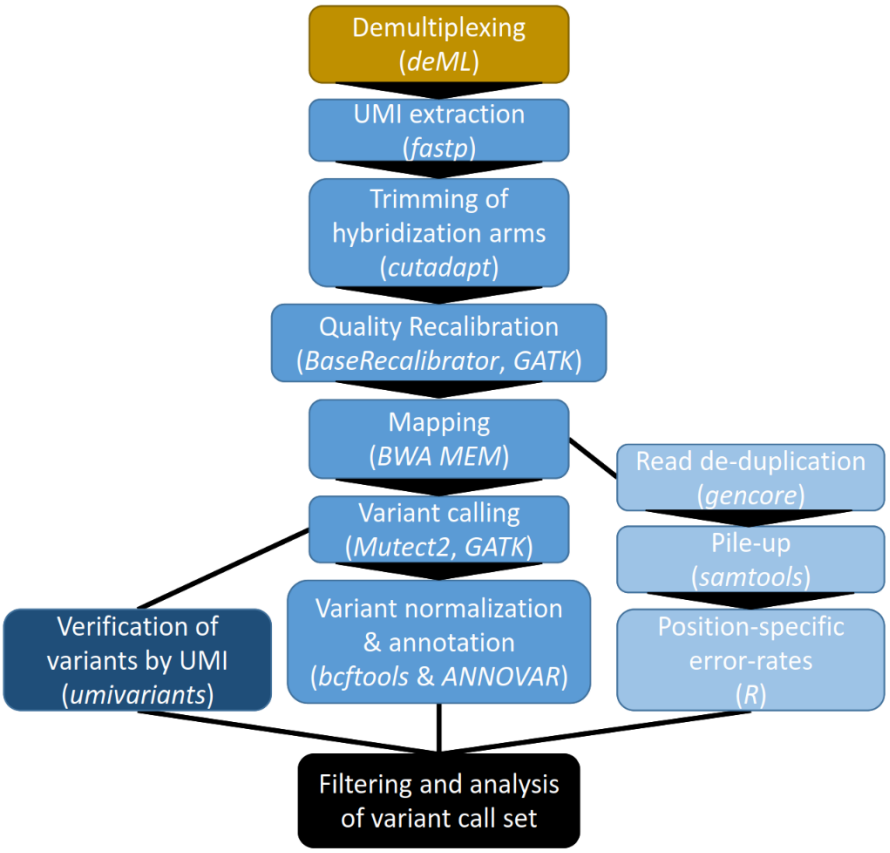


Figure 26: Overview of the smMIP analysis pipeline. The first part of the analysis pipeline consists of UMI-extraction, trimming of hybridization arm sequences from the raw reads, mapping via BWA-MEM, as well as Base Quality Score Recalibration and variant calling via Mutect2. Subsequently detected variants are normalized (bcftools) and annotated (ANNOVAR). These processes can be started for hundreds of samples in parallel with a single Bash script. A second script is used to generate a position specific error rate for all positions within the target areas based on the UMI-deduplicated reads (gencore). Additionally, the R package 'umivariants' verifies the UMI support for all variants detected within the non-deduplicated set of all reads. Finally, filtering of variants based on all acquired information in order to reduce false-positive variant calls is carried out in R.

In a first step all sequencing libraries are being demultiplexed using *deML* [*87*] and named in a systematic way that allows to easily identify technical replicates derived from the same sample. The downstream processing of the sequencing data can be started via a bash script for all sequenced samples at once (Figure 26).

4.2.5.1 The SLURM workload manager enables batch processing of samples

The SLURM (Simple Linux Utility for Ressource Management) workload manager [*131*] is a tool that enables flexible scheduling of processes on linux servers. Processes can be submitted to the SLURM controller and are executed when enough resources are available on the server.

Using a bash script for submission of jobs to the SLURM workload manager on the server, the further processing of raw sequencing data to the final variant calls can be started using by a single command. This script contains all necessary processing steps needed for variant calling within the smMIPs sequencing. Additionally, job-arrays within the SLURM environment are used to allow simultaneous submission of hundreds of jobs. Hence, analysis for all acquired sequencing data can be started at once with only a predefined number of samples being processed in parallel in order to leave server resources for other users.

4.2.5.2 Handling of UMIs in raw sequencing reads

Within the first step of the automated processing pipeline the UMIs present within the first five nucleotides within forward and reverse sequencing reads get trimmed and moved to the header-line of the respective read within the fastq-file using *fastp* [*132*]. As technical duplicates of each biological sample are merged afterwards, the UMIs are artificially appended by 'AT' or 'GC' in order to prevent the UMI-collisions between the technical replicates. Additionally, this measure allows to preserve the information about which UMI was derived from which replicate and could potentially be further utilized, e.g. to prioritize variants detected in both replicates of a sample.

4.2.5.3 Trimming of hybridization arm sequences

After handling of the UMI sequences, sequencing reads derived from technical library duplicates of the same biological sample are merged by concatenating the respective fastq files.

The resulting sequencing reads still carry the hybridization arms used to anneal the probes to their target regions during hybridization on both ends of the fragment. These sequences are part of the oligonucleotide-probes themselves and therefore cannot contain any genotyping information, even if variants would be present in these regions. Sequences from hybridization arms that overlap with the capture sequence of a neighboring probe would therefore skew the allele frequency of any variant within this overlap. Hence, arm sequences are removed using cutadapt [*88*] in the linked adapters trimming mode. Here, a list of paired ligation- and extension-arm sequences, corresponding to each smMIP probes' arms included in the panel, is used for trimming. Sequences at the beginning of the reads have to closely match one of the combinations known to be present within the probe pool for every pair of reads in order to get trimmed. All read-pairs that do not match one of these expected combinations are

discarded at this step in order to exclude any artefactual or contaminating reads from downstream analyses.

4.2.5.4 Mapping and variant calling based on duplicated sequencing reads

Trimmed read-pairs are subsequently mapped to the reference genome using *BWA MEM* [*133*]. For downstream processing of mapped reads the Genome Analysis Toolkit (GATK), published and maintained by the BROAD institute, is utilized. Before variant calling, the Base Quality Scores of all reads are being recalibrated using the GATK BaseRecalibrator. For variant calling GATK *Mutect2* was chosen as it allows variant calling without a matched normal sample and shows a high sensitivity and precision also at low variant allele frequencies. This is achieved by local re-alignment of mapped reads and assembly of possible haplotypes instead of only looking for mismatches to the reference sequence within the mapped reads, resulting in improved precision especially in difficult regions. All variant calls are being generated based on raw reads, without utilization of the UMI information and additionally using parameters to increase the sensitivity of variant calling. Thereby a set of all possible variants that may be present within the sample is being created. After normalization using bcftools, all variants get annotated *ANNOVAR* [*90*], adding information about known SNPs, pathogenic variants and deleteriousness of mutations.

4.2.5.5 Validation of potential variants based on UMI information

At this point, the created list of variants is based on non-deduplicated sequencing reads, completely ignoring the UMI information in order to create a list of possible variants with the highest possible sensitivity.

In order to utilize UMI information for error-correction "*umivariants*" was utilized. This R package was developed by Ilse Valtierra [*93*] and allows to verify a list of potential variants by building single-molecule consensus reads. For every potential variant, reads covering the specific genomic position are grouped by their UMI and a consensus read is calculated. Thereby the potential variants discovered in the raw sequencing data are being re-evaluated based on the UMI consensus reads, which allows to reduce the number of false-positive variants originating from errors introduced by sequencing and PCR amplification as well as more accurate determination of variant allele frequencies.

4.2.5.6 Calculation of position-specific error rates to reduce false positives at very low variant allele frequencies

Many false-positive variant calls can be present at very low frequencies within sequencing data due to random errors derived for example from PCR. When introduced at the gap-fill reaction or very early cycles of PCR these technical artefacts cannot be corrected using the UMIs of the smMIP sequencing data. Furthermore, sequence contexts like homopolymer stretches of repetitive elements may lead to more error-prone positions within the genomic target regions.

To reduce the number of false-positive variant calls at very low frequencies, calculation of a position-specific error rate has been integrated into the pipeline.

In a separate process the mapped reads are being de-duplicated based on their UMI using the software package *gencore*. Pile-ups based on the de-duplicated reads are generated using *samtools* to count matches and mismatches to the reference at every position. Data from all processed samples is subsequently used to create a background-error rate for each position within the genomic targets. In order to exclude mismatches derived from real genetic variants from calculations of the error rate, positions that show more than 5% mismatches in a given sample are excluded for calculation of the overall error rate. The calculated error rates at each covered genomic position as observed within all processed samples are summarized in a text-file output.

4.2.5.7 Final processing and filtering of detected variants

The set of detected variants obtained from *Mutect2* is merged with the UMI-based evaluation of the variant set from *umivariants*. A text file containing all information of the detected variants and their UMI-based evaluation is the final output of the analysis pipeline and can subsequently be imported into R for further analysis.

As a first step for downstream analysis within R, the obtained position-specific error rates are utilized to estimate the likelihood of being caused by technical noise for each potential variant. For that a p-value based on Poisson distribution is calculated using the unique coverage, the number of UMIs supporting the variant together with the observed error rate across all samples at this position. The p-values are subsequently corrected for multiple testing by the Benjamini-Hochberg procedure. Using this approach allows to identify and discard low-frequency variants that are likely derived from technical artefacts.

The resulting final data set can be further filtered based on custom criteria, e.g. in order to discard germline mutations and minimize possible false-positive calls.

4.2.6 The established smMIP workflow shows robust performance across a cohort of 561 individuals and allows for balanced multiplexing of library pools

As a first application for the established smMIP assay a cohort study using samples derived from 386 AML patients in long-term remission was carried out. 159 samples from age-matched healthy individuals were included as controls to which the frequency of detected clonal hematopoiesis as well as its mutational spectrum is to be compared. Six samples with known mutations derived from AML patients at diagnosis were included as quality controls. For additional quality control library preparations for 16 samples of the long-term survivor cohort were repeated in a separate batch to assess reproducibility of variant detection. Lastly, four gDNA dilution series with a total of 16 samples were prepared to assess sensitivity and precision of the variant frequencies for the assay.

The gDNA samples of 551 pseudonymized individuals were provided by Klaus Metzeler and Frank Ziemann and used for preparation of NGS libraries utilizing the established smMIP assay. Library preparations for patient samples and additional quality controls were carried out in technical duplicates, resulting in a total of 1102 libraries, which were prepared in seven batches of 150 – 192 libraries.

Two days were needed to prepare one batch consisting of up to 192 libraries from up to 96 biological samples, clearly demonstrating the suitability of the smMIP assay for high-throughput projects.

Only five drop-out samples for which library preparations failed were observed, corresponding to less than 1% of all samples. The most plausible explanation for these drop-outs is inaccurate quantification of gDNA concentrations, which resulted in too high dilutions of the gDNAs and hence too few template DNA as input for the hybridization reactions.

Generally, this low dropout rate reflects the robustness of the established workflow for preparation of libraries using the new smMIP panel and emphasizes its utility in conducting cohort sequencing studies at low prices and short turn-around times.

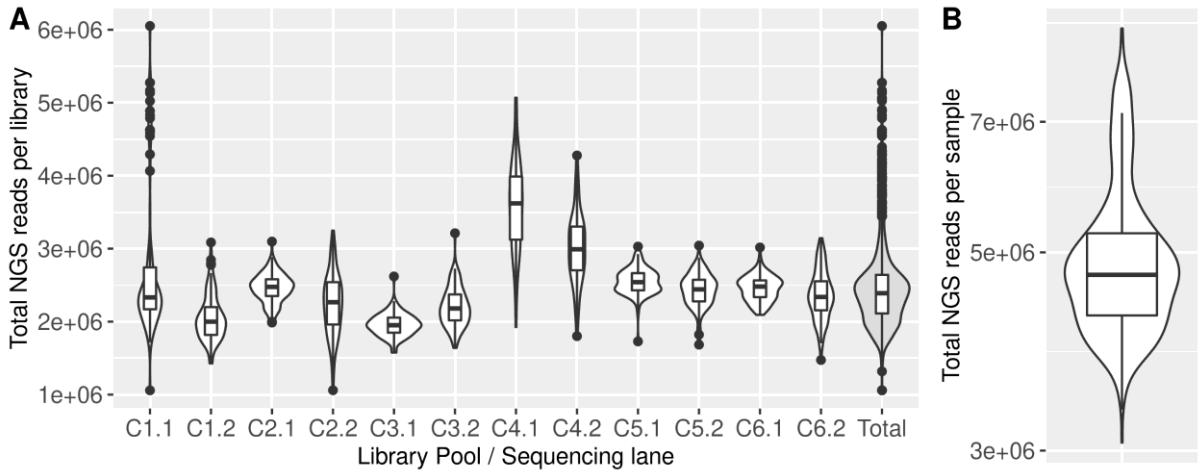


Figure 27: Sequencing reads acquired for samples throughout the sequencing batches of the individual cohort samples. Total reads per library within each library pool (A) were well balanced. Batch C1.1 showed an increased spread of sequencing reads across libraries which was caused by complications at library DNA quantification before pooling. (B) Total sequencing reads for both technical replicates per sample were above the planned 4 million reads in most cases.

The total sequencing reads obtained per sequencing library within each sequencing batch (Figure 27A) show a very narrow distribution. Overall, the anticipated two million sequencing reads were obtained for most of the sequencing libraries. Due to an overall lower yield of sequencing reads for some flowcell lanes batches C1.2 and C3.1 include many libraries that got slightly less than 2 million sequencing reads. However, as one million reads are still expected to provide sufficient sequencing depth, all sequenced samples could be included for downstream analyses.

The total reads obtained for both technical library replicates per sample (Figure 27B) show that a median of about 4.8 million raw sequencing reads down to a minimum of 3 million reads were acquired. Hence, all samples within the cohort received sufficient sequencing reads to ensure enough coverage across target regions and enable sensitive variant calling.

In summary, the established workflow showed robust performance for library preparations within the first cohort study. Additionally, the library pooling and sequencing layout allowed to relatively uniform sequencing depth across all samples and sequencing runs

4.2.6.1 Dilution series indicate high sensitivity of the established smMIP assay down to 0.7% VAF

In order to assess the precision of variant allele frequencies, pairs of patient samples differing in heterozygous SNP alleles were chosen to create dilution series as a quality control. By serially diluting the gDNA from one patient with gDNA from another patient low frequency variants with expected frequencies of 8.3%, 2.8%, 1.4% and 0.7% were created. These variants were used to test the precision of obtained VAFs, as well as to ensure sufficient sensitivity for low frequency variants. A total of four

dilution series were prepared and used for library preparations. Analysis was carried out using the newly established smMIP analysis pipeline.

All variants investigated in the dilution series were correctly identified by the analysis pipeline, even at the highest dilution step with an expected VAF of 0.7%, indicating a high sensitivity of the smMIP assay. The determined VAFs at the highest dilution ranged from 0.28 to 1.63% with a median of 0.7% (Figure 28A). The unique coverages for the SNPs analyzed at the highest dilution step ranged from 184x to 7011x with a median of 1674x. In order to assess whether the deviations between observed and expected VAFs are caused by sampling effects and differing sequencing depths across the analyzed SNPs, the 95% confidence interval for UMIs carrying the variant given different sequences depths and the expected variant allele frequency of 0.7% were calculated according to binomial distribution (Figure 28B). Most measured VAFs were within the range expected from sampling noise, however three measurements fell outside the intervals.

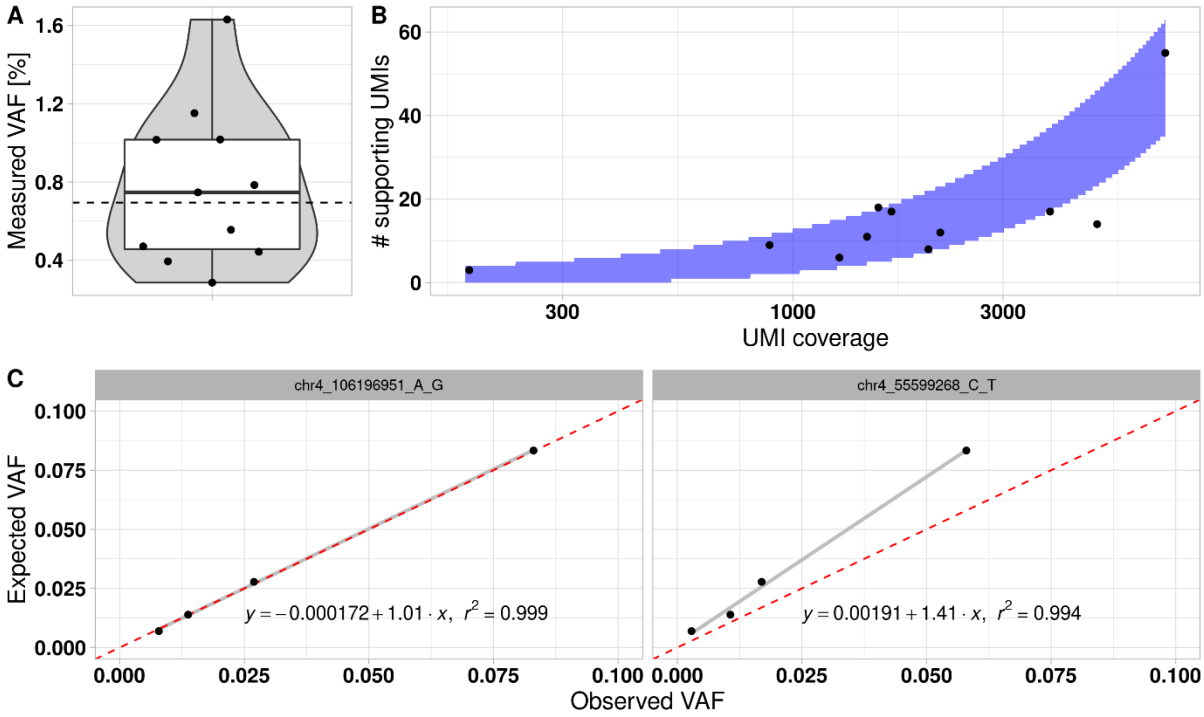


Figure 28: Data for the dilution series experiment used as quality control. (A) VAFs of the variants detected at the highest dilution factor. The dashed line indicates the expected VAF of 0.7%. (B) Number of detected UMIs supporting the diluted variant in relation to the unique sequencing coverage achieved. The blue area indicates the 95% confidence interval for the number of expected UMIs supporting the variant based on sampling effect estimated by the binomial distribution. (C) Expected VAF in relation to observed VAF for two exemplary SNPs analyzed within the same samples of one dilution series. Correlation of determined VAFs are generally very high throughout all analyzed dilution steps. While linear regressions for most SNPs showed the expected slope of one (left panel), some SNPs showed slopes differing from this expectation (right panel).

In order to further exploit these unexpected deviations in to the expected VAFs, linear regression were carried out for each SNP. When comparing the concordance of VAFs between different dilution steps per SNP (Figure 28C), linear regressions generally showed precise fits with adjusted R^2 values being above 0.96 for all 11 analyzed SNPs with a median value 0.995. Interestingly, some of the regressions

showed slopes differing from the expectation (Figure 28C right panel). These deviations cannot be caused by inaccurate dilutions of the gDNAs, as SNPs within one dilution step are derived from one library preparation and hence the same gDNA template. Consequently, the same change to the expected slope would be present in all tested SNPs of one dilution series, which is not the case. Only seven of the eleven analyzed SNPs accurately resemble dilutions of heterozygous SNPs while regressions for the other four SNPs show significant deviations in their slopes.

All four gDNA samples used for creation of the dilution series were derived from individuals of the long-term AML remission cohort. Within three of these samples low frequency variants were detected indicating the presence of subclones at frequencies between 1.2 and 4.4%. A loss of heterozygosity including the respective SNP positions within these subclones can lead to an increase or decrease in the overall variant allele fraction present within the diluted sample, which could explain the observed deviations. Another possible explanation for this observation might be related to patients having received an allogeneic hematopoietic stem cell transplantation (HSCT). Even years after transplantation, residual recipient cells may still contribute to hematopoiesis [*134, 135*]. If donor and recipient differ in some SNPs, this chimeric hematopoiesis can explain the observed differences as the variant allele frequencies of specific heterozygous SNPs consequently differs from the expected 50% in the patient sample.

The fact that not all putative SNPs analyzed within the dilution series were initially present at 50% allele frequency explains that some of the allele frequencies observed at the highest dilutions steps differed from the expectation.

In summary, the results of the dilution series quality control demonstrate that the established smMIP panel is able to reliably detect variants even at very low frequencies of 0.7% allele frequency.

4.2.6.2 Comparison of smMIP variant calls to Haloplex data shows good agreement but difficulties to detect larger insertions and deletions

In order to assess the reliability of variant calls using the established smMIP panel and analysis pipeline, data for 36 individuals was cross-validated using data obtained from using a commercial sequencing panel.

This subset of samples included in the cohort had already been sequenced by Maja Rothenberg-Thurley (Klinikum der Universität München) using a custom commercial Haloplex assay for targeted re-sequencing. The panel targets over 106 kb in 68 genes recurrently mutated within myeloid

neoplasms [*28*]. To see whether the smMIP panel is able to reliably detect variants previously detected using the Haloplex panel, all variants detected in the target regions shared by both assays were compared. Variant calls generated using the smMIP assay were provided to Dr. Maja Rothenberg-Thurley (Klinikum der Universität München) who carried out the comparison in order to adhere to the data protection guidelines related to the data acquired during routine diagnostics in the clinics.

Overall 54 variants were detected within the smMIP data, whereas 49 variants were detected in the Haloplex data. 44 of these variants could be detected within the smMIP variants calls as well as in the Haloplex data (Figure 29).

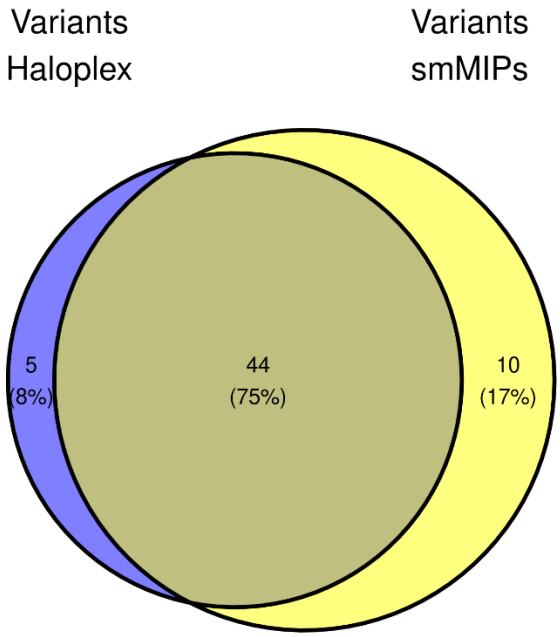


Figure 29: Venn-Diagram of variants detected within 36 individuals using the commercial Haloplex sequencing panel and/or the newly established smMIP assay. 44 variants were detected by the smMIP panel as well as by the Haloplex panel. Another 10 variants were only detected by the smMIP assay and five variants detected in the Haloplex data were not included in the smMIP variants calls.

The majority of variants detected by the Haloplex panel were also detected by the established smMIP panel (44/49 variants), indicating that the sensitivity of the established assay is comparable to the commercial Haloplex panel. However, five variants identified using the Haloplex panel were not present within the filtered smMIP variant calls.

Two of these variants located at the SRSF2 hotspot locus were present within the initial variant calls but filtered out in post-processing steps. Hence, these variants were detected but not included in the final smMIP variant call set. The SRSF2 hotspot locus is currently hard to genotype using the smMIP assay, as probes in this area show low hybridization efficiency. Consequently, low unique coverages only allow to call higher frequency variants, as variants need to be supported by at least 3 UMIs using the current preliminary filtering strategy.

Two other undetected variants represent *FLT3*-ITDs (internal tandem duplications) in the form of 27 and 52 nucleotides long insertions. One possible explanation is that the insertions led to reduced hybridization efficiency due to the increased target capture size which requires the template gDNA to bend stronger. The last variant not detected by the smMIPs comprises a 52 nucleotide deletion within the CALR gene at about 30% variant allele frequency. Due to the length and positioning of the deletion the DNA sequences used to hybridize both probes that should cover this region is either partially or completely missing. Hence, this mutation can technically not be detected using the current set of smMIPs.

All ten variants that were exclusively detected by the smMIPs panel were below the detection minimal detection threshold of 1% variant allele frequency used for the Haloplex panel. For five of these variants raw Haloplex data were manually checked for the presence of these mutations. In three of five cases raw reads carrying the variant previously identified using smMIPs could be observed. Hence, most of the variants exclusive to the smMIP variant set are most likely caused by their very low variant allele frequencies although it cannot be excluded that some of these mutations could still represent technical artefacts.

In summary, comparison between the smMIP assay and the commercial Haloplex assay indicates good agreement of both methods for detected variants down to allele frequencies of about 1%, clearly demonstrating the assay's competitive performance for detection of small subclones by targeted sequencing of their driver mutations.

4.2.6.3 Re-sequencing of patient samples hints towards elevated error rates at very low allele frequencies below 2%

Another aspect for the reliability of the smMIP assay is the reproducibility of variant detection. For this purpose 16 individuals that had already sequenced in prior library batches were again used for library preparation within a new batch. In order to have fully independent technical replication new gDNA from archived blood samples was prepared by Sebastian Tschuri (Klinikum der Universität München) to serve as a template for a new round of library preparations.

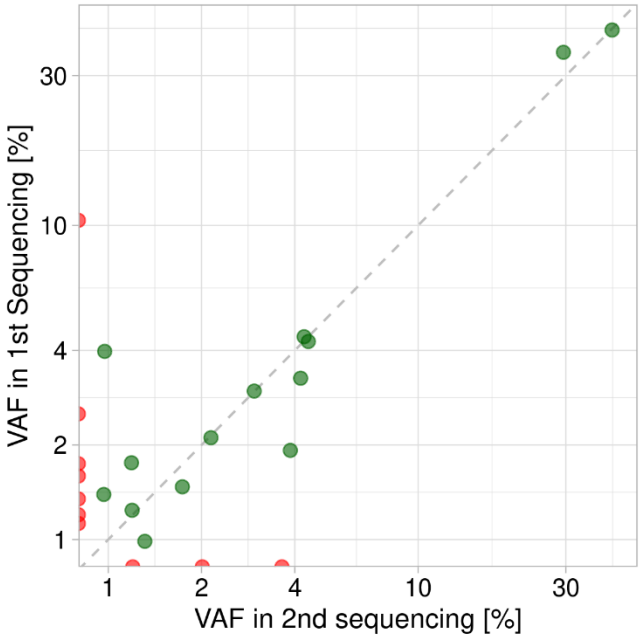


Figure 30: Variants detected within 16 patient samples that were used twice for independent library preparations and sequencing runs 14 variants (58%) were detected in both batches, while 7 variants (29%) were exclusive to the first and 3 variants (12%) were exclusive to the second batch.

For this analysis only variants above 1% VAF were considered in order to exclude most artefacts associated with very low frequency mutations. A total of 24 variants were detected of which 14 mutations were successfully called in both sequencings (Figure 30). However, seven mutations were exclusively detected in the first sequencing and three other mutations remained exclusive to the re-sequencing libraries.

Six of these ten variants were detected below 2% allele frequency, represented by three to seven UMIs, and located within low-complexity sequence contexts. Hence, they might be representing technical artefacts derived from polymerase or sequencing errors not discarded by the current filtering strategy. Interestingly, two of these variants were present within the mapped reads of the other batch but not within the UMI-deduplicated reads, indicating that these variants are likely derived from technical noise.

The rest of four mutations observed only in one experimental batch had variant allele frequencies over 2%. One mutation with 10.4% variant allele frequencies could be detected when looking at the raw and de-duplicated mapped reads of the other batch at the same variant allele frequencies. As the unique sequencing coverage was well above 1000x the absence of this variant has to be related to *Mutect2*, used to call variants within the established pipeline. The remaining three mutations detected with 2%, 2.5% and 3.6% variant allele frequencies in only one of both data sets could also be observed within the raw variant call set of the other batch but were only supported by one to two UMIs. For example, the variant at 2.4% VAF (*DNMT3A*, chr2:25463283, A>G) was present in the initial data set with 5 UMIs support

at a unique coverage of 199x. However, the unique coverage obtained for the re-sequencing data was at 107x with only one UMI supporting the variant. Subsequently this variant was discarded from the final variant call set by the preliminary established filters, which require support by at least three UMIs to exclude noise from technical artefacts. Hence, the inconsistencies in detection of these variants is caused by their low frequencies in combination with differences in coverages between the batches and the current preliminary thresholds for variants filtering.

Taken together, variant calls of the experimental replicates showed good agreement. Six mutations below 2% frequency could not be detected within data of the other respective batch. All of these variants are part of low-complexity regions (e.g. chr2:25471068 CT>C with 1.6% VAF; CCCCGGGCCC-CT>C-GGTTTTCTTCC) suggesting that they are representing technical artefacts not removed by utilizing UMI-consensus reads. However, all mutations above 2% variant allele frequency were detected in data of both sequencing batches, although some were not part of the final variant call set due to the applied filtering or issues related to variant calling using *Mutect2*. Additionally, variants below 2% frequency may become difficult to detect in regions with low unique coverage as coverage across targets can vary between repeated batches of library preparation and sequencing.

However, the data of the re-sequencing quality controls showed reliable and reproducible detection of variants down to 2% allele frequency in areas with sufficient coverage. Using the current sequencing depth the established smMIP-panel's sensitivity for reliable detection of variants is therefore estimated to be comparable to the Haloplex assay utilized at the Klinikum der Universität München, for which a cutoff of 2% VAF for reliable identification of variants has been reported [*28*] with the advantage of being significantly more cost-efficient.

5 Discussion

5.1 The established cellular barcoding approach has proven to be an efficient way to directly investigate subclonal heterogeneity within AML Pdx samples

Two different constructs for cellular barcoding have been successfully cloned, quality controlled and successfully applied for studying clonal heterogeneity within an AML Pdx mouse model.

The established workflow allowed to clone high-complexity plasmid pools with over 10 million barcodes estimated to be present within the high complexity DNA-barcode plasmid pool and over 700,000 barcodes within the expressed barcode plasmid pool. Quality controls showed that both pools can be utilized to barcode thousands of cells with no or minimal re-usage of barcode tags and therefore allow to precisely quantify cell numbers.

Both construct have been utilized in several pilot experiments, demonstrating the power of the barcoding approach to investigate heterogeneous cell populations. Application of the established barcoding pools allowed to quantify the bottleneck for serial-transplantation of AML Pdx cells in order to verify that subclonal heterogeneity of AML Pdx samples can theoretically be retained over many passages within the utilized mouse model. Furthermore barcoding allowed to directly observe differential response of subclones towards in vivo chemotherapeutic treatment within multiple biological replicates of the same barcoded Pdx line population, thereby robustly proving the presence of resistant subclones that are less sensitive to *in vivo* chemotherapy within this Pdx sample. Moreover, using an experimental setup for limited dilution transplantations with barcoded AML Pdx cells showed that cellular barcoding can potentially decrease the number of experimental mice necessary to estimate the frequency of leukemia initiating cells within these samples. Additionally, barcoding allowed to identify samples derived from a single leukemic cell that engrafted in this experiment. These single-cell isolates were determined to represent genetically different subclones based on known marker mutations. The single-cell clones were further characterized based on their transcriptomes, exomes and proteomes as well as their phenotypes in competitive *in vivo* experiments. Results of these experiments, for which the established cellular barcoding constructs provided the groundwork, are published in the *Journal of Hematology & Oncology* [*120*] (see *7. Publications*).

5.1.1 Barcoding allows to reduce the number of experimental mice needed for determination of LIC frequencies

The use of cellular barcodes in a limiting dilution transplantation assays has shown, that barcodes enable a robust read-out for the number of engrafted cells that allows to estimate the LIC frequency more precisely than binary read-out of engraftment rates. As the number of engrafted cells can be quantitatively assessed the number of cells used for transplantation can be increased, thereby reducing the number of mice not showing successful engraftment. This advantage potentially allows to significantly reduce the number of mice needed to determine the LIC frequency of AML samples. The LIC frequencies obtained via the barcode analysis were comparable to those derived from engraftment rates when excluding outliers most probably derived from technical errors. Regulations and laws about the protection of experimental vertebrate animals, e.g. 'EU-Tierschutzrichtlinie', 'Tierschutz-Versuchstierverordnung' and 'Tierschutzgesetz', have implemented the 3R-principles [*136*] requiring to replace animals where they are not necessary, reduce the number of animals utilized and refine experimental methods in order to minimize stress. Therefore the usage of cellular barcoding for determination of LIC frequencies should be highly recommended for determination of LIC frequencies as it allows to reduce the number of mice needed.

5.1.2 Comparison of engrafting cells upon transplantation of low and high number of cells

Compared to the determination of LIC frequencies at extremely low cell numbers the bottleneck observed for serial transplantations of Pdx lines at high cell numbers, i.e. the fraction of cells showing successful engraftment and proliferation, is significantly more important for utilization of the AML Pdx mouse model. Knowledge about this bottleneck is essential in order to estimate whether the subclonal heterogeneity of specific Pdx lines can be conserved throughout multiple serial passages at a given cell number.

In a first proof-of-principle experiment, the fraction of engrafting cells with 550,000 cells transplanted was estimated to be about 0.04%, corresponding to 1 in 2,500 transplanted cells. In contrast, the LIC frequency determined in another experiment utilizing very low cell numbers at limiting dilutions was estimated to be 1 in 1620. This fraction of leukemia initiating cells would generally also be expected to show successful engraftment upon transplantation. Yet, the fraction of engrafting cells at 550,000 transplanted cells is significantly lower.

The observed difference in engrafting cell fractions could potentially be derived from the additional filtering of detected barcodes applied for determining the bottleneck for serial transplantations, for which many barcodes at very low frequencies had been detected. It has been shown that barcode read-out via PCR can lead to formation of artificial false-positive barcode calls due to PCR errors and PCR-mediated recombination [*108, 114*]. Although most errors introduced by PCR and the sequencing itself are supposed to be corrected for by clustering of barcodes based on their hamming, the presence of false-positive barcode sequences cannot be excluded as the true sequences of barcodes within plasmid and lentiviral library are not entirely known. Therefore barcodes that were detected in only one of three technical replicates per sample were excluded from further analysis in order to avoid underestimation of the passaging bottleneck. Due to the unexpected strong skewing within relative barcode frequencies this filtering might have discarded true barcodes present at very low frequencies.

However, the size of the bottleneck, i.e. the fraction of cells engrafting and proliferating upon transplantation of high cell numbers, could differ from the LIC frequency determined when transplanting very few cells due to biological reasons. Upon injection of leukemic cells into the bloodstream, cells preferentially home within the bone marrow, which provides an optimal microenvironment. Within the bone marrow LICs home to so-called 'niches' which represent specialized compartments that provide optimal microenvironments for long-term maintenance of HSCs. Consequently, LICs directly compete with healthy HSCs for niche spaces upon arrival within the bone marrow [*137*]. Similarly, a higher number of transplanted LICs may result in increased competition for optimal niches among these leukemic stem cells, resulting in an overall decreased engraftment efficiency.

Given the current experimental data it cannot be excluded that such additional competition for bone marrow niches decreases the engraftment efficiency for transplantation at higher cell numbers. Therefore the number of cells engrafting at higher cell numbers which are regularly used for serial passaging of Pdx lines, i.e. the 'passaging bottleneck', represents an important read-out to determine the suitability of the passaging setup to retain the subclonal complexity within Pdx lines.

To determine whether competitive effects among leukemic cells are causative for the lower fractions of cells engrafting after transplantation of high cell numbers compared to very low cell numbers as observed in the initial experiments, a new experiment could be carried out. The same barcoded cell population would need to be transplanted into several experimental mice at low as well as high cell numbers. This approach ensures that differences in relative engraftment efficiencies are not caused by differences in fitness of the transplanted Pdx cells that might be caused by utilizing Pdx cells derived from different passage numbers or the experimental procedures itself, e.g. thawing of Pdx cells or *in vitro* lentiviral transduction.

5.1.3 Limitations of the current barcoding constructs

The newly established barcoding systems demonstrated good performance in multiple proof-of-principle experiments. However some limitations could still be enhanced.

The high complexity DNA barcode was cloned undirectionally, but due to initial problems the PCR amplification strategy had to be adapted in order to allow for efficient removal of adapter-primer dimers that would interfere with sequencing. As both insert orientations are equally likely to happen within the ligation reaction, only half of the barcodes within the pool are detectable using the established PCR. Thus the current DNA barcode analysis is limited to about half of the barcoded cell population and blind to other half. This drawback is less relevant when using expanded cell populations for investigating differential responses of barcoded cells by their frequency within the population (see 4.1.6) as all experimental mice share the same population of barcoded cells.

However, when barcodes are used without prior expansion in multiple mice, for example to determine the overall number of engrafting cells (see 4.1.3), the read-out becomes less precise. Although the number of detected barcodes can be doubled in order to estimate the total number of engrafted cells, the fraction of detectable barcodes will be differing between biological replicates due to additional sampling variance. As this sampling variance is dependent on the sampling size, i.e. the number of barcoded cells injected into recipient mice, the number of estimated barcodes will vary more strongly between biological replicates when carrying out experiments with low cell numbers.

Hence, for experiments comprising transplantation of low numbers of barcoded cells the expressed barcode construct is more suitable. For this construct all barcodes are extracted using the established PCR amplification strategy. Here, the barcode is positioned within the 3'-UTR of the lentiviral marker gene and hence also detectable via 3'-scRNAseq. Therefore, barcode frequencies can be measured population wide based on targeted amplification of bulk gDNA thereby providing a measure of fitness, e.g. under the influence of chemotherapeutic treatment. Additionally, fractions of the cell population can be used for preparation of single-cell RNA sequencing libraries. As the expressed barcodes are part of the marker transcripts, they can be directly detected within the reverse-transcribed cDNA. The subsequent association of transcriptomic data to cellular fitness of specific clones, as determined by the population-wide barcode readout, represents a first step beyond pure observation of subclones towards their characterization. In recent publications, this approach enabled further insights into the fate determination in hematopoiesis [*138*] and could furthermore prove distinct transcriptional responses of subclones towards chemotherapy in an ALL Pdx model [*139*]. Although this advantage has not been utilized within the first experiments, bulk RNAseq data of the generated single-cell isolates indicated that the expressed barcodes can be detected within Prime-seq libraries (data not shown).

However, to avoid destabilization of the marker transcript the barcode within the 3'-UTR was kept as short as possible, thereby limiting the maximum theoretical complexity of the barcode. Its application is therefore limited to lower cell numbers in the range of a few hundred barcoded cells, which is already sufficient for many AML Pdx experiments.

In summary, most experimental setups within the AML Pdx model system are feasible using one of both established barcoding constructs. The established barcoding constructs therefore represent a reliable foundation for further barcoding experiments in order to elucidate clonal evolution within the AML Pdx model used by my collaborators at the group of Prof. Dr. Irmela Jeremias (Helmhotz Zentrum München).

5.1.4 A new high-complexity expressed barcode construct combines advantages of both initial constructs

In order to combine the advantages of both barcoding constructs presented herein the established workflow for cloning of high-complexity plasmid pools was used to create a third barcoding construct. Recent studies have utilized longer expressed barcodes within the 3'-UTR of marker transcripts and did not report on decreased expression levels of marker genes due to interference by barcode sequences [*138-140*]. A new construct was therefore designed and cloned that includes a high-complexity expressed barcode, thereby combining the benefits of both previous barcoding plasmid pools described herein.

The new construct is based on the pBA439 vector [*141*] (gift from Jonathan Weissman, Addgene plasmid #85967) which provides puromycin resistance as well as the fluorescent TagBFP as marker genes. A high-complexity barcode with 22 variable positions was cloned into the 3'-UTR of the marker transcript directly upstream of the bovine Growth-Hormone polyadenlyation-signal (bGH-polyA signal), thereby potentially optimizing the barcode coverage within 3'-RNAseq libraries (Figure 31 A).

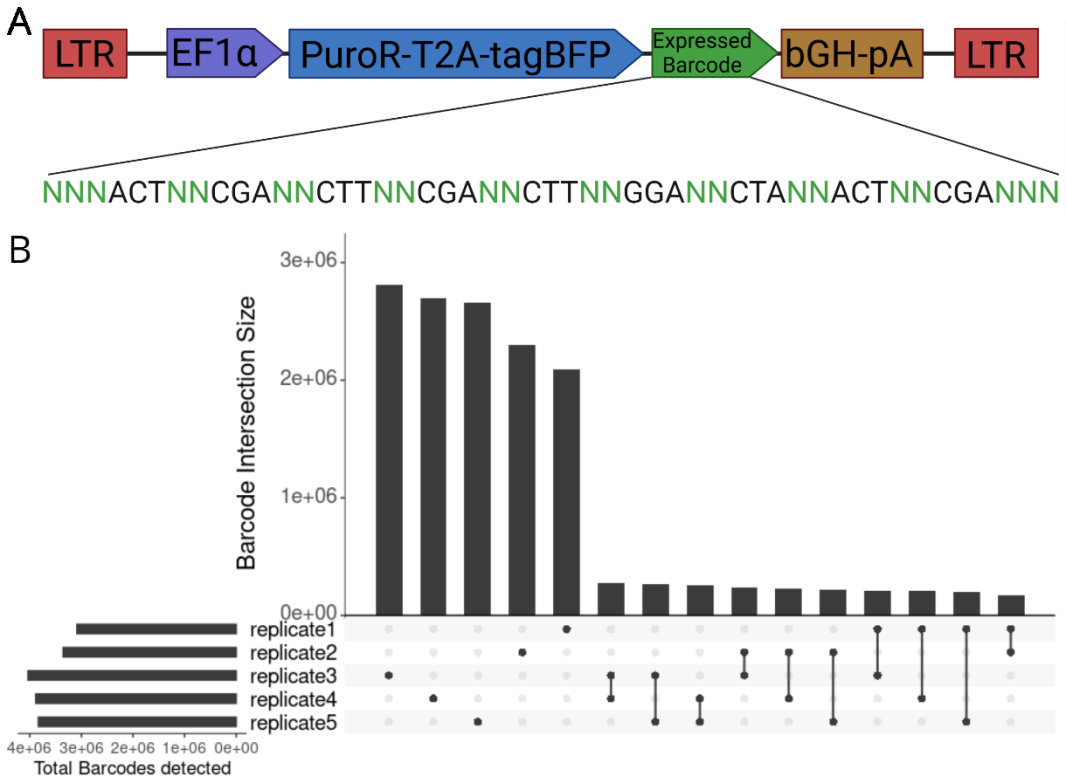


Figure 31: A new barcoding construct combines advantages of both previous constructs using a high-complexity expressed barcode. (A) Schematic overview of the lentiviral insert. Puromycin resistance and the tagBFP fluorescence marker genes allow selection of barcoded cells. The expressed barcode containing 22 variable nucleotide positions is located in the 3'UTR of the marker transcript directly upstream of the bovine growth-hormone poly-adenylation signalling sequence (bGH-polyA). (B) Sequencing of the cloned plasmid pool in five technical replicates indicated a complexity of at least 15 million different barcodes, with 2.1 to 2.8 million barcodes being unique within each replicate. Only overlaps between any two replicates are shown, ranging from 175 – 270,000 barcodes.

The maximum complexity of over 17.5 trillion barcodes (4^{22} possible barcode variants) ensures a high hamming distance between the barcodes. Thus clustering of detected barcodes by sequence similarity can be used to reduce false-positive barcode sequences introduced by PCR and sequencing errors even at high numbers of barcoded cells, consequently increasing precision for determining absolute counts of barcodes. Colony counts from plating of the transformation reactions (data not shown) and sequencing of the plasmid pool (Figure 31B) suggest a minimum complexity of about 15 million barcodes.

Cloning of the barcode insert into the vector was carried out directionally, utilizing two pairs of restriction enzymes that generate compatible sticky-ends. Therefore all barcodes within the pool can be detected by standard PCR even if one primer is located outside the barcode insert, which enables detection and analysis of the whole population of barcoded cells. Additionally, suitable primer pairs that allow for efficient amplification of barcodes without off-target amplification in human and murine gDNA background have also been established.

In summary, the new high-complexity expressed barcode plasmid pool combines advantages of both previous barcode constructs making it a universal tool for future barcoding experiments.

5.1.5 Further advances in methodological approaches enable genotyping and isolation of barcoded clones

Expressed barcodes enable the association of transcriptomes and cellular fitness and provide a first step towards further characterization of subclones in barcoding experiments. Additionally, most known driver mutations are located in exonic regions of genes thereby affecting the respective protein's function. Hence, most driver mutations are present within the full-length cDNA prepared in most bulk and single-cell RNA sequencing library preparation workflows and potentially allow genotyping of the cells to associate the transcriptome to a given subclone.

However, most RNAseq methods rely on sequencing of only the 3'-ends of transcripts in order to provide a cost efficient way for measuring transcript levels [*123, 142-145*]. The detection of coding mutations is therefore mostly limited to variants located at the 3'-end of the transcript, which significantly limits the number of mutations that can be detected using conventional 3'-RNAseq methods. Most of these approaches introduce cell-specific barcodes to the cDNA ends at the reverse-transcription step, which subsequently enables pooling and processing of all cells within single reaction and thereby decrease costs for library preparation. Hence, targeted amplification of specific transcript regions for genotyping via PCR on the cDNA would results in loss of the cell-barcode sequence which is located at the cDNA ends. Recently, new techniques have enabled the preservation of cell-barcodes while theoretically enabling genotyping of variants across any transcript position [*146*]. Although this approach may be problematic when genotyping transcripts that are expressed at relatively low levels due to increased drop-outs, it may serve as a very useful tool in future barcoding experiments. For example, when increased resistance towards chemotherapeutic treatment is observed in a fraction of barcoded cells, these cells could be directly linked to a known subclone defined from its mutational markers as defined from bulk sequencing of the sample. This approach would therefore allow to further close the gap between cellular barcodes and the identity of cells which they represent.

One step further, in many cases it may prove advantageous to isolate specific barcoded clones due to their unique behavior observed in experiments. Recently, multiple methods have been described that enable enrichment and isolation of cells carrying specific barcodes from a whole population of barcoded cells. These methods additionally express the cellular barcodes as part of a single-guide RNA (sgRNA). Upon identification of interesting subclones, the barcoded cell population can be lentivirally transduced or transfected with an additional construct. This construct enables a Cas9-mediated activation of expression of additional markers, such as fluorophores or antibiotic resistance. The barcode-sgRNA of the subclone of interest matches a sequence upstream of the additional marker gene, thereby directing a Cas9 to the new marker and inducing its expression. [*147, 148*].

Experimentally, barcoded cell populations are expanded in a first step in order to ensure that every barcode is represented by multiple cells. An aliquot of the starting cell population is frozen and the actual experiment, e.g. exposure of barcoded cells to a selection pressure, is carried out. Interesting subclones that either show increased or decreased cellular fitness, can be identified by the barcode frequencies of the cell populations. The frozen aliquot of the initial barcoded cell population can be thawed, expanded and transduced with the second lentiviral construct that enables the Cas9-mediated selection of specific clones. This approach greatly simplifies the process of isolating phenotypically interesting subclones and hence allows to further investigate their phenotypic, as well as genomic features.

As of now these techniques have only been utilized in cell lines *in vivo* and mostly within *in vitro* cultures [*147-149*]. Their usage for leukemic Pdx lines within the mouse model system might pose additional challenges. However, successful establishment of this approach for the Pdx mouse model would allow for a fast and efficient isolation of interesting subclones and hence enable further targeted analysis of cells showing specific phenotypes, e.g. an increased resistance towards *in vivo* treatment.

5.2 The established smMIP panel allows for sensitive and highly cost-efficient sequencing of hundreds of samples

A smMIP panel targeting more than 16 kb in exons and hotspot regions of 24 recurrent driver genes involved in clonal hematopoiesis and AML has been created and used for processing of over 550 samples for a first cohort study. A re-designed smMIP backbone sequence enables the use of standard sequencing primers and subsequently allows to sequence prepared libraries as spike-ins on flowcells of high-output sequencers like NovaSeq, thereby further decreasing costs. A cost-efficient indexing strategy was set up to facilitate multiplexing of up to 192 libraries per flowcell lane while minimizing the risk of sample cross-contaminations through index swapping.

Compared to the commercial Haloplex assay for which reagents usually cost about 175€ (16,790€ per 96 reactions) [*150*] the established smMIP assay costs only about 7€ per sample (not including initial costs for oligonucleotide probes), while including technical duplicates of prepared libraries. This drop of costs to about 4% compared to the commercial assay allows to process 25 times more samples at the same price, thereby enabling cohort studies with a larger number of individuals that would otherwise be prohibitively expensive without dedicated funding, especially for smaller research groups.

Additionally to the robustly preforming workflow for preparation of sequencing libraries an automated computational analysis pipeline has been established to enable processing of acquired sequencing data for hundreds of samples with little hands-on time. The analysis pipeline combines technical library duplicates for each sample, while preserving the information of which read originated from which replicate. Variant calling is based on raw sequencing reads using *Mutect2* to provide a set of variants detected within the data. These variant calls are subsequently verified by *umivariants*, which utilizes the UMI information included in the data to reassess presence of the variants after computing single-molecule consensus reads. In order to allow for efficient filtering of low frequency technical artefacts, position-specific error rates are calculated across all samples, allowing to calculate the likelihood of an observed variant being derived from background errors.

The results of quality controls carried out using gDNA samples of the cohort study indicate a high sensitivity, reliably detecting variants at 0.7% allele frequencies. Comparisons of variant calls to data derived from the commercial Haloplex sequencing panel indicate good agreement between both methods, despite problems for smMIPs to detect larger insertions and deletions. To test reproducibility of obtained variant calls, library preparations from gDNA of 16 samples were repeated at a later time point. Here, about 60% of variants above 1% allele frequency could be detected in both batches. However, some variants were missing in either one of the variant call sets due to issues in the variant

calling pipeline and differences in sequencing coverage between the batches. Although the set of variants obtained for the cohort study may still contain false-positive variants especially at low variant allele frequencies below 1% after applying the established preliminary filters, the data is absolutely suitable for further analysis. The obtained variant calls from the cohort sequencing are currently being further analyzed and manually curated according to best practices [*151*] by Simon Krauß and Prof. Dr. Klaus Metzeler (Universitätsklinikum Leipzig) and are planned to be published within a separate research article.

The smMIP panel is also further utilized in other projects at the Universitätsklinikum Leipzig, due to its combination of high sensitivity and cost-efficiency. If needed, the panel can be customized to exclude non-informative targets or include new targets by designing and including new smMIPs in the future.

5.2.1 Possible enhancements to the current smMIP assay and library preparation workflow

The smMIP assay has been set up to provide an optimized cost-efficiency. After demonstrating good performance and high-throughput capabilities in a first cohort study the smMIP panel will be further utilized in other projects. Hence, it may prove beneficial to further enhance its performance.

5.2.1.1 Extending the smMIP panel to double-tiling to enable detection of further technical artefacts

Currently an optimized single-tiling strategy was used to target the genomic regions of interest on only one strand, resulting in fewer probes needed to cover these areas and thereby decreasing the initial costs for the panel. However, oxidative damage of the gDNA during its preparation or during the smMIP hybridization reaction may introduce technical artefacts resulting in altered bases after PCR amplification. Most prominently cytosine deamination resulting mutation of a C:G pair to a T:A pair [*152*] or oxidation of Guanine to 8-oxo-Guanine which can pair with Adenine, resulting in mutation of G:C to T:A [*153, 154*]. Although these artifacts are most pronounced when working with formalin-fixed paraffin-embedded (FFPE) samples [*155*] or including shearing of gDNA in the library preparation procedure [*154*], they are also present within fresh clinical samples and contribute to technical noise [*53*]. As usually only one base of a specific base-pair is prone to oxidation the other strand retains the original sequence. When covering both DNA strands throughout the target regions, *ex vivo* gDNA oxidation can be detected by only being present within reads derived from one of both strands thereby enabling detection and removal of oxidative artifacts. Additionally, double-tiling would result in more

probes hybridizing with closer proximity which increases the ability to capture and detect bigger deletions or insertions, like the internal tandem duplication in *FLT3* (FLT3-ITD).

A strong improvement in precision could hence be achieved by designing and ordering new smMIP probes and extend the panel to double-tiling, targeting both DNA strands throughout the complete target regions.

5.2.1.2 Longer sequencing reads to enable coverage of library fragments from forward as well as reverse reads can further decrease technical noise

Another source of technical errors that can impair precision of variant calls at low allele frequencies are introduced by the sequencing itself. It has been shown that specific sequence motifs, e.g. ACGGCGGT, can introduce context-specific errors and result in significantly elevated error rates. However, as the reverse complements of such motifs are usually not inducing additional errors, a reverse read covering this motif would not show increased error rates at this position [*156*]. Overlapping forward and reverse reads of a paired-end sequencing setup can hence be used to identify and subsequently filter these artefacts [*157*]. For sequencing of the smMIP libraries of the first cohort study a 100-nt paired-end sequencing layout was chosen in order to optimize sequencing costs. The smMIPs had been designed with a capture size of 120 nucleotides, which comprises the hybridization arms and gap-fill sequence copied from the gDNA template, as well as a total of 10 degenerate nucleotides that serve as UMI. Consequently the ends of the gap-fill sequence of each smMIP are not covered by both forward and reverse reads and are therefore prone to elevated error rates.

For future studies longer reads, e.g. 120 to 150 nucleotides, may be utilized in order to fully cover the whole captured sequences with the forward and reverse read of each read-pair. Using spike-ins on high-output flow-cells of sequencers like NovaSeq 6000, which provide lower prices per million sequencing reads, can compensate the increased costs for the longer sequencing reads. This simple improvement is already used for current further experiments using the established smMIP panel at the Universitätsklinikum Leipzig.

5.2.1.3 Higher amounts of template gDNA within hybridization reactions can increase the number of unique capture events

For the current cohort sequencing study detection of variants at 1 – 2% allele frequency was considered sufficient, as it is comparable to other commonly used panel sequencing technologies. When higher sensitivities are necessary, multiple potential improvements to the library preparation workflow could be tested.

Increasing the amount of gDNA template within the hybridization reactions can increase the overall number of successful capture events and hence also the unique coverage within the final sequencing data. Although most commonly about 100 ng gDNA template are utilized [*158-162*], it has been shown that higher gDNA input of up to 500 ng per hybridization reaction can significantly increase the yield of successful target-captures [*163*]. Increasing the amount of template utilized for the hybridization reactions can provide a simple way to increase the assays sensitivity. For gDNA samples with a minimal concentration of 25 ng/µl template amounts per hybridization could be doubled from 100 ng to 200 ng with the currently used reaction setup. In order to exclude elevated rates of probe-probe interactions by increased molecular crowding within the reaction that can lead to self-ligated, circularized smMIPs without capturing any genomic sequence, a small pilot experiment would need to be conducted. Importantly, as more UMIs will be present within the library the sequencing depth needs to be increased in order to obtain enough reads per UMI to enable creation of consensus reads. Increasing the assay's sensitivity by utilizing more hybridization products therefore simultaneously increases sequencing costs. Hence, it is necessary to balance sensitivity and costs based on the requirements of the conducted experiments.

5.2.1.3 The usage of a high-fidelity polymerase for the gap-fill within the hybridization reaction can decrease the background error rate

Another major factor within the smMIP hybridization reaction is the gap-fill, i.e. the copying of genomic sequences between the probes' hybridization arms that is carried out by a DNA polymerase.

The established workflow uses the *Hemo KlenTaq* (New England Biolabs) polymerase, which represents a derivative of Taq-Polymerase lacking the first 280 amino acids and having additional internal mutations in order to make it more resistant to PCR inhibitors [*164*] like hemoglobin. This makes the polymerase a good choice when template gDNAs for smMIP hybridizations are isolated from whole blood, as potential impurities do not influence its function. Additionally, the optimal extension

temperature of the *Hemo KlenTag* is at 68°C [*165*], which is lower than the 72°C optimal for most other DNA polymerases used in PCR. Hence, this polymerase is supposed to retain more activity at the hybridization temperature of 60°C. However, newer DNA polymerases engineered for highest fidelity generally show 50 – 300-fold lower error rates than traditional derivatives of the Taq-polymerase according to their manufacturers [*166-168*]. These polymerases have also been shown to reduce error rates within consensus reads when used at the UMI-integration step in a similar methodological approach [*169*], thereby further decreasing background noise within the de-duplicated sequencing data. Consequently, a smMIP panel used to detect minimal residual disease in acute myeloid leukemia samples utilizing a combination of 500 ng template gDNA in addition to Q5 HiFi DNA polymerase for the smMIP hybridization reaction reported detection for variants below 0.1% allele frequency [*170*]. However, it is important to note that about 80 million sequencing reads were needed in order to reach this sensitivity in the mentioned study, again emphasizing the need to balance sensitivity and costs for the study of larger cohorts.

5.2.2 Optimization of the computational analysis pipeline to improve precision of variant calls

To complement the established smMIP panel and its high-throughput library preparation workflow a computational pipeline that enables processing of many samples with minimal hands-on time has been established. The analysis pipeline is completely adapted to the established smMIP workflow and can be started by a single bash script to automatically process hundreds of samples by utilizing the SLURM workload manager on a linux server. The computational pipeline already proved very useful for processing of the sequencing data derived from the first cohort study. However, some issues remain to be resolved and more features could be added in order to further simplify downstream processing of obtained variant call sets, as well as comparisons between samples and improve robustness of variant calls.

5.2.2.1 Integration of a second variant caller to increase sensitivity and prioritize variants

The smMIP quality control experiments indicated good reproducibility for variant calls for allele frequencies of at least 1% when comparing independent library preparations within different batches for the same samples. However, one variant at 10% allele frequency was not consistently called in both datasets, although the mutations were clearly present within the mapped reads. This indicates that the lack of detection was caused by computational problems at the variant calling step using *Mutect2* within

the established analysis pipeline. Further testing is needed in order to determine the reason why this variant was not correctly called and to adjust the parameters used for variant calling accordingly.

It has been suggested that multiple variant callers should be used in order to increase the reliability for detecting mutations within analysis pipelines used in clinical settings [*151*]. Hence, integration of a second variant caller into the pipeline would be a second option to decrease the false negative rate of mutation calls. Variant calling by *Mutect2* which detects variants based on local re-alignment of mapped reads could be complemented by a pile-up based variant caller. Pile-up based variant callers, such as the commonly used *VarScan 2* [*171*], solely rely on mismatches between mapped sequencing reads and the reference genome and should therefore reliably detect the variant at 10% VAF described above. Call sets from both variant callers could be evaluated based on UMI consensus sequences and afterwards merged into one final call set. Information about detection of particular variants by only one or both variant callers may additionally help to prioritize and classify high confidence variants, in order to simplify the manual curation of detected variants.

5.2.2.2 Improving the final filtering of variant calls to balance sensitivity and precision

For quality control repeated library preparations and sequencings for 16 samples were carried out. One variant at about 2% was part of the variant set of only one batch, as it was detected with five supporting UMIs at a unique sequencing depth of 199x. However, the variant was also present within the second batch but supported only by one UMI at a unique coverage of 107x. The current filtering of variants includes a threshold of at least three UMIs that are necessary to support a variant in order to exclude false-positive variant calls derived from technical noise. Hence, this variant was filtered from the final set of variants in this batch.

This emphasizes that the currently established set of filters used for final processing of the obtained variant call sets from the computational pipeline may be not optimal yet. In order to retain more sensitivity, especially in regions with low unique coverage, the number of UMIs needed to support a given variant could be based on the observed coverage at this position.

Fortunately, 77 control samples derived from hip surgeries, have meanwhile been utilized in another study in order to analyze the prevalence of clonal hematopoiesis within these patients using the Haloplex assay established at the Klinikum der Universität München [*172*]. The filtering of variants detected by the smMIP assay could therefore be optimized using the Haloplex data of these samples in order to increase precision for variant calling. As the sensitivity of the Haloplex panel is limited to 1-2% allele frequency, depending on sequencing coverage, this optimization cannot be carried for variants at very

low frequencies below 1%. Nevertheless, the Haloplex data enables further comparison of variant calls in order to increase confidence in variants detected by the smMIP assay and refinement of variant filtration, especially in regions with lower unique sequencing coverage.

5.2.2.3 Summarizing the obtained unique sequencing depths to identify limits for reliable detection of variants

The re-sequencing of patient samples as quality control to assess reproducibility of variant calls showed that some variants are not consistently called within both data sets. This discrepancy was caused by differing unique sequencing coverages between the sequencing batches. A lower sequencing coverage resulted in variants not reaching the currently used thresholds of three variant-supporting UMIs for the final filtration of variants.

In addition to the refinement of the filtering thresholds, as described above, summarizing the obtained unique sequencing depth across all target regions within a given sample should be integrated as an additional output for the analysis pipeline. This data could be used to calculate the limit of detection for variants across the target areas for each sample given the applied filtering criteria. A mapped BAM file with all read-pairs, de-duplicated based on their UMI information, is already part of the computational pipeline's output. The unique coverage can therefore easily be obtained by integrating e.g. *samtools* to summarize the sequencing depth in target regions within the de-duplicated reads. This information about unique coverage can subsequently be used to estimate the sensitivity to detect variants across the targeted genomic regions, subsequently allowing a better direct comparison of detected variants between samples.

5.2.2.3 Summary statistics about performance of individual probes can help to identify bad performing probes and potential undetected variants

The established computational pipeline currently focuses on pre-processing of sequencing data as well as variant calling and annotation. However, the current output does not include summaries about the performance of each individual smMIP.

Currently, the trimming of hybridization arm sequences from read-pairs produces are log file that contains information about how often which sequences were successfully trimmed. Extracting this information allows to identify the number of raw sequencing reads per probe that were detected within

the sample. In addition, the number of raw read-pairs and UMI-consensus reads could be derived from the mapped bam files produced as output of the current analysis pipeline. The expected read-start positions after mapping are specific for every smMIP and could therefore be utilized to generate information about the total and unique on-target reads for each probe.

The raw reads per probe, detected upon initial trimming of hybridization arms from sequencing reads, would allow to also use the computational pipeline for rebalancing experiments when new probes are added to the pool, e.g. when including additional genomic targets for re-genotyping. In combination with the reads detected at the expected mapping position for a specific probe, these data would allow to easily assess the relative performance of each smMIP within the panel, as well as to identify probes that show elevated rates of off-target mappings, either due to unambiguous mapping or off-target capture events within the hybridization reaction.

Additionally, hybridization efficiencies per smMIP could be summarized for all processed samples. If specific probes perform significantly worse within a specific sample, this could indicate the presence of undetected variants, e.g. caused by deletions within the genomic regions used to hybridize the smMIPs as observed for *CALR* (see 4.2.5.2), which could be further analyzed by additional re-genotyping methods if necessary. These statistics could therefore also contribute to a better characterization of samples that were sequenced using the smMIP panel.

5.2.2.4 Comparison of the established computational pipeline to recently published smMIP analysis tools

As smMIPs gained popularity within the last years due to their customizability, cost-efficiency, scalability and sensitivity to detect low frequency variants a few specialized analysis pipelines have been published recently. These pipelines allow for automated processing of smMIP data by including trimming of hybridization arms and de-duplication of reads based on UMIs into the pipelines [*173, 174*]. However, these pipelines are either not specifically intended to identify low-frequency variants [*173*] or classify the calling of these variants as experimental [*174*]. Interestingly, both of these pipelines rely on the use of the well-established GATK suite or even use *Mutect2* to call low-frequency variants within the sequencing data, thereby confirming the choice of *Mutect2* as a suitable variant caller.

Even more recently *smMIP-tools*, an analysis pipeline specifically designed for detection of low-frequency variants using smMIPs has been published, reporting robust detection of variants down to 0.5% allele frequency [*175*]. The increased precision is achieved utilizing additional error-suppression by computing error rates to calculate the likelihoods for a detected variant being caused by technical

background noise, similarly to the approach implemented in the established pipeline. However, error rates are computed allele-specific, i.e. for all possible transversions and transitions at each position, which allows for even more specific error profile filtering. This refinement could easily be implemented into the currently established computational pipeline and could further enhance classification of background errors used for variant filtration.

Additionally UMI-singletons that consist of single reads, as well as UMI-consensus reads calculated for UMIs that were observed in multiple reads are both utilized for pile-up based variant calling. Variants detected only in UMI-singletons but not in UMI-consensus sequences are separately output in a low-confidence variant call set. Currently, UMIs detected with only one read-pair are being ignored when evaluating variants detected by *Mutect2* using the *umivariants* package in the established pipeline. However, despite the missing possibility for error-correction the utilization of these 'singleton-UMIs' can contribute to increase sensitivity especially in regions with low sequencing depth.

Furthermore, raw sequencing reads are mapped to the reference genome without trimming of hybridization arms beforehand within the *smMIP-tools* workflow. Instead, sequencing reads are assigned to the smMIP they originated from based on their mapping positions and the respective arm sequences are ignored when performing the pile-up based variant calling. This approach has the advantage of minimizing 'edge-effects' known from amplicon sequencing data. Here, SNVs or InDels near the end of the sequencing reads can lead to softclipping of bases, which are usually ignored for downstream processing and variant calling [*176, 177*]. As pile-up based variant calling is solely based on mismatches to the reference within mapped reads, this approach ensures that reads are correctly mapped in order to detect variants located next to the hybridization arms. In contrast, the variant calling using *Mutect2* within the newly established computational pipeline relies on local-assembly of sequencing reads that also utilizes soft-clipped bases in order to maximize sensitivity. Hence, trimming of hybridization arm sequences from sequencing reads is not expected to significantly affect variant calling for the currently established analysis pipeline. However, if additional pile-up based variant callers are going to be integrated into the current pipeline (see 5.2.2.1) it might prove beneficial to remove the sequences derived from the smMIP-arms only after mapping to the reference.

6 Conclusion and Outlook

For analysis of cell compositions within AML Pdx models, two barcode libraries were cloned and successfully used within first pilot experiments. Here, the barcoding approach not only allowed to observe differential response of subclones towards *in vivo* chemotherapeutic treatment but also to save laboratory mice and costs by substituting classical LDTA assays using the quantitative barcode read-out for the determination of LIC frequencies. Furthermore, the barcoding assay was used to estimate the bottleneck for engraftment of leukemic cells in order to estimate whether subclonal heterogeneity can be maintained throughout serial passages for a specific AML Pdx line, which represents an important information for the use of the mouse model. Additionally, subclones derived from single-cells of a single AML patient's samples could be isolated allowing for detailed analysis of genomic features as well as *in vivo* characteristics of subclones. This study has recently been published in the Journal of Hematology & Oncology. For future experiments a new barcoding construct has been cloned that contains a high-complexity expressed barcode, which combines advantages of both previous constructs and thus can be used as a universal tool for all future barcoding experiments.

The established cellular barcoding assays already proved to be a powerful tool for investigation of subclonal heterogeneity within AML Pdx lines. One potential long term perspective lies in the isolation of more treatment resistant subclones from other AML Pdx lines in order to create a library of genotypically and phenotypically distinct adverse subclones. The subclones could subsequently be mixed and used to investigate their responses towards new therapeutics, allowing to further link effectiveness of new treatment options to known genotypes.

As a complementary approach enabling the investigation of subclonal heterogeneity within patient samples a highly-sensitive smMIP panel for targeted re-genotyping of recurrent AML and CHIP driver genes was established. The published approach was enhanced by utilization of a custom multiplexing strategy allowing for sequencing of up to 192 samples on one flowcell lane. Futhermore the probe-backbone was re-designed to enable sequencing of samples as spike-ins on big flowcells, thereby decreasing sequencing costs for carrying out screening of large cohorts. In combination with reagents costs for library preparations being reduced about 25-fold compared to the commercial Haloplex assay, which is being utilized for similar studies at the Klinikum der Universität München, the new assay provides a highly cost efficient and sensitive way for targeted re-genotyping of AML and CH driver genes. Although some smaller issues remain to be resolved the established smMIP assay demonstrated a robust workflow capable of high-throughput processing of hundreds of samples with minimized hands-on time, both on the wet lab as well as the computational side, in combination with a generally high

sensitivity that enables detection of subclonal variants below 1% variant allele frequency. The variant calls acquired from the cohort study of AML patients in long-term remission are currently being analyzed and planned to be published within a separate research article.

Due to the robust performance, high sensitivity and cost-efficiency demonstrated within the first sequencing project the smMIP panel is now utilized in additional projects and will be developed further by Simon Krauss at the Universitätsklinikum Leipzig within the group of Prof. Dr. Klaus Metzeler. Its high cost-efficiency will allow to increase cohort sizes in future sequencing studies and thereby contribute to new findings in the field of AML and clonal hematopoiesis.

In summary, two complementary approaches for analysis of subclonal heterogeneity were successfully established and will continually be used for further studies by the respective collaboration partners Prof. Dr. Irmela Jeremias and Prof. Dr. Klaus Metzeler.

7 Publications

- "**Adverse stem cell clones within a single patient's tumor predict clinical outcome in AML patients"**. Zeller C*, **Richter D***, Jurinovic V, Valtierra-Gutiérrez I A, Jayavelu A K, Mann M, Bagnoli J W, Hellmann I, Herold T, Enard W, Vick B, Jeremias I. Journal of Hematology & Oncology 15, 25 (2022).
  * contributed equally
- "**In vivo PDX CRISPR/Cas9 screens reveal mutual therapeutic targets to overcome heterogeneous acquired chemo-resistance**". Anna-Katharina Wirth, Lucas Wange, Sebastian Vosberg, Erbey Özdemir, Christina Zeller, **Daniel Richter**, Daniela Senft, Ehsan Bahrami, Ashok Kumar Jayavelu, Wolfgang Enard, Tobias Herold, Irmela Jeremias
  (*in revision, Leukemia*)
- "**Prime-seq, efficient and powerful bulk RNA-sequencing**", Aleksandar Janjic & Lucas E. Wange, Johannes W. Bagnoli, Johanna Geuder, Phong Nguyen, **Daniel Richter**, Beate Vieth, Binje Vick, Irmela Jeremias, Christoph Ziegenhain, Ines Hellmann, Wolfgang Enard. Genome biology 23, 88 (2022).
- **"Regulatory and coding sequences of TRNP1 co-evolve with cortical folding in mammals"**. Zane Kliesmete, Lucas Esteban Wange, Beate Vieth, Miriam Esgleas, Jessica Radmer, Matthias Hülsmann, Johanna Geuder, **Daniel Richter**, Mari Ohnuki, Magdalena Götz, Ines Hellmann, Wolfgang Enard
  (*https://www.biorxiv.org/content/10.1101/2021.02.05.429919v2 - manuscript in preparation*)
- **"T cell-expressed microRNAs critically regulate germinal center T follicular helper cell function and maintenance in acute viral infection in mice"**. Zeiträg J, Dahlström F, Chang Y, Alterauge D, **Richter D**, Niemietz J, Baumjohann D. Eur J Immunol. 2020 Sep 30.

8 References

1. Hyuna Sung *et al.*, Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians* **71**, 209-249 (2021).
2. Seyed Hossein Hassanpour, Mohammadamin Dehghani, Review of cancer from perspective of molecular. *Journal of Cancer Research and Practice* **4**, 127-129 (2017).
3. Douglas Hanahan, Robert A. Weinberg, The Hallmarks of Cancer. *Cell* **100**, 57-70 (2000).
4. Douglas Hanahan, Robert A Weinberg, Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646-674 (2011).
5. Michael R. Stratton, Peter J. Campbell, P. Andrew Futreal, The cancer genome. *Nature* **458**, 719-724 (2009).
6. P. C. Nowell, The clonal evolution of tumor cell populations. *Science* **194**, 23-28 (1976).
7. Lauren M. F. Merlo, John W. Pepper, Brian J. Reid, Carlo C. Maley, Cancer as an evolutionary and ecological process. *Nature Reviews Cancer* **6**, 924-935 (2006).
8. Marco Gerlinger *et al.*, Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nature Genetics* **46**, 225-233 (2014).
9. Jia-Jie Hao *et al.*, Spatial intratumoral heterogeneity and temporal clonal evolution in esophageal squamous cell carcinoma. *Nature Genetics* **48**, 1500-1507 (2016).
10. Anna Schuh *et al.*, Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood* **120**, 4191-4196 (2012).
11. Marco Gerlinger *et al.*, Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *New England Journal of Medicine* **366**, 883-892 (2012).
12. Peng-Chan Lin *et al.*, Intratumor Heterogeneity of MYO18A and FBXW7 Variants Impact the Clinical Outcome of Stage III Colorectal Cancer. *Frontiers in Oncology* **10**, (2020).
13. L. Ding *et al.*, Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481**, 506-510 (2012).
14. Luc G. T. Morris *et al.*, Pan-cancer analysis of intratumor heterogeneity as a prognostic determinant of survival. *Oncotarget* **7**, 10051-10063 (2016).
15. Rebecca A. Burrell, Charles Swanton, Tumour heterogeneity and the evolution of polyclonal drug resistance. *Molecular Oncology* **8**, 1095-1111 (2014).
16. L. I. Shlush, S. Zandi, S. Itzkovitz, A. C. Schuh, Aging, clonal hematopoiesis and preleukemia: not just bad luck? *International journal of hematology* **102**, 513-522 (2015).
17. D. C. Link, M. J. Walter, 'CHIP'ping away at clonal hematopoiesis. *Leukemia* **30**, 1633-1635 (2016).
18. Cathy C. Laurie *et al.*, Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nature Genetics* **44**, 642-650 (2012).
19. S. Jaiswal *et al.*, Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. *The New England journal of medicine* **377**, 111-121 (2017).
20. D. P. Steensma *et al.*, Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. *Blood* **126**, 9-16 (2015).
21. M. Xie *et al.*, Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nature medicine* **20**, 1472-1478 (2014).
22. G. Genovese *et al.*, Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *The New England journal of medicine* **371**, 2477-2487 (2014).
23. J. S. Welch *et al.*, The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150**, 264-278 (2012).
24. L. Hartmann, K. H. Metzeler, Clonal hematopoiesis and pre-leukemia -genetics, biology and clinical implications. *Genes, chromosomes & cancer*, (2019).

25. Rebecca L. Siegel, Kimberly D. Miller, Hannah E. Fuchs, Ahmedin Jemal, Cancer Statistics, 2021. *CA: A Cancer Journal for Clinicians* **71**, 7-33 (2021).
26. Rory M. Shallis, Rong Wang, Amy Davidoff, Xiaomei Ma, Amer M. Zeidan, Epidemiology of acute myeloid leukemia: Recent progress and enduring challenges. *Blood Reviews* **36**, 70-87 (2019).
27. I. De Kouchkovsky, M. Abdul-Hay, 'Acute myeloid leukemia: a comprehensive review and 2016 update'. *Blood Cancer Journal* **6**, e441-e441 (2016).
28. K. H. Metzeler *et al.*, Spectrum and prognostic relevance of driver gene mutations in acute myeloid leukemia. *Blood* **128**, 686-698 (2016).
29. Jeffery M Klco *et al.*, Functional Heterogeneity of Genetically Defined Subclones in Acute Myeloid Leukemia. *Cancer Cell* **25**, 379-392 (2014).
30. Xiaoli Wu *et al.*, Prognostic significance of FLT3-ITD in pediatric acute myeloid leukemia: a meta-analysis of cohort studies. *Molecular and Cellular Biochemistry* **420**, 121-128 (2016).
31. S. P. Whitman *et al.*, FLT3 internal tandem duplication associates with adverse outcome and gene- and microRNA-expression signatures in patients 60 years of age or older with primary cytogenetically normal acute myeloid leukemia: a Cancer and Leukemia Group B study. *Blood* **116**, 3622-3626 (2010).
32. Alexander E. Perl, The role of targeted therapy in the management of patients with AML. *Hematology. American Society of Hematology. Education Program* **2017**, 54-65 (2017).
33. National Cancer Institute, https://seer.cancer.gov/statfacts/html/amyl.html, Last accessed: March 2022
34. Himisha Beltran *et al.*, Divergent clonal evolution of castration-resistant neuroendocrine prostate cancer. *Nature medicine* **22**, 298-305 (2016).
35. Mariam Jamal-Hanjani *et al.*, Tracking the Evolution of Non–Small-Cell Lung Cancer. *New England Journal of Medicine* **376**, 2109-2121 (2017).
36. Dan A Landau *et al.*, Evolution and Impact of Subclonal Mutations in Chronic Lymphocytic Leukemia. *Cell* **152**, 714-726 (2013).
37. Juhi Ojha *et al.*, Deep sequencing identifies genetic heterogeneity and recurrent convergent evolution in chronic lymphocytic leukemia. *Blood* **125**, 492-498 (2015).
38. Matthew K. H. Hong *et al.*, Tracking the origins and drivers of subclonal metastatic expansion in prostate cancer. *Nature communications* **6**, 6605 (2015).
39. Norio Shiba *et al.*, Whole-exome sequencing reveals the spectrum of gene mutations and the clonal evolution patterns in paediatric acute myeloid leukaemia. *British journal of haematology* **175**, 476-489 (2016).
40. Hideki Makishima *et al.*, Dynamics of clonal evolution in myelodysplastic syndromes. *Nature Genetics* **49**, 204-212 (2017).
41. Heidi L. Rehm, Disease-targeted sequencing: a cornerstone in the clinic. *Nature Reviews Genetics* **14**, 295-300 (2013).
42. Rasmus Nielsen, Joshua S. Paul, Anders Albrechtsen, Yun S. Song, Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics* **12**, 443-451 (2011).
43. A. L. Young *et al.*, Quantifying ultra-rare pre-leukemic clones via targeted error-corrected sequencing. *Leukemia* **29**, 1608-1611 (2015).
44. J. Jee *et al.*, Rates and mechanisms of bacterial mutagenesis from maximum-depth sequencing. *Nature* **534**, 693-696 (2016).
45. M. W. Schmitt *et al.*, Sequencing small genomic targets with high efficiency and extreme accuracy. *Nature methods* **12**, 423-425 (2015).
46. I. Kinde, J. Wu, N. Papadopoulos, K. W. Kinzler, B. Vogelstein, Detection and quantification of rare mutations with massively parallel sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 9530-9535 (2011).
47. E. M. Elnifro, A. M. Ashshi, R. J. Cooper, P. E. Klapper, Multiplex PCR: optimization and application in diagnostic virology. *Clin Microbiol Rev* **13**, 559-570 (2000).

48. Illumina Inc., https://support.illumina.com/bulletins/2020/04/maximum-read-length-for-illumina-sequencing-platforms.html, Last accessed: December 2021
49. Donavan T. Cheng *et al.*, Detection of Mutations in Myeloid Malignancies through Paired-Sample Analysis of Microdroplet-PCR Deep Sequencing Data. *The Journal of Molecular Diagnostics* **16**, 504-518 (2014).
50. Florian Mertes *et al.*, Targeted enrichment of genomic DNA regions for next-generation sequencing. *Briefings in Functional Genomics* **10**, 374-386 (2011).
51. Iwanka Kozarewa, Javier Armisen, Andrew F. Gardner, Barton E. Slatko, C.L. Hendrickson, Overview of Target Enrichment Strategies. *Current Protocols in Molecular Biology* **112**, 7.21.21-27.21.23 (2015).
52. M. Nilsson *et al.*, Padlock probes: circularizing oligonucleotides for localized DNA detection. *Science* **265**, 2085-2088 (1994).
53. J. B. Hiatt, C. C. Pritchard, S. J. Salipante, B. J. O'Roak, J. Shendure, Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome research* **23**, 843-854 (2013).
54. Emily H. Turner, Choli Lee, Sarah B. Ng, Deborah A. Nickerson, Jay Shendure, Massively parallel exon capture and library-free resequencing across 16 genomes. *Nature methods* **6**, 315-316 (2009).
55. B. J. O'Roak *et al.*, Recurrent de novo mutations implicate novel genes underlying simplex autism risk. *Nature communications* **5**, 5595 (2014).
56. P. Hardenbol *et al.*, Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nature biotechnology* **21**, 673-678 (2003).
57. B. J. O'Roak *et al.*, Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338**, 1619-1622 (2012).
58. Madeleine R. Geisheker *et al.*, Hotspots of missense mutation identify neurodevelopmental disorder genes and functional domains. *Nature Neuroscience* **20**, 1043-1051 (2017).
59. Manon S. Oud *et al.*, Validation and application of a novel integrated genetic screening method to a cohort of 1,112 men with idiopathic azoospermia or severe oligozoospermia. *Human Mutation* **38**, 1592-1605 (2017).
60. H. P. Koeffler, D. W. Golde, Human Myeloid Leukemia Cell Lines: A Review. *Blood* **56**, 344-350 (1980).
61. Y. Matsuo *et al.*, Two acute monocytic leukemia (AML-M5a) cell lines (MOLM-13 and MOLM-14) with interclonal phenotypic heterogeneity showing MLL-AF9 fusion resulting from an occult chromosome insertion, ins(11;9)(q23;p22p23). *Leukemia* **11**, 1469-1477 (1997).
62. V. C. Daniel *et al.*, A primary xenograft model of small-cell lung cancer reveals irreversible changes in gene expression imposed by culture in vitro. *Cancer research* **69**, 3364-3373 (2009).
63. Christina Krupka *et al.*, CD33 target validation and sustained depletion of AML blasts in long-term cultures by the bispecific T-cell–engaging antibody AMG 330. *Blood* **123**, 356-365 (2014).
64. Dong Lin *et al.*, High Fidelity Patient-Derived Xenografts for Accelerating Prostate Cancer Discovery and Drug Development. *Cancer research* **74**, 1272-1283 (2014).
65. Sarah Ebinger *et al.*, Characterization of Rare, Dormant, and Therapy-Resistant Cells in Acute Lymphoblastic Leukemia. *Cancer Cell* **30**, 849-862 (2016).
66. B. Vick *et al.*, An advanced preclinical mouse model for acute myeloid leukemia using patients' cells of various genetic subgroups and in vivo bioluminescence imaging. *PloS one* **10**, e0120925 (2015).
67. Christopher Walsh, Constance L. Cepko, Widespread Dispersion of Neuronal Clones Across Functional Regions of the Cerebral Cortex. *Science* **255**, 434-440 (1992).
68. Rong Lu, Norma F. Neff, Stephen R. Quake, Irving L. Weissman, Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. *Nature biotechnology* **29**, 928-933 (2011).
69. S. H. Naik *et al.*, Diverse and heritable lineage imprinting of early haematopoietic progenitors. *Nature* **496**, 229-232 (2013).

70. Daniela B. Zanatta *et al.*, Genetic barcode sequencing for screening altered population dynamics of hematopoietic stem cells transduced with lentivirus. *Molecular Therapy — Methods & Clinical Development* **1**, 14052 (2014).
71. L. V. Nguyen *et al.*, DNA barcoding reveals diverse growth kinetics of human breast tumour subclones in serially passaged xenografts. *Nature communications* **5**, 9 (2014).
72. L. V. Nguyen *et al.*, Barcoding reveals complex clonal dynamics of de novo transformed human mammary cells. *Nature* **528**, 267-271 (2015).
73. H. E. Bhang *et al.*, Studying clonal dynamics in response to cancer therapy using high-complexity barcoding. *Nature medicine* **21**, 440-448 (2015).
74. O. Nolan-Stevaux *et al.*, Measurement of Cancer Cell Growth Heterogeneity through Lentiviral Barcoding Identifies Clonal Dominance as a Characteristic of In Vivo Tumor Engraftment. *PloS one* **8**, e67316 (2013).
75. K. Cornils *et al.*, Clonal competition in BcrAbl-driven leukemia: how transplantations can accelerate clonal conversion. *Mol Cancer* **16**, 120 (2017).
76. D. Pal *et al.*, Long-term in vitro maintenance of clonal abundance and leukaemia-initiating potential in acute lymphoblastic leukaemia. *Leukemia* **30**, 1691-1700 (2016).
77. A. Elder *et al.*, Abundant and equipotent founder cells establish and maintain acute lymphoblastic leukaemia. *Leukemia* **31**, 2577-2586 (2017).
78. M. E. Belderbos *et al.*, Clonal selection and asymmetric distribution of human leukemia in murine xenografts revealed by cellular barcoding. *Blood* **129**, 3210-3220 (2017).
79. K. Klauke *et al.*, Tracing dynamics and clonal heterogeneity of Cbx7-induced leukemic stem cells by cellular barcoding. *Stem cell reports* **4**, 74-89 (2015).
80. N. Rohland, D. Reich, Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome research* **22**, 939-946 (2012).
81. Evan A. Boyle, Brian J. O'Roak, Beth K. Martin, Akash Kumar, Jay Shendure, MIPgen: optimized modeling and design of molecular inversion probes for targeted resequencing. *Bioinformatics (Oxford, England)* **30**, 2670-2672 (2014).
82. INTERNATIONAL HAPMAP PROJECT GM18505 B-Lymphocyte cell line, Human Genetic Cell Repository at the Coriell Institute for Medical Research, https://www.coriell.org/0/Sections/Search/Sample_Detail.aspx?Ref=GM18505&Product=CC, Last accessed: October 2021
83. Integrated DNA Technologies Inc., https://sfvideo.blob.core.windows.net/sitefinity/docs/default-source/supplementary-product-info/xgen-dual-index-umi-adapter-barcode-sequences.xlsx?sfvrsn=9a3d0c07_6, Last accessed: October 2021
84. Integrated DNA Technologies Inc., https://eu.idtdna.com/pages/products/next-generation-sequencing/adapters/xgen-dual-index-umi-adapters-tech-access/unique-dual-indexed-sequencing-adapters-with-umis-effectively-eliminate-index-cross-talk-and-significantly-improve-sensitivity-of-massively-parallel-sequencing, Last accessed: Novermber 2021
85. M. Costello *et al.*, Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC genomics* **19**, 332 (2018).
86. M. Zuker, Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic acids research* **31**, 3406-3415 (2003).
87. G. Renaud, U. Stenzel, T. Maricic, V. Wiebe, J. Kelso, deML: robust demultiplexing of Illumina sequences using a likelihood-based approach. *Bioinformatics* **31**, 770-772 (2015).
88. Marcel Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011* **17**, 3 (2011).
89. S. T. Sherry, M. Ward, K. Sirotkin, dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome research* **9**, 677-679 (1999).
90. K. Wang, M. Li, H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38**, e164 (2010).

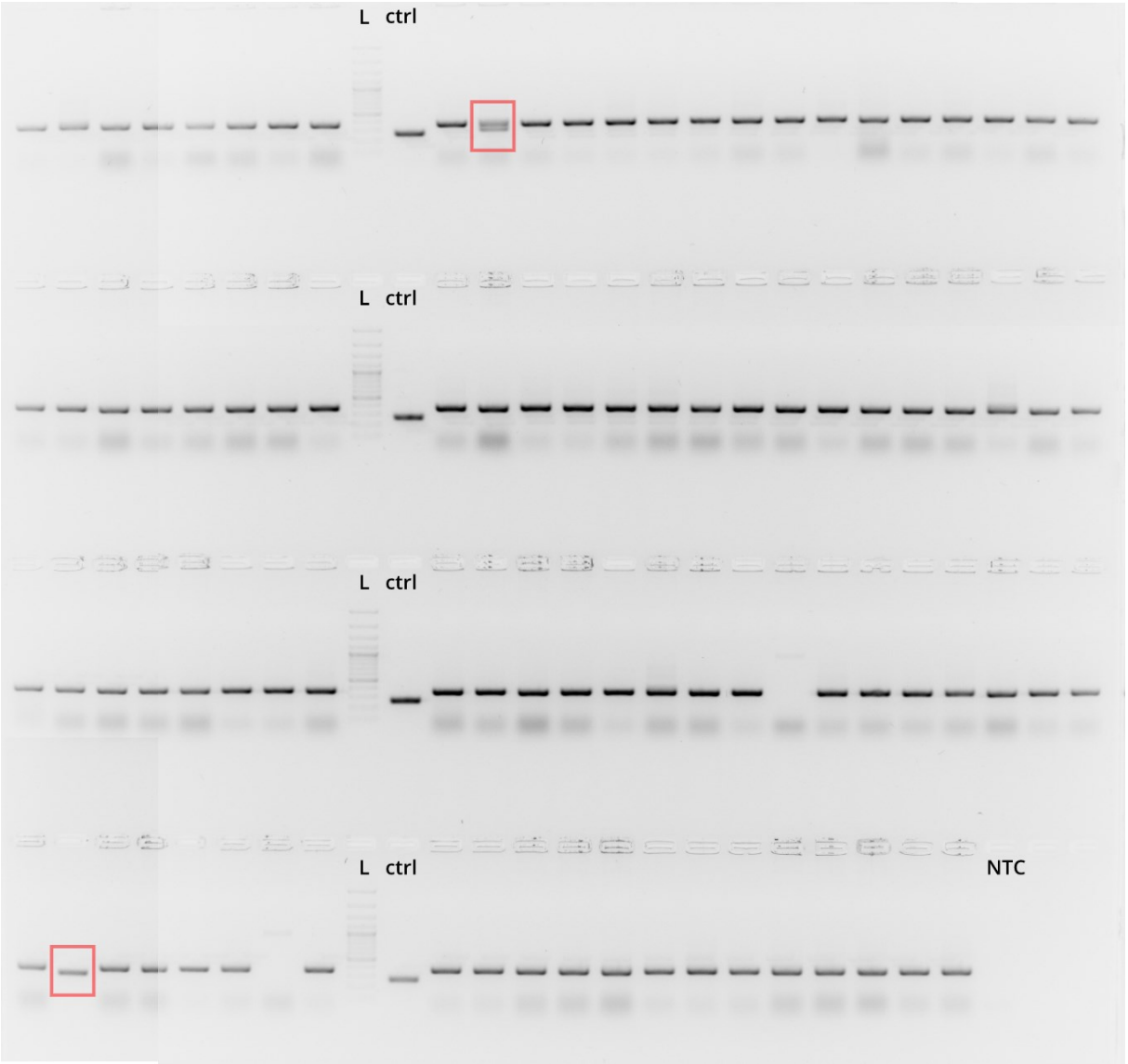
91. N. A. O'Leary *et al.*, Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research* **44**, D733-745 (2016).
92. Shifu Chen *et al.*, Gencore: an efficient tool to generate consensus reads for error suppressing and duplicate removing of NGS data. *BMC bioinformatics* **20**, 606 (2019).
93. Ilse Ariadna Valtierra Gutiérrez, Inferring the Clonal Identity of Single Cells from RNA-seq Data with Unique Molecular Identifiers, Ludwig Maximilians Universität München, München (2020), (10.5282/edoc.28558).
94. James T. Robinson *et al.*, Integrative genomics viewer. *Nature biotechnology* **29**, 24-26 (2011).
95. S. Ebinger *et al.*, Characterization of Rare, Dormant, and Therapy-Resistant Cells in Acute Lymphoblastic Leukemia. *Cancer Cell* **30**, 849-862 (2016).
96. N. Terziyska *et al.*, In vivo imaging enables high resolution preclinical trials on patients' leukemia cells growing in mice. *PloS one* **7**, e52798 (2012).
97. A. Stahlberg *et al.*, Simple multiplexed PCR-based barcoding of DNA for ultrasensitive mutation detection by next-generation sequencing. *Nature protocols* **12**, 664-682 (2017).
98. Lu Zhao, Zhimin Liu, Sasha F Levy, Song Wu, Bartender: a fast and accurate clustering algorithm to count barcode reads. *Bioinformatics* **34**, 739-747 (2017).
99. R Core Team. (R Foundation for Statistical Computing, 2015).
100. Hadley Wickham *et al.*, Welcome to the Tidyverse. *Journal of open source software* **4**, 1686 (2019).
101. Hadley Wickham, *ggplot2: elegant graphics for data analysis*. (Springer New York, 2009).
102. Jake R Conway, Alexander Lex, Nils Gehlenborg, UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938-2940 (2017).
103. Christopher A. Miller *et al.*, Visualizing tumor evolution with the fishplot package for R. *BMC genomics* **17**, 880 (2016).
104. Y. Akimov, D. Bulanova, S. Timonen, K. Wennerberg, T. Aittokallio, Improved detection of differentially represented DNA barcodes for high-throughput clonal phenomics. *Mol Syst Biol* **16**, e9195 (2020).
105. E. Zorita, P. Cusco, G. J. Filion, Starcode: sequence clustering based on all-pairs search. *Bioinformatics* **31**, 1913-1919 (2015).
106. Chrstina Maria Zeller, Single Stem Cell Clones of an Acute Myeloid Leukaemia Patient Display Functional Heterogeneity In Vivo, Monography, Ludwig-Maximilians University, (2020),
107. Y. Hu, G. K. Smyth, ELDA: extreme limiting dilution analysis for comparing depleted and enriched populations in stem cell and other assays. *Journal of immunological methods* **347**, 70-78 (2009).
108. L. Thielecke *et al.*, Limitations and challenges of genetic barcode quantification. *Scientific reports* **7**, 43249 (2017).
109. S. Han, D. Kim, AtRTPrimer: database for Arabidopsis genome-wide homogeneous and specific RT-PCR primer-pairs. *BMC bioinformatics* **7**, 179 (2006).
110. L. V. Bystrykh, M. E. Belderbos, Clonal Analysis of Cells with Cellular Barcoding: When Numbers and Sizes Matter. *Methods in molecular biology* **1516**, 57-89 (2016).
111. G. Sezonov, D. Joseleau-Petit, R. D'Ari, Escherichia coli physiology in Luria-Bertani broth. *J Bacteriol* **189**, 8746-8749 (2007).
112. S. Ebinger *et al.*, Plasticity in growth behavior of patients' acute myeloid leukemia stem cells growing in mice. *Haematologica* **105**, 2855-2860 (2020).
113. L. V. Bystrykh, G. de Haan, E. Verovskaya, Barcoded vector libraries and retroviral or lentiviral barcoding of hematopoietic stem cells. *Methods in molecular biology* **1185**, 345-360 (2014).
114. C. T. Deakin *et al.*, Impact of next-generation sequencing error on analysis of barcoded plasmid libraries of known complexity and sequence. *Nucleic acids research* **42**, e129 (2014).
115. Tsvee Lapidot *et al.*, A cell initiating human acute myeloid leukaemia after transplantation into SCID mice. *Nature* **367**, 645-648 (1994).
116. J. M. Gerber *et al.*, A clinically relevant population of leukemic CD34(+)CD38(-) cells in acute myeloid leukemia. *Blood* **119**, 3571-3577 (2012).

117. N. Goardon *et al.*, Coexistence of LMPP-like and GMP-like leukemia stem cells in acute myeloid leukemia. *Cancer Cell* **19**, 138-152 (2011).
118. T. Lapidot *et al.*, A cell initiating human acute myeloid leukaemia after transplantation into SCID mice. *Nature* **367**, 645-648 (1994).
119. J. E. Sarry *et al.*, Human acute myelogenous leukemia stem cells are rare and heterogeneous when assayed in NOD/SCID/IL2Rγc-deficient mice. *The Journal of clinical investigation* **121**, 384-395 (2011).
120. Christina Zeller *et al.*, Adverse stem cell clones within a single patient's tumor predict clinical outcome in AML patients. *Journal of Hematology & Oncology* **15**, 25 (2022).
121. Illumina Inc., https://emea.support.illumina.com/downloads/illumina-adapter-sequences-document-1000000002694.html, Last accessed: October 2021
122. Aleksandar Janjic *et al.*, Prime-seq, efficient and powerful bulk RNA-sequencing. *bioRxiv*, 2021.2009.2027.459575 (2021).
123. Johannes W. Bagnoli *et al.*, Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. *Nature communications* **9**, 2937 (2018).
124. P. H. Sudmant *et al.*, An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75-81 (2015).
125. R. L. Levine, P. J. M. Valk, Next-generation sequencing in the diagnosis and minimal residual disease assessment of acute myeloid leukemia. *Haematologica* **104**, 868-871 (2019).
126. Illumina Inc, https://emea.illumina.com/systems/sequencing-platforms/hiseq-2500/specifications.html, Last accessed: 11/2021
127. M. Kircher, S. Sawyer, M. Meyer, Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic acids research* **40**, e3 (2012).
128. Rahul Sinha *et al.*, Index switching causes "spreading-of-signal" among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. *bioRxiv*, 125724 (2017).
129. J. A. Griffiths, A. C. Richard, K. Bach, A. T. L. Lun, J. C. Marioni, Detection and removal of barcode swapping in single-cell RNA-seq data. *Nature communications* **9**, 2667 (2018).
130. D. Vodak *et al.*, Sample-Index Misassignment Impacts Tumour Exome Sequencing. *Scientific reports* **8**, 5307 (2018).
131. Andy B. Yoo, Morris A. Jette, Mark Grondona, in *Job Scheduling Strategies for Parallel Processing,* D. Feitelson, L. Rudolph, U. Schwiegelshohn, Eds. (Springer Berlin Heidelberg, Berlin, Heidelberg, 2003), pp. 44-60.
132. S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884-i890 (2018).
133. Heng Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*, (2013).
134. N. Schaap *et al.*, Long-term follow-up of persisting mixed chimerism after partially T cell-depleted allogeneic stem cell transplantation. *Leukemia* **16**, 13-21 (2002).
135. A. Stikvoort *et al.*, Chimerism patterns of long-term stable mixed chimeras posthematopoietic stem cell transplantation in patients with nonmalignant diseases: follow-up of long-term stable mixed chimerism patients. *Biol Blood Marrow Transplant* **19**, 838-844 (2013).
136. William Moy Stratton Russell, Rex Leonard Burch, *The principles of humane experimental technique*. (Methuen, 1959).
137. Chen Glait-Santar *et al.*, Functional Niche Competition Between Normal Hematopoietic Stem and Progenitor Cells and Myeloid Leukemia Cells. *Stem cells* **33**, 3635-3642 (2015).
138. C. Weinreb, A. Rodriguez-Fraticelli, F. D. Camargo, A. M. Klein, Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science* **367**, (2020).
139. Humberto Contreras-Trujillo *et al.*, Deciphering intratumoral heterogeneity using integrated clonal tracking and single-cell transcriptome analyses. *Nature communications* **12**, 6522 (2021).
140. C. E. Eyler *et al.*, Single-cell lineage analysis reveals genetic and epigenetic interplay in glioblastoma drug resistance. *Genome biology* **21**, 174 (2020).

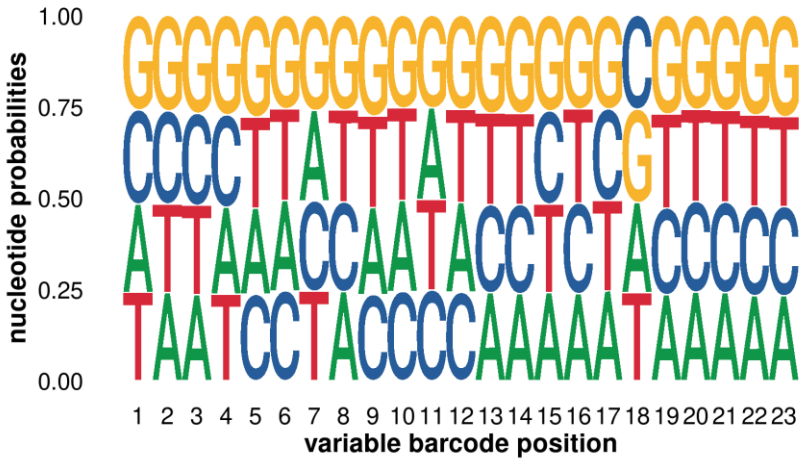
141. B. Adamson *et al.*, A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* **167**, 1867-1882.e1821 (2016).
142. Grace X. Y. Zheng *et al.*, Massively parallel digital transcriptional profiling of single cells. *Nature communications* **8**, 14049 (2017).
143. Tamar Hashimshony *et al.*, CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome biology* **17**, 77 (2016).
144. Hadas Keren-Shaul *et al.*, MARS-seq2.0: an experimental and analytical pipeline for indexed sorting combined with single-cell RNA sequencing. *Nature protocols* **14**, 1841-1862 (2019).
145. Yohei Sasagawa *et al.*, Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. *Genome biology* **19**, 29 (2018).
146. Anna S. Nam *et al.*, Somatic mutations and cell identity linked by Genotyping of Transcriptomes. *Nature* **571**, 355-360 (2019).
147. Catherine Gutierrez *et al.*, Multifunctional barcoding with ClonMapper enables high-resolution study of clonal dynamics during tumor evolution and treatment. *Nature Cancer* **2**, 758-772 (2021).
148. David Feldman *et al.*, CloneSifter: enrichment of rare clones from heterogeneous cell populations. *BMC Biology* **18**, 177 (2020).
149. Aziz M. Al'Khafaji, Daniel Deatherage, Amy Brock, Control of Lineage-Specific Gene Expression by Functionalized gRNA Barcodes. *ACS Synth. Biol.* **7**, 2468-2474 (2018).
150. Agilent Technologies Inc., https://www.agilent.com/en/product/next-generation-sequencing/amplicon-based-next-generation-sequencing-ngs/amplicon-amplification-assay/haloplex-custom-kits-232855, Last accessed: January 2022
151. Daniel C. Koboldt, Best practices for variant calling in clinical sequencing. *Genome Medicine* **12**, 91 (2020).
152. Guoli Chen, Stacy Mosier, Christopher D. Gocke, Ming-Tseh Lin, James R. Eshleman, Cytosine deamination is a major cause of baseline noise in next-generation sequencing. *Molecular diagnosis & therapy* **18**, 587-593 (2014).
153. William A Beard, Vinod K Batra, Samuel H Wilson, DNA polymerase structure-based insight on the mutagenic properties of 8-oxoguanine. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis* **703**, 18-23 (2010).
154. M. Costello *et al.*, Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic acids research* **41**, e67 (2013).
155. M. Pogoda *et al.*, Single Molecule Molecular Inversion Probes for High Throughput Germline Screenings in Dystonia. *Front Neurol* **10**, 1332 (2019).
156. Manuel Allhoff *et al.*, Discovering motifs that induce sequencing errors. *BMC bioinformatics* **14**, S1 (2013).
157. Nicholas Stoler, Anton Nekrutenko, Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics and Bioinformatics* **3**, (2021).
158. Elise M Bekers *et al.*, Identification of novel GNAS mutations in intramuscular myxoma using next-generation sequencing with single-molecule tagged molecular inversion probes. *Diagnostic pathology* **14**, 15 (2019).
159. S. Cantsilieris, H. A. Stessman, J. Shendure, E. E. Eichler, Targeted Capture and High-Throughput Sequencing Using Molecular Inversion Probes (MIPs). *Methods in molecular biology* **1492**, 95-106 (2017).
160. R. Gallon *et al.*, A sensitive and scalable microsatellite instability assay to diagnose constitutional mismatch repair deficiency by sequencing of peripheral blood leukocytes. *Hum Mutat*, (2019).
161. J. Zhang *et al.*, A molecular inversion probe-based next-generation sequencing panel to detect germline mutations in Chinese early-onset colorectal cancer patients. *Oncotarget* **8**, 24533-24547 (2017).

162. M. I. Perez Millan *et al.*, Next generation sequencing panel based on single molecule molecular inversion probes for detecting genetic variants in children with hypopituitarism. *Mol Genet Genomic Med*, (2018).
163. J. K. Yoon *et al.*, microDuMIP: target-enrichment technique for microarray-based duplex molecular inversion probes. *Nucleic acids research* **43**, 9 (2015).
164. Milko B. Kermekchiev, Lyubka I. Kirilova, Erika E. Vail, Wayne M. Barnes, Mutants of Taq DNA polymerase resistant to PCR inhibitors allow DNA amplification from whole blood and crude soil samples. *Nucleic acids research* **37**, e40-e40 (2009).
165. New England Biolabs Inc., https://international.neb.com/protocols/2012/09/06/pcr-guidelines-for-hemo-klentaq-m0332, Last accessed:
166. Thermo Fisher Scientific Inc., https://www.thermofisher.com/order/catalog/product/F549L, Last accessed: January 2022
167. Roche Molecular Systems Inc., https://rochesequencingstore.com/catalog/kapa-hifi-plus-dntps/, Last accessed: January 2022
168. ThermoFisher Scientific Inc., https://www.thermofisher.com/order/catalog/product/12351010, Last accessed: January 2022
169. S. Filges, E. Yamada, A. Stahlberg, T. E. Godfrey, Impact of Polymerase Fidelity on Background Error Rates in Next- Generation Sequencing with Unique Molecular Identifiers/Barcodes. *Scientific reports* **9**, 7 (2019).
170. A. Waalkes, K. Penewit, B. L. Wood, D. Wu, S. J. Salipante, Ultrasensitive detection of acute myeloid leukemia minimal residual disease using single molecule molecular inversion probes. *Haematologica* **102**, 1549-1557 (2017).
171. D. C. Koboldt *et al.*, VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research* **22**, 568-576 (2012).
172. J. S. Hecker *et al.*, CHIP & HIPs: Clonal Hematopoiesis is Common in Hip Arthroplasty Patients and Associates with Autoimmune Disease. *Blood*, (2021).
173. Philip Kleinert, Beth Martin, Martin Kircher, HemoMIPs—Automated analysis and result reporting pipeline for targeted sequencing data. *PLOS Computational Biology* **16**, e1007956 (2020).
174. Nils Koelling *et al.*, amplimap: a versatile tool to process and analyze targeted NGS data. *Bioinformatics* **35**, 5349-5350 (2019).
175. Jessie J F Medeiros *et al.*, SmMIP-tools: a computational toolset for processing and analysis of single-molecule molecular inversion probes-derived data. *Bioinformatics*, (2022).
176. Chun Hang Au, Dona N. Ho, Ava Kwong, Tsun Leung Chan, Edmond S. K. Ma, BAMClipper: removing primers from alignments to minimize false-negative mutations in amplicon next-generation sequencing. *Scientific reports* **7**, 1567 (2017).
177. Ravi Vijaya Satya, John DiCarlo, Edge effects in calling variants from targeted amplicon sequencing. *BMC genomics* **15**, 1073 (2014).

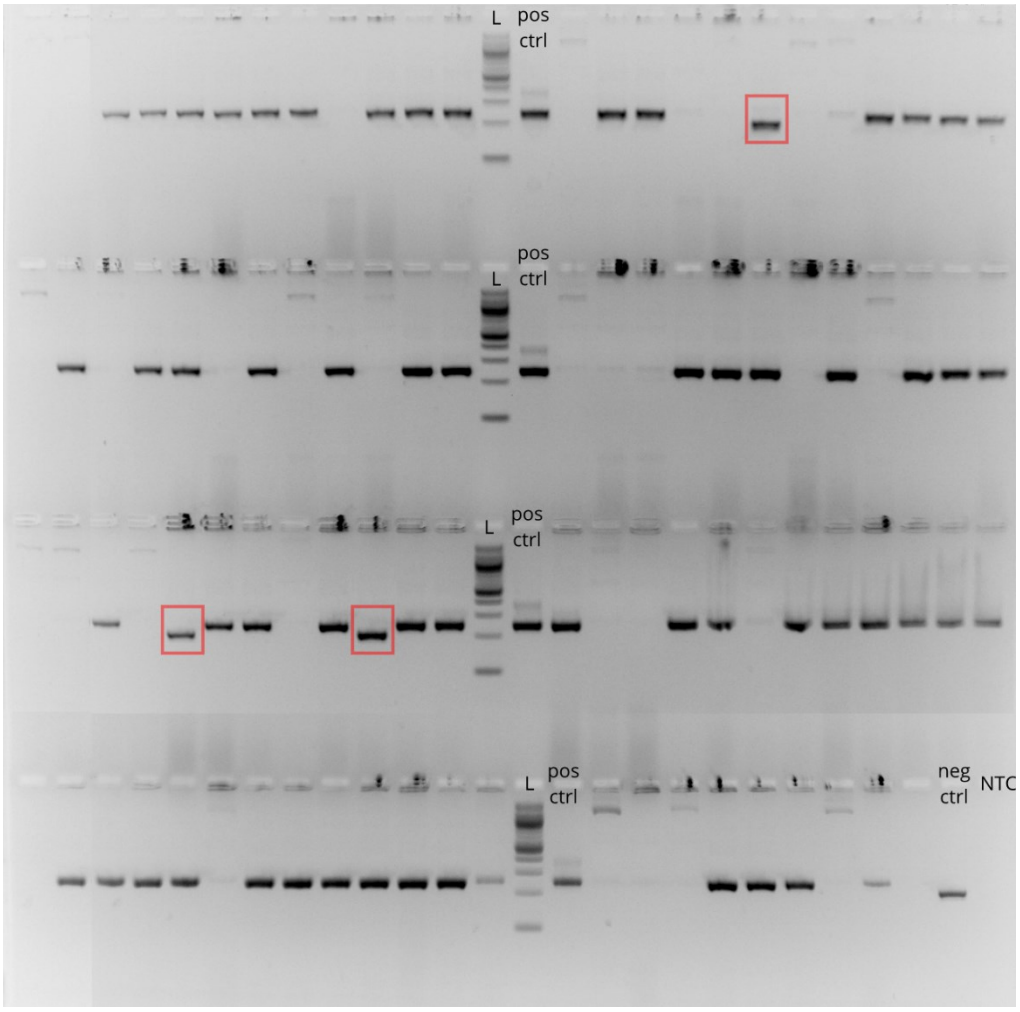
9 Supplemental material



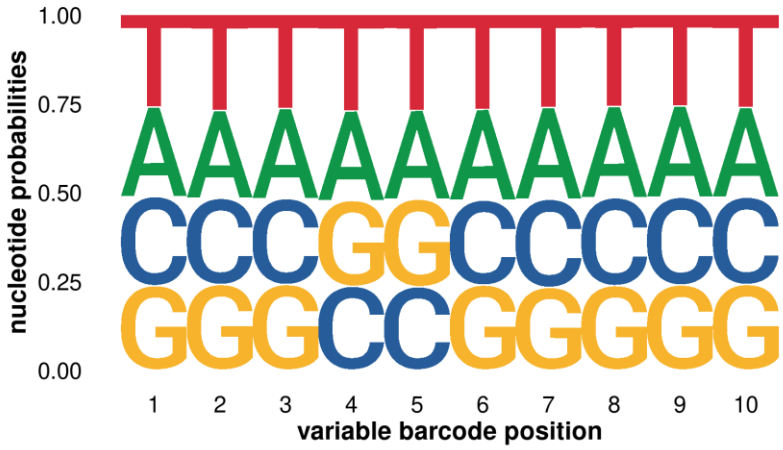
Supplemental Figure 1: Gel documentation of the products from screening PCRs of E.coli colonies that were transformed using the DNABC plasmid. PCR primers used annealed outside the barcode insert, producing an amplicon of 367bp for plasmid carrying the barcode and 267bp for plasmid without barcode integration. For easier identification of correct amplicon sizes products from a control PCR on clean plasmid not having an integrated barcode ('ctrl') were used as size comparison. Out of 91 successfully screened colonies, only two samples (red boxes) showed deviations from the expected amplicon size. This corresponds to 97.8% of plasmids within the pool carrying the correct insert. 'L' = Ladder, 'NTC' = Non-Template Control.



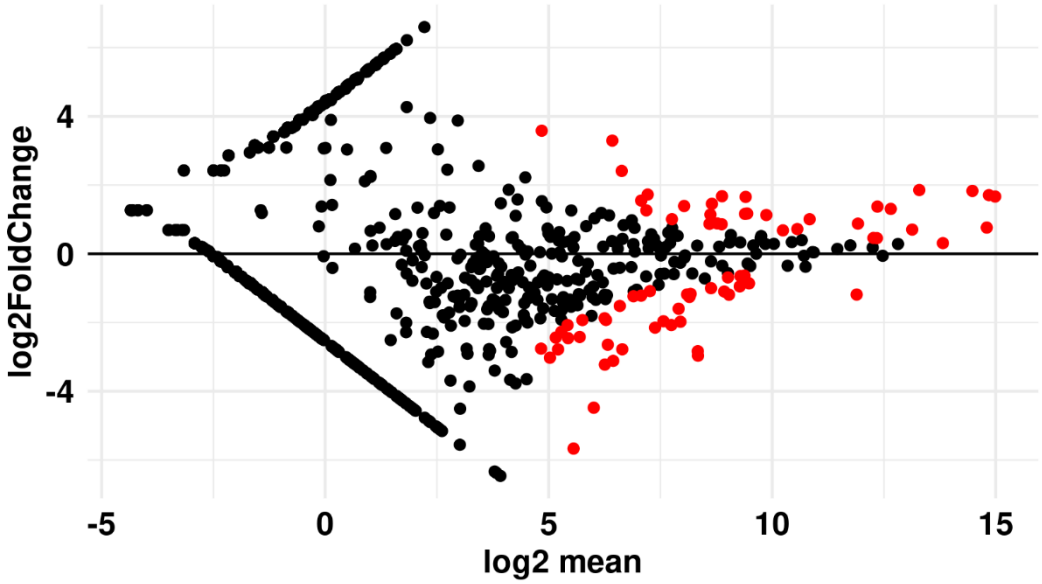
Supplemental Figure 2: Sequence logo of the variable positions within the detected high-complexity barcodes (from replicate 2). Nucleotide frequencies are evenly distributed at each position despite a slight over-representation of G throughout the barcode.



Supplemental Figure 3: Screening PCR for successful integration of the expressed barcode insert into the plasmid. E.coli colonies transformed with the expressed barcode plasmid pool were picked and screened for the presence of the barcode insert via PCR. Out of 63 successfully screened colonies only 3 indicated absence of the barcode insert (red boxes). This corresponds to 95.24% of the plasmids within the pool carrying the correct insert. 'L='Ladder, 'pos ctrl'=amplicon from plasmid carrying the correct barcode insert, 'neg ctrl' = amplicon from plasmid without the insert, 'NTC' = Non-Template Control



Supplemental Figure 4: Sequence logo showing the variable positions of the detected expressed barcodes within the final plasmid pool. Base frequencies are evenly distributed at all variable positions. Only a slight over-representation of T and A is consistently observed across all positions



Supplemental Figure 5: Differentially represented barcodes between the Start and Control groups of the in vivo treatment experiment (see 4.1.6.2). 73 barcodes had statistically different frequencies detected when comparing biological replicates between the Start and Control groups. 33 barcodes were enriched, indicating increased fitness due to higher proliferation compared to the average cell within the barcoded leukemic cell population. In contrast, 40 barcodes were depleted after the outgrowth, indicating slower proliferation compared to the population's mean. These results indicate that intratumoral heterogeneity, i.e. differences in cellular fitness among the subclones, also influences the outgrowth of leukemic cells of the AML-491 Pdx line without additional selective pressure of the in vivo chemotherapeutic treatment.

Supplemental Table 1: Oligonucleotide probes included in the final smMIP panel. Probe sequences are listed in 5'-3' orientation. RC refers to the relative concentration of the respective probe within the smMIP pool. For initial setup all probes were pooled equimolar ('1x'). To balance target capture efficiencies within the panel relative molar concentrations of probes showing decreased performance were increased up to 50-fold ('50x').

Target	Probe name	RC	Probe sequence
ASXL1	ASXL1_0311	1x	AGAGAGGCGGCCACCACTGCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCACGGAGTCCTCCTGCCGGGG
ASXL1	ASXL1_0313	50x	AGCCCAGGGGAGGCCCGAGCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGGGGGGGGTGGCCCGGGTGG
ASXL1	ASXL1_0315	50x	AGAGAGGACCTGCCTTCTCTGAGAANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNAGTGTACGTCAGATCT
ASXL1	ASXL1_0317	1x	GTCCTCCCAAACCTCAGTAGCTGANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGAGGGCTACAGTTGGAC
ASXL1	ASXL1_0319	25x	AGACAATGGTCCCATTCTGTCTCTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGCAGCCTCAGTTGCATC
ASXL1	ASXL1_0321	1x	TCCTCACCGACTGATTGCCTGCAGANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGAGATGATACATTAGAG
ASXL1	ASXL1_0322_SNP_a	1x	AGGGAAAGTGATACTAGACAAGAAAACTTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCAGTTCCACACCTGAA
ASXL1	ASXL1_0322_SNP_b	1x	AGGGAAAGTGATACTAGACAAGAAAACTTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCAGTTCCACTCCTGAA
ASXL1	ASXL1_0323	1x	GTGGTTTGATACGTGAAAGTTGAANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGGGTCCTGGCGCCAGTC
ASXL1	ASXL1_1443499	1x	AGGTGGCAGAGGCAGCAGCAGTGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGGGCGAGAGGTCACCAC
ASXL1	ASXL1_0327	25x	AGGCCTCACCACCATCACCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCTCCCCATTTAGAGGATAAGGC
ASXL1	ASXL1_0329	1x	AGCTCTGGACATGGCAGTTCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGGAGTTGGGAGGCATCTCCT
ASXL1	ASXL1_0331	1x	AGGCCTGGCATGGCTGGTCCCCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNAGTGGTGCCAGACTCACA
ASXL1	ASXL1_0333_SNP_a	1x	GTCCTTGCTCCTCATCATCACNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTCATAGTGGGATGACTGTC
ASXL1	ASXL1_0333_SNP_b	1x	GTCCTTGCTCCTCATCATAACNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTCATAGTGGGATGACTGTC
BRAF hotspot	BRAF_0001	1x	AGAAATATATCTGAGGTGTAGTAAGTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTTCCAGACAACTGTTCA

Supplemental Table 1(continued)

Target	Probe name	RC	Probe sequence
BRAF hotspot	BRAF_0002	1x	TCCATTTTGTGGATGGTAAGAATTGANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTCATGAAGACCTCACA
CALR Exon 9	CALR_EXON9_0176	1x	AGGATGAGGAGGATGAGGAGGACNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCAGGACGAGGAGCAGAG
CALR Exon 9	CALR_EXON9_0178	1x	CGCGCCAAATAATGTCTCTGTGAGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGGAGGAAGATGTCCCC
CALR Exon 9	CALR_EXON9_0179	50x	GGCCTTGCCCCCTGCCAGCCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTGCGTTTCTTGTCTTCTTCC
CALR Exon 9	CALR_EXON9_0181	1x	GTCCTCATCATCCTCCTTGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCCTGGAGGCAGGCCTCTCTAC
CBL	CBL_0063	1x	TCGGTATTATATAGCCTTTACTGATACANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGTGGACACCTCATGTG
CBL	CBL_0064_SNP_a	1x	ACTTTTACTTTTTTTTGATCTCTAGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCGACTTTTTTCAGCTA
CBL	CBL_0064_SNP_b	1x	ACTTTTACTTTTTTTTGGTCTCTAGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCGACTTTTTTCAGCTA
CBL	CBL_0068	1x	AGTTGGAATGTGGAGCCCATCTCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGATCCGTACCTGCCAGG
CBL	CBL_0071	1x	ACCCCAAAAGCCAGGCCACCCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGCAAAAGATAGTAACAGATGCA
CBL	CBL_0072	25x	TCTGAAAATACTTAAAATATTAATCTACTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNNTGGGTTCAGTACCTTT
CSF3R	CSF3R_1_0003	1x	AGTGCCCTGGCCCTGGGCTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCAGGGTCCCCAAGGGGCTGGC
CSF3R	CSF3R_1_0005	50x	AGGTCTGGACCAGAGTGGGGAGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGGGGCTGCCCAGCAGCTG
CSF3R	CSF3R_1_0007_SNP_a	50x	AGCCCCTCTTGGCGGGCCTCACCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTCCCAGTCTGGCACCAG
CSF3R	CSF3R_1_0007_SNP_b	50x	AGCCCCTCTTGGCAGGCCTCACCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTCCCAGTCTGGCACCAG
CSF3R	CSF3R_2_0009	1x	GGCTCCAGGCCATGGAGGACNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTCTCCCCTTACCTGGGGTCA

Supplemental Table 1(continued)

Target	Probe name	RC	Probe sequence
DNMT3A Exon 10	DNMT3A_EX10_0297	50x	AGTGGGCTGCTGCACAGCAGGAGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNACTCTGGCTCGTCATCG
DNMT3A Exon 10	DNMT3A_EX10_0299	1x	AGTCACCTTGACCTCTCCAGGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCACCTTCTGAGACTCCCCGG
DNMT3A Exon 10	DNMT3A_EX10_0300	1x	GGGGGCCTTCCACTGCCAGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGTGGGGTCCGATGCTGGGGAC
DNMT3A Exon 10	DNMT3A_EX10_0057	1x	TGGCCAGCTCTTTCCGGGGGCCTTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGTGGGGTCCGATGCTGGGG
DNMT3A Exon 10	DNMT3A_EX10_0058	1x	GGCGTGGTAGCCACAGTGGGGGATNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNAAAGAGCTGGCCACGG
DNMT3A Exon 10	DNMT3A_EX10_0302	1x	GCATCCCCCACTGTGGCTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGTCATTGCAGGAATGAATGCTG
DNMT3A Exon 11	DNMT3A_EX11_0291	50x	AGTTTCCCCCACACCAGCTCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGTCTCCGAACCACATGACCCAG
DNMT3A Exon 11	DNMT3A_EX11_0293	1x	AGGTTCTTGATCCCAGGGCCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGGAGCCGAGCAGCTGAAGGCAC
DNMT3A Exon 11	DNMT3A_EX11_0295	1x	AGGCCGCATTGTGTCTTGGTGGATGANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGTAATGATTTCTGCTC
DNMT3A Exon 12	DNMT3A_EX12_0284	1x	GGCCTGGTGGAACGCACTGCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCTCCTAGTGCTCTAGGCTCC
DNMT3A Exon 12	DNMT3A_EX12_0286	50x	GGCAGGGGCTGGGAGCCTCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNAGCTCAGCGGCATCAGCTTCT
DNMT3A Exon 12	DNMT3A_EX12_0288	1x	AGCCATCTACGAGGTCCTGCAGGTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGTCCCCCCAGGTGTGT
DNMT3A Exon 13	DNMT3A_EX13_0278	1x	GTTCTGCACCTCCACGGCCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGGGTCCCAGAAAGCTGGGTGC
DNMT3A Exon 13	DNMT3A_EX13_0280	1x	GTCAGGACAGGCTGGAAGGCAGATNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTGGCAGTGTCACTCTC
DNMT3A Exon 13	DNMT3A_EX13_0282_SNP_a	1x	GCTTCCAGCCTTCTGGCCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGCGCGGGGAAGCTGTTCCCGGT
DNMT3A Exon 13	DNMT3A_EX13_0282_SNP_b	1x	GCTTCCAGCCTTCTGGCCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGCGCAGGGAAGCTGTTCCCGGT
DNMT3A Exon 14	DNMT3A_EX14_0272	1x	GTGGAGGTGGTGCGTAGGCAGCTGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCCTCTCCAGAAGCAGG

Supplemental Table 1(continued)

Target	Probe name	RC	Probe sequence
DNMT3A Exon 14	DNMT3A_EX14_0050	1x	CTGCTTCTGGAGAGGGTGGCACCAGGNNNNNAG ATCGGAAGAGCGTGTGTATAAGAGACAGNNNNN ACCTCCACCAGCCAAAAA
DNMT3A Exon 14	DNMT3A_EX14_1068189	1x	GCTTTTTGGCTGGTGGAGGTGGTGNNNNNAGATC GGAAGAGCGTGTGTATAAGAGACAGNNNNNGAT GCAGGCCTCCTGGT
DNMT3A Exon 14	DNMT3A_EX14_0276	1x	AGGAGATTATTGATGAGCGCACAAGNNNNNAGA TCGGAAGAGCGTGTGTATAAGAGACAGNNNNNC ACGGACATGTGGGTG
DNMT3A Exon 14	DNMT3A_EX14_0277	50x	GGCAGCTGCCTACGCACCACNNNNNAGATCGGAA GAGCGTGTGTATAAGAGACAGNNNNNTTTCCTGT CAGCCTGTAACT
DNMT3A Exon 15	DNMT3A Exon 16	50x	GGGAACTCTGGCACTCCTNNNNNAGATCGGAAG AGCGTGTGTATAAGAGACAGNNNNNCAGGCGCG GCCTCCTCTGACGC
DNMT3A Exon 16	DNMT3A_EX16_0267_SNP_ a	1x	GGCCTCTCCCTCCCCGGGCNNNNNAGATCGGAAG AGCGTGTGTATAAGAGACAGNNNNNCTCCAACGA AGAGGGGGTGTTC
DNMT3A Exon 16	DNMT3A_EX16_0267_SNP_ b	1x	GGCCTCTCCCTCCCTGGGCNNNNNAGATCGGAAG AGCGTGTGTATAAGAGACAGNNNNNCTCCAACGA AGAGGGGGTGTTC
DNMT3A Exon 16	DNMT3A_EX16_0268	50x	ACACCCACCCAGGAGAGGTGCCGTNNNNNAGATC GGAAGAGCGTGTGTATAAGAGACAGNNNNNCTT TCCAGACATCTGC
DNMT3A Exon 17	DNMT3A_EX17_0263	1x	AGGGTCAGAAACCACCAGGACNNNNNAGATCGG AAGAGCGTGTGTATAAGAGACAGNNNNNACCTC ACGGCCCCCACAGCA
DNMT3A Exon 17	DNMT3A_EX17_0264	1x	GTTGTGGCCTCCAGTGGTCTCCTNNNNNAGATCG GAAGAGCGTGTGTATAAGAGACAGNNNNNCAGT ACGACGACGACGG
DNMT3A Exon 18	DNMT3A_EX18_0256	1x	AGTCCTCTCGCCGCCGCAGCNNNNNAGATCGGAA GAGCGTGTGTATAAGAGACAGNNNNNCAAGGCT CAGCCAAGGGAGCTC
DNMT3A Exon 18	DNMT3A_EX18_0258	1x	AGCAGAGGAGACTCTCAGCCCNNNNNAGATCGG AAGAGCGTGTGTATAAGAGACAGNNNNNCAGTT CCAGGGGTCTTCCT
DNMT3A Exon 18	DNMT3A_EX18_0260	50x	GGCCCTCCCGGCTCCAGANNNNNAGATCGGAAG AGCGTGTGTATAAGAGACAGNNNNNTGGGGCCG GGGGCTGCCCAGGC
DNMT3A Exon 19	DNMT3A_EX19_0252	50x	TCCTCTTCTCAGCTGGGACAGGTNNNNNAGATCG GAAGAGCGTGTGTATAAGAGACAGNNNNNGGCT CCCCCATCCTGGGA
DNMT3A Exon 19	DNMT3A_EX19_0255	1x	AGCCCATCCGGGTGCTGTCTCTCNNNNNAGATCG GAAGAGCGTGTGTATAAGAGACAGNNNNNGGGA GCTTGGGACACCG
DNMT3A Exon 20	DNMT3A_EX20_0246	50x	GTGATGGAGTCCTCACACACCTCNNNNNAGATCG GAAGAGCGTGTGTATAAGAGACAGNNNNNGGGC TGCGCCCCACAGC
DNMT3A Exon 20	DNMT3A_EX20_0250	1x	GGACGTCCGCAGCGTCACACNNNNNAGATCGGA AGAGCGTGTGTATAAGAGACAGNNNNNACTTGG GCATTCAGGTGGAC

Supplemental Table 1(continued)

Target	Probe name	RC	Probe sequence
DNMT3A Exon 20	DNMT3A_EX20_0251	1x	GCCTCGGAGGTGTGTGAGGACNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGGGTGCCAGGGAGATGGCTC
DNMT3A Exon 21	DNMT3A_EX21_0243	1x	GGCCTGCTGTCCAGGGACANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGAGCAGGGTTGACGATGGAGA
DNMT3A Exon 21	DNMT3A_EX21_0244	1x	AGCTGGTGCTTCCGCACANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCGATCTGGTGATTGGGGGCAGTC
DNMT3A Exon 22	DNMT3A_EX22_0236	50x	GGCCACCACATTCTCAAAGANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCAGCAGTCCAAGGTAGAAGC
DNMT3A Exon 22	DNMT3A_EX22_0237	50x	AGAACTCAAAGAAGAGCCGGCCAGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTCTTGTCACTAACGCC
DNMT3A Exon 22	DNMT3A_EX22_0238	1x	AGTGGTGTGGCTCGGGCACNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCGCGCATCATGCAGGAGGCGG
DNMT3A Exon 23	DNMT3A_EX23_0233	1x	AGATGAGCCAAGGAGGAGCATGANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGGAAGGTTACCCCAGAAG
DNMT3A Exon 23	DNMT3A_EX23_0234	1x	GGACTGCAGGTGGGATGACCCANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNAGAAGTGTCAGCTGCACAC
DNMT3A Exon 24	DNMT3A_EX24_0229	1x	AGGGCCCCAGCTGCACGACNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCCAGACACTCCTGCAGCTCCA
DNMT3A Exon 24	DNMT3A_EX24_0230	1x	AGCGTCTAGAACCTCTGCTGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNAGTCTCTCTTCTGCCTCCTAG
DNMT3A Exon 25	DNMT3A_EX25_0222	1x	AGACAGGAAAATGCTGGTCTTTGCCCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCAATCAGAACAGCCAC
DNMT3A Exon 25	DNMT3A_EX25_0224	1x	GTCATGCGTCTACCAAATATGCCANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTGGAGTTTGACCTCGT
DNMT3A Exon 25	DNMT3A_EX25_0226	1x	AGAGGACATCTTATGGTGCACTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCTCTTCATCTAGTTCAGCA
DNMT3A Exon 26	DNMT3A_EX26_23177	1x	AGCGACACAAAGTTAAACAAACAAACANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCCGGTCATGGAGCGTG
DNMT3A Exon 26	DNMT3A_EX26_0218_SNP_a	1x	AGGCTGCCCGGAAGCCGTCTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGCCAAGCGGCTCATGTTGGAGAC
DNMT3A Exon 26	DNMT3A_EX26_0218_SNP_b	1x	AGGCTGCCCGGAAGCTGTCTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGCCAAGCGGCTCATGTTGGAGAC
DNMT3A Exon 26	DNMT3A_EX26_0220	1x	GCCACCTCTTCGCTCCGCTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCAGGGTATTTGGTTTCCCAGTC

Supplemental Table 1(continued)

Target	Probe name	RC	Probe sequence
FLT3 (ITD region)	FLT3_ITD_0090	1x	AGAAACATTTGGCACATTCCATTCTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCTTTTTCCAAAAGCACCT
FLT3 (ITD region)	FLT3_ITD_0092_SNP_a	50x	AGCTGGCTTTCATACCTAAATTGCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTCTCTTGGAAACTCCCA
FLT3 (ITD region)	FLT3_ITD_0092_SNP_b	50x	AGCTGGCTTTCATACCTAAATTGCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTCTCTTGGGAACTCCCA
FLT3 (ITD region)	FLT3_ITD_0093	1x	AGGAGTCTCAATCCAGGTTGCCGTCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTGCACGTACTCACCAT
FLT3 (ITD region)	FLT3_ITD_0095	25x	AGCATTTCTTTTCCATTGGAAAATCTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGTTGATTTCAGAGAAT
FLT3 (N676 hotspot)	FLT3_N676_0087	1x	GTGGCTTCCCAGCTGGGTCATCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCTCTTTGTCATCAAGCTAC
FLT3 (TKD hotspot)	FLT3_TKD_0085	1x	AGCACGTTCCTGGCGGCCAGGTCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCAGCCTCACATTGCCCCT
GNB1 (c57 hotspot)	GNB1_c57_0002	1x	GGGGCTGCTCTGTTGTCCTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGGAGGACACTGCGGGGGCACC
IDH1 hotspot	IDH1_hotspot_1359015	1x	ACCCATCCACTCACAAGCCGGGGGANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGAAAAAAAAAACATGCAAAA
IDH1 hotspot	IDH1_hotspot_1358656	1x	GCATGTTTTTTTTTTCATGGCCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGGCTTGTGAGTGGATGGG
IDH2 (c140-172 hotspot)	IDH2_c140_172_0097	1x	GGCTCCCGGAAGACAGTCCCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCTCTCCACCCTGGCCTACCT
IDH2 (c140-172 hotspot)	IDH2_c140_172_0100	1x	ACATCCCACGCCTAGTCCCTGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTCACAGAGTTCAAGCTGAAG
JAK2 hotspot	JAK2_617_0003	1x	AGGCTTTCTAATGCCTTTCTCANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTCTTTGAAGCAGCAAGTA
JAK2 hotspot	JAK2_617_0004	1x	AGCTTGCTCATCATACTTGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNAAACAGATGCTCTGAGAAAGG
KIT Exon 17	KIT_EX17_0389	1x	GGTCTAGCCAGAGACATCANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTGGTTTTCTTTTCTCCTCCAA
KIT Exon 17	KIT_EX17_0390	1x	CTCTGCTTGACAGTCCTGCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCCTTACTCATGGTCGGATCACAA
KIT Exon 8	KIT_EX8_0386	1x	CTTATAATGCAGAGGGGAAGGACTGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCTCCAATGTGTGGCAG

Supplemental Table 1(continued)

Target	Probe name	RC	Probe sequence
KIT Exon 8	KIT_EX8_0387	1x	AGAAATGGCCATATGTCAGAGTGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTCTATTGTGGGCTCTGGG
KRAS (c12-13 hotspot)	KRAS_c12_13_0076	1x	AGTTTATATTCAGTCATTTTCAGCANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCCTCTATTGTTGGATCA
KRAS (c12-13 hotspot)	KRAS_c12_13_0077	1x	ACTGGTGCAGGACCATTCTTTGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGTAGGCAAGAGTGCCTTGAC
KRAS (c58-61 hotspot)	KRAS_c58_61_0074	1x	GGGAGGGCTTTCTTTGTGTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGAAGCAAGTAGTAATTGATGG
NPM1 hotspot	NPM1_hotspot_0205	1x	ATTTCATTTCTGTAACAGTTGATATCTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCCAGGCTATTCAAGAT
NPM1 hotspot	NPM1_hotspot_4162406	1x	GAAAAAAAAAAAAAGAAATGTGGTTAAGGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNNAGACGGAAAATTTTTT
NRAS (c12-13 hotspot)	NRAS_c12_13_0013	1x	TCCAACCACCACCAGTTTGTACTCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNACCACTGGGCCTCACCTC
NRAS (c58-61 hotspot)	NRAS_c58_61_0012	1x	GGCTTCCTCTGTGTATTTGCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNAGTGGTTATAGATGGTGAAAC
PPM1D Exon 6	PPM1D_ex6_0030	1x	ACCTGCCCTGGTTCGTAGCAATGCCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGTTGAGTTCTGGGATAAATT
PPM1D Exon 6	PPM1D_ex6_0039	1x	AGGTAAAACAAATAAGAAAAAATTTATCCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTTGCTACGAACCAGGG
PPM1D Exon 6	PPM1D_ex6_0037	1x	GGAAATCCTTTACTTCATCAACACAGGAANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGAAGTAGTGGTGCTCA
PPM1D Exon 6	PPM1D_ex6_0038	1x	CTGGGAAATGAGGTTTTTCCAAACTTAGGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGCAGACTTAGGGGCCA
PPM1D Exon 6	PPM1D_ex6_970662	1x	GGTGAGTTTAACAGAGTTCTTTCGCTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNAAACCTCATTTCCCAGAT
PPM1D Exon 6	PPM1D_ex6_0159	1x	GCTGAGATAGCTCGAGAGANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNAGTCACTGGAGGAGGATCCAT
PPM1D Exon 6	PPM1D_ex6_0161	1x	AGCCTTCCAATTGGCCTTGTGCCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTGTCCAAGGTGTAGTCA
PPM1D Exon 6	PPM1D_ex6_0163	1x	AGGACATTAGAAGAGTCCAATTCTGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCAAACACTGTCATGGAC
PPM1D Exon 6	PPM1D_ex6_0164	1x	AGCCTGCAAGTCTCCCCANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGAACCCCTCCAACAAACTTTAAA

Supplemental Table 1(continued)

Target	Probe name	RC	Probe sequence
PPM1D Exon 6	PPM1D_ex6_0167	1x	AGGCATTGCTACGAACCAGGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCAATTTTCTTCAAGTGGTTC
PPM1D Exon 6	PPM1D_ex6_0169_SNP_a	1x	AGAATCATGTATCCTTAAAGTCAGGGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTTTGGCCAGGAGTTGAC
PPM1D Exon 6	PPM1D_ex6_0169_SNP_b	1x	AGAATCATGTATCCTTAAAGTCAGGGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTTTGGCCAGGAATTGAC
PPM1D Exon 6	PPM1D_ex6_0172	1x	CCATTTCGTCTATGCTTCTTCANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGGCCCCTAAGTCTGCGTCG
PTPN11 Exon 13 hotspot	PTPN11_EX13_hotspot_0084	1x	GTCAATATCGCAGTCAACACCTACGANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNAATGCTGGACCGCCAT
PTPN11 Exon 13 hotspot	PTPN11_EX3_hotspot_0079	1x	AGAAATTTGCCACTTTGGCTGAGTTGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCCTCCCTTTCCAATGG
PTPN11 Exon 13 hotspot	PTPN11_EX3_hotspot_0082	1x	AGTAATCACCAGTGTTCTGAATNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGACATCTCCATTCTTCTCT
RUNX1 Exon 4B	RUNX1_EX4B_0371	50x	AGTTGGGGCTGTCGGTGCGCACNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCCGCCTGTCCTCCCACCAC
RUNX1 Exon 4B	RUNX1_EX4B_0375	50x	AGCTCCTACCAGACGGCGACAGGGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCTCAGCGCGGTGGAAGG
RUNX1 Exon 4B	RUNX1_EX4B_0377	50x	CGCACTGGCGCTGCAACAAGACNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGAGGAGCGGCGACCGCAG
RUNX1 Exon 4B	RUNX1_EX4B_0092	1x	GCTTGCCGGCCAGGGCAGCGCCGGCGTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGAGGAAGTTGGGGCTG
RUNX1 Exon 4B	RUNX1_EX4B_0091	1x	GCGGTGGAAGGCGGCGTGANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNATGCTGCGGTCGCCGCTCCTC
RUNX1 Exon 4B	RUNX1_EX4B_0094	1x	GTGGAGGTGCTGGCCGACCACNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCAGGCAAGATGAGCGAGGCG
RUNX1 Exon 4B	RUNX1_EX4B_0098	1x	GGACCCTGCAAACAGCTCCTACCANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTCGCTCATCTTGCCTGG
RUNX1 Exon 4B	RUNX1_EX4B_0093	1x	GACGCCGGCGCTGCCCTGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGCTTGTTGTGATGCGTATCCCCG
RUNX1 Exon 5	RUNX1_EX5_0085	1x	CCCACATTTCAAAATTCTAGTGATTTCTGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNNTGACCTCAGGTTTGTC
RUNX1 Exon 5	RUNX1_EX5_0086	1x	CCTGGTTCTTCATGGCTGCGGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGTGGGTTTGTTGCCATGAAACG

Supplemental Table 1(continued)

Target	Probe name	RC	Probe sequence
RUNX1 Exon 5	RUNX1_EX5_0088	1x	CCAGTCAAAGGACAAATGCAGACATCAGGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGCTCAGCCGAGTAGTT
RUNX1 Exon 5	RUNX1_EX5_1949872	1x	GCACTCTGGTCACTGTGATGGCTGGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNACTACACAAATGCCCTA
RUNX1 Exon 5	RUNX1_EX5_0367	1x	AGCCGAGTAGTTTTCATCANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGTTTTGACAGATAACGTACCT
RUNX1 Exon 5	RUNX1_EX5_0370	1x	AGCTGAGAAATGCTACCGCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNNCATTTGTCCTTTGACTGGTGT
RUNX1 Exon 6	RUNX1_EX6_0364	1x	GGGACACGATAGAGAACAAAACNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGGGGCCCATCCACTGTGA
RUNX1 Exon 6	RUNX1_EX6_0365	1x	GGGGCTGGTACACCCTCCAGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTCACTGTCTTCACAAACCCA
RUNX1 Exon 7	RUNX1_EX7_0357	1x	GGTGGTGTGGGCTGACCCTCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGTGGGCTCCATCTGGTACTT
RUNX1 Exon 7	RUNX1_EX7_0361	50x	GTGCCTCCCTGAACCACTCCACNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGAGCTTGTCCTTTTCCGAG
RUNX1 Exon 7	RUNX1_EX7_0362	50x	AGCAGCTGCGGCGCACAGCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTCCCCTCCCCTCCCTGCTCCC
RUNX1 Exon 8	RUNX1_EX8_0083	1x	AGCTGAGCTGGGGTGGAAGGTCCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNATTTCACCTGGACGTGCCAG
RUNX1 Exon 8	RUNX1_EX8_0080	1x	GCGTTGCTGGGTGCACAGAAGGAGAGGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNAGCTCAGCTGCAAAGA
RUNX1 Exon 8	RUNX1_EX8_0351	1x	AGAAGGAGAGGCAATGGATCCCAGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTTCAAGTGGCTTACTT
RUNX1 Exon 8	RUNX1_EX8_0353	1x	AGAAATGAGTGGCCCTTGTTCANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCCACGGTGGGGATGGTTGG
RUNX1 Exon 8	RUNX1_EX8_0355	50x	AGCAACGCCCATTTCACCTGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGATTCTCTTCAGATACAAGG
RUNX1 Exon 9	RUNX1_EX9_0335	50x	GGCCTCCACCACGTCGCTCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCCAGCCGGGCCAGGCCTGGCG
RUNX1 Exon 9	RUNX1_EX9_0339	1x	AGCCCATGGCCGACATGCCGANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNACAGGTGGTAGGAGGGCGAG
RUNX1 Exon 9	RUNX1_EX9_0341	50x	GTCAGGTCGGGTGCCGCTGCANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGTGACCGGCGTCGGGGAGT

Supplemental Table 1(continued)

Target	Probe name	RC	Probe sequence
RUNX1 Exon 9	RUNX1_EX9_0346	50x	AGCGCTCGCCGCCGCGCANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCGTCGCAAGCGCAGGGAGGCCC
RUNX1 Exon 9	RUNX1_EX9_0348	50x	GGCCACGCGCTACCACACCTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCCTCCATCTCCGACCCCCGCA
RUNX1 Exon 9	RUNX1_EX9_0350	50x	AGCGACCCGCGCCAGTTCCCCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCCTGGGGCAGAGGGAAGAGC
RUNX1 Exon 9	RUNX1_EX9_0066	1x	GTCGCTCTGGTTCGGGAGGCTGGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGGGCCAGGCCTGGCGCCT
RUNX1 Exon 9	RUNX1_EX9_0067	1x	AGCGCTCGCCGCCCACCANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTTGCTGTGGCTGCCCTCGGCCTCCACC
RUNX1 Exon 9	RUNX1_EX9_0068	1x	GTGGTAGGAGGGCGAGCTGGCTTGGANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGGTGGAGGCGTTGGTG
RUNX1 Exon 9	RUNX1_EX9_0069	1x	GGGCGGCGGCAGGTAGGTGTGGTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNAGAACTGGTAGGAGCCGG
RUNX1 Exon 9	RUNX1_EX9_0074	1x	AGCGCTCGCCGCCGCGCATNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNAAGCGCAGGGAGGCCCGTTCC
SF3B1 (K666 hotspot)	SF3B1_K666_hotspot_0306	1x	ACCAGTGTGTCTCGCTTGCCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTACATTACAACTTACCATGTTC
SF3B (K700 hotspot)	SF3B1_K700_hotspot_0305	25x	AGCAACTCCTTATGGTATCGAATCTTTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGTTAAAACCTGTGTTTG
SRSF2 (p95 hotspot)	SRSF2_p95_hotspot_0046	10x	GCGGCCGTAGCGCGCCATTTGCACCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTTTACCTGCGGCTCCGG
SRSF2 (p95 hotspot)	SRSF2_p95_hotspot_0047	1x	ACCGCCACCCCGCAGGTACGGGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTGGACGGGGCCGTGCTGG
SRSF2 (p95 hotspot)	SRSF2_p95_hotspot_1028291	1x	GTCCAGCACGGCCCCGTCCATGGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGTACCTGCGGGGTGGCGG
SRSF2 (p95 hotspot)	SRSF2_p95_hotspot_1024770	10x	ACGGACGCCGGAGCCGCAGGTAAANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGGTGCAAATGGCGCGC
TET2 Exon 10	TET2_EX10_0525	1x	TCTGCCTTTATACAAAGTCTCTGACNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGGCTGTAATGTCTTACT
TET2 Exon 10	TET2_EX10_0527	1x	AGAAGCCAAGAAAGCTGCAGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNAGGAGAAAAAACGGAGTGGTG
TET2 Exon 10	TET2_EX10_0528	1x	AGTCAGCCCCATCACGTACANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCCAGTCAAGACTTGCCGACAAAGGA

Supplemental Table 1(continued)

Target	Probe name	RC	Probe sequence
TET2 Exon 10	TET2_EX10_0529	1x	AGGTAAATTTAATGTAAAGCATTTGTAGNNNNNA GATCGGAAGAGCGTGTGTATAAGAGACAGNNNN NCTGGAGAACAGCTCAA
TET2 Exon 10	TET2_EX10_0532	1x	ACGTGAAGCTGCTCATCCTCNNNNNAGATCGGAA GAGCGTGTGTATAAGAGACAGNNNNNAAAAGAA CTCAGTACCTGAA
TET2 Exon 10	TET2_EX10_0535	1x	GGGCTGACTTTTCCTTTTCATTTNNNNNAGATCGG AAGAGCGTGTGTATAAGAGACAGNNNNNATACC ACACAACACATTT
TET2 Exon 11	TET2_EX11_0537	1x	GGATCCACCAATCCATACANNNNNAGATCGGAAG AGCGTGTGTATAAGAGACAGNNNNNCTCTACAGA AGCAGCCACCAC
TET2 Exon 11	TET2_EX11_0539	1x	TCTAATCCCATGAACCCTTACCCTGNNNNNAGATC GGAAGAGCGTGTGTATAAGAGACAGNNNNNCAA ACTCTTCACACACT
TET2 Exon 11	TET2_EX11_0540	1x	AGTGGACAACTGCTCCCCANNNNNAGATCGGAAG AGCGTGTGTATAAGAGACAGNNNNNGCAGGTTC ATATTTGAATTCT
TET2 Exon 11	TET2_EX11_0542	50x	AGCCAGAGTTTTACATCTNNNNNAGATCGGAAGA GCGTGTGTATAAGAGACAGNNNNNCAGTCTCAGC CGATGGATCTGT
TET2 Exon 11	TET2_EX11_0547	1x	AGATGCTTCCAGCTCTTAACCANNNNNAGATCGG AAGAGCGTGTGTATAAGAGACAGNNNNNCAGTG CAGCTCCGGGCATGTTC
TET2 Exon 11	TET2_EX11_0548	1x	AGTGATGCTAATGGTCAGGAAAAGNNNNNAGAT CGGAAGAGCGTGTGTATAAGAGACAGNNNNNGA CATGCTTTCCCACAC
TET2 Exon 11	TET2_EX11_0549	1x	CGATGAGGTCTGGTCAGACANNNNNAGATCGGA AGAGCGTGTGTATAAGAGACAGNNNNNCTTGTGT CCAAGGAGGCTTA
TET2 Exon 11	TET2_EX11_0550	1x	TCAATTCTCATTGAGTGTGCNNNNNAGATCGGAA GAGCGTGTGTATAAGAGACAGNNNNNGGCTTCT GGTGCAGAGGACAA
TET2 Exon 11	TET2_EX11_0551	1x	GGATCTCCCTCGTCTTTTACCANNNNNAGATCGGA AGAGCGTGTGTATAAGAGACAGNNNNNCGTGGC TCCAACTCATGGG
TET2 Exon 11	TET2_EX11_0552	1x	CTGAAAAAGCCCGTGAGAAAGAGGAANNNNNAG ATCGGAAGAGCGTGTGTATAAGAGACAGNNNNN AGGAATCACCCCACCA
TET2 Exon 11	TET2_EX11_0553	1x	AGTGAAACGGGAGCCTGCTGNNNNNAGATCGGA AGAGCGTGTGTATAAGAGACAGNNNNNCTCTTTG GGAAGCCAAAATGG
TET2 Exon 11	TET2_EX11_0554	1x	AGGACCATGTCCGTGACCACANNNNNAGATCGGA AGAGCGTGTGTATAAGAGACAGNNNNNCCTCAG AAATCCCATGGCAAAAA
TET2 Exon 11	TET2_EX11_0555	1x	GTCACAGGGCCTTACAACAGANNNNNAGATCGG AAGAGCGTGTGTATAAGAGACAGNNNNNGAGCC CACTTACCTGCGTTTC
TET2 Exon 11	TET2_EX11_0556	1x	AGACCACAACCAACCTGTCNNNNNAGATCGGAAG AGCGTGTGTATAAGAGACAGNNNNNCTCCACAGT AACTACATCTCC

Supplemental Table 1(continued)

Target	Probe name	RC	Probe sequence
TET2 Exon 11	TET2_EX11_0557	1x	AGGGTAAGAGAGAACAGAANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGTCTCTGCTGCTGCTGGGGCT
TET2 Exon 11	TET2_EX11_0559	1x	AGAAGCAGAATAAGAGTTGACAGANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCATAGGGCTGGTGCTT
TET2 Exon 11	TET2_EX11_0562	1x	AGGTTTCCATTGCATTGATATGANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCTTAGACAGAGGGTCTTG
TET2 Exon 11	TET2_EX11_0564	1x	AGTGTATGGATGGGTGGTAGACTGAGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNACAACTGCTGAAACCA
TET2 Exon 11	TET2_EX11_0565	1x	ATATTTTGGTTTCCATAACCTAAGTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNAGTGGCCATCCATCTCA
TET2 Exon 11	TET2_EX11_0566	1x	TCCCTACATGATGTACATTTGGTCTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNATAGTCCATGTTTGGA
TET2 Exon 11	TET2_EX11_0567	1x	GTAATCTAGAGGTGGCTCCCATGANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNACATGCCCGGAGCTGCAC
TET2 Exon 11	TET2_EX11_0568	1x	GTGAGAAGGTGAATGATGTTCACNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCTGTGTGGGAAAGCATG
TET2 Exon 4A	TET2_EX4A_0400	1x	AGCCCAAGAAAATGCAGTTAAAGATTTCANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNACTTCGGGGTAAGCCA
TET2 Exon 4A	TET2_EX4A_0401	1x	AGATTCTGAATGAGCAGGAGGGGAANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNAGAATCTGTGAGTTCTG
TET2 Exon 4A	TET2_EX4A_0402	1x	AGGCAGTGCTAATGCCTAATGGTGCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCAGTGGGCCTGAAAATC
TET2 Exon 4A	TET2_EX4A_0403	1x	GTGAACTCCTGGAAAAAACACTGTCTCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNACCATGACAAGAACAT
TET2 Exon 4A	TET2_EX4A_0404	50x	AGAAAACCACATCTCACATAAATGCCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNACAGTTTCTGCCTCTTCC
TET2 Exon 4A	TET2_EX4A_0405	50x	TCACCCATCGCATACCTCAGGGCAGATNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTCCAGATTGTGTTTCC
TET2 Exon 4A	TET2_EX4A_0406	1x	GTGAGTGAGGCCTGTGATGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNATGAGTTGTCCTGTGAGATCA
TET2 Exon 4A	TET2_EX4A_0408	1x	CCTGCAGAAAATAACATCCAGGGAACCACNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCAGTAAACTAGCTGCA
TET2 Exon 4A	TET2_EX4A_0410	1x	GTGCTTACTTCAAGCAAAGCTCAGTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNAGCGTCTGGTGAAGAA

Supplemental Table 1(continued)

Target	Probe name	RC	Probe sequence
TET2 Exon 4A	TET2_EX4A_0413	50x	AGTAACACAACACTTTTAAGGGAAGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCTCCTCCTCTTCCACAGG
TET2 Exon 4A	TET2_EX4A_0414	50x	TCCATCTACACATGTATGCAGCCCTTCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNACACCACCACTACCCC
TET2 Exon 4A	TET2_EX4A_0415	1x	AGACTGCAGGGACAATGACNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNACCACCTTCCCAGAGTCCTAA
TET2 Exon 4A	TET2_EX4A_0416	1x	AGAACACCTCAAGCATAACCCANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTCTGAAAGGCCTCAGAATAAT
TET2 Exon 4A	TET2_EX4A_0417	1x	AGCAGTTGATGAGAAACAAAGAGCANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTCCATTGTGTTCTGAG
TET2 Exon 4A	TET2_EX4A_0418	1x	ACACAGCACTATCTGAAACCANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGGAGAGCTACAGGACAACTGC
TET2 Exon 4A	TET2_EX4A_0419	1x	GTAATGAGGCATCACTGCCATCANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNACACGAGATCTTGTGCCC
TET2 Exon 4A	TET2_EX4A_0429	1x	AGAAATTCCCCTTATAGTCAGACCATGANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGTTTTCATGGTGAAAA
TET2 Exon 4A	TET2_EX4A_0431	1x	AGACCCAAAACTTGCATCACANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCAAGTGCATGCAAAATACAGG
TET2 Exon 4A	TET2_EX4A_0433	1x	AGATATGTCTGGTCAACAAGCTGCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGATCCCAAAGCAAGATCT
TET2 Exon 4A	TET2_EX4A_0435	50x	AGCATGCTGCTCTAAGGTGGCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCTCAGCAAAGGTACTTGAT
TET2 Exon 4A	TET2_EX4A_0436	50x	AGTCAGATGCACAGGCCANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCCTCCCCAGAAGGACACTCAAA
TET2 Exon 4A	TET2_EX4A_0438	50x	GTGCAGCAAAAGAGCATCATTGAGACNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCAAGCCACATGCCTGT
TET2 Exon 4A	TET2_EX4A_0440	1x	ACTAGACAAACCACTGCTGCAGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNAGCAGTTTCACGCCAAGT
TET2 Exon 4A	TET2_EX4A_0442	1x	AGAGTCACCTTCCAAATTACTAGANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNACTTGATAGCCACACCC
TET2 Exon 4A	TET2_EX4A_0444_SNP_a	25x	TCCAGAATTAGCAAATTTATCTTCAGANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTTTATTGGATACACCT
TET2 Exon 4A	TET2_EX4A_0444_SNP_b	25x	TCCATAATTAGCAAATTTATCTTCAGANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTTTATTGGATACACCT

Supplemental Table 1(continued)

Target	Probe name	RC	Probe sequence
TET2 Exon 4A	TET2_EX4A_0445	50x	AGGCTCATAAAAATCTGAAGCTTACNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGAGACACATGGCGTTT
TET2 Exon 4A	TET2_EX4A_0446	1x	CCATCCACAAGGCTGCCCTCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTTCTGTCTGGCAAATGGGAG
TET2 Exon 4A	TET2_EX4A_0447	1x	GTGATGGTATCAGGAATGGACNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGCCACTTGGTGTCTCCATT
TET2 Exon 4A	TET2_EX4A_0448	50x	ACTTCTGGATGAGCTCTCTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTTTCTTGTGTAAAGTCAGGAC
TET2 Exon 4A	TET2_EX4A_0105	1x	GTAGAGGGTATTCCAAGTGTTTGCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGAGACACCAAGTGGCACT
TET2 Exon 4A	TET2_EX4A_0106	1x	CTCTCTGGGCTCCTTCAGATNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCGTGTGAGTCCTGACTTTACACAAG
TET2 Exon 4A	TET2_EX4A_0157	1x	ACTTTTGAAAGAGTGCCACTTGGTGTCTCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCCTCCATTTTGCAAAC
TET2 Exon 4A	TET2_EX4A_0158	1x	TCTACTTTCTTGTGTAAAGTCAGGACTCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCTTTTGGTCTTGTTTC
TET2 Exon 4A	TET2_EX4A_0450	50x	TCCTCCATTTTGCAAACACTTGGANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGAAGTTACGTCTTTCTC
TET2 Exon 4A	TET2_EX4A_0451	50x	AGCCTTTTGGTCTTGTTTCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCTCACAGATTCTTTCTTATCA
TET2 Exon 4A	TET2_EX4A_0458	1x	AGCTCAGAGTTAGAGGTCTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTCTGAAAGGAACAGGTATTTA
TET2 Exon 4A	TET2_EX4A_0460	1x	CTGATTTTTGTTGTTGTAGTTGTTCTGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTGCTGCTGGAACCTGA
TET2 Exon 4A	TET2_EX4A_0462	1x	ACCGTTCAGAGCTGCCACCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTTGTGATGGTGGTGGTGGTGT
TET2 Exon 4A	TET2_EX4A_0463	25x	AGTGGCAGAAAAGGAATCCTTAGTGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTTCAGAGTGCTTTTTC
TET2 Exon 4A	TET2_EX4A_0465	1x	AGTGTTGTGTTACTTTGGTTGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTCAGAAAGCATCGGAGAAGGG
TET2 Exon 4A	TET2_EX4A_0470	1x	GGGCCTTCAATTCAATCCATCCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGGAGGTCATTTGATTGGAG
TET2 Exon 4A	TET2_EX4A_0471	1x	AGAATTGATGGCAGTGATGCCTCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCTGGGTGTAAGCTTGCCT

Supplemental Table 1(continued)

Target	Probe name	RC	Probe sequence
TET2 Exon 4A	TET2_EX4A_0472	1x	GGGAGCCCCCCAGGCATGTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGTTGGTCCACTGTACCTTGGG
TET2 Exon 4A	TET2_EX4A_0473	1x	ACTGCCCTTGATTCATTTCANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNAGTGACTGCACATGAGCTTT
TET2 Exon 4A	TET2_EX4A_0474	1x	GTAAATGGTCTGTTTTGGAGAAGTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGTTTCAACACTGGGGAC
TET2 Exon 4A	TET2_EX4A_0475	50x	AGTTTTTCAGTTTGGGAATCTGCTCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGTTTATGAGGCTTATGT
TET2 Exon 4A	TET2_EX4A_0476	1x	AGTTTGAAAATGGCTCAGTCTCTGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNATTTTTGCTGCTGTTGC
TET2 Exon 4A	TET2_EX4A_0477	1x	GGTTTTGAGGGAGATGTGAACTCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGATCCTTCTCTTTGCTG
TET2 Exon 4A	TET2_EX4A_0478	50x	TCATTGTTGCTTTGGGGGTGAGGANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTGAGTCTCGAACTCGC
TET2 Exon 4A	TET2_EX4A_0480	1x	GTACTTCCTCCAGTCCCATTTGGACANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNACTAGGTGTGTATTGT
TET2 Exon 4A	TET2_EX4A_0482	1x	AGTTCAGGATGTGTAGTCTGTTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGTTCTTGAAAGCACCTGT
TET2 Exon 4A	TET2_EX4A_0484	25x	TCTATTTTTATATCCCTGTAGAACTGANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCTGGTCAGGCACAGGA
TET2 Exon 4A	TET2_EX4A_0487	1x	AGACTCAGTTTGGGGTTGCTGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCTTTTTCCATGTTTTGTTTTC
TET2 Exon 4A	TET2_EX4A_0489	1x	TCACAGCTTGCAGGTGGATTCTCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTTTGAGAGTAAGAGCCT
TET2 Exon 4A	TET2_EX4A_0491	1x	CCCTGACATTTCAACTTTTACTTGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTGGTGTCTTTTCTGAAG
TET2 Exon 4A	TET2_EX4A_0493	1x	AGAACAGAAGCAGCTGTTCTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTCTGCAAGATGGGAAATCATA
TET2 Exon 4C	TET2_EX4C_0497	1x	AGGGCAGATTAACGTTTATCCTTTTGTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGATGAAGGTCCTTTTT
TET2 Exon 4C	TET2_EX4C_0498	1x	AGATTATTATTTTAGACCTCAATTATACTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNNCTCTAATAGCTGCCAC
TET2 Exon 5	TET2_EX5_0501	50x	TCTTGGGCATTTTGATTTGTAAATCTGANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNAGTCATCTATACTGGT

Supplemental Table 1(continued)

Target	Probe name	RC	Probe sequence
TET2 Exon 5	TET2_EX5_0502	1x	ACCCACAGAAACACACACACNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCTTAGCAATAGGACATCCCT
TET2 Exon 6	TET2_EX6_0504	1x	GGCTGCAGTGATTGTGATTCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTGTTTGGGATGGAATGGTGAT
TET2 Exon 6	TET2_EX6_0506	1x	TCTCCCCTCTTTGCGGCCACTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGACAAACTCTACTCGGAGCTT
TET2 Exon 6	TET2_EX6_0508	1x	TCACAGGTGTGGCCAGCTCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCGTATTTCCTCAGCGTCTCGG
TET2 Exon 7	TET2_EX7_0510	1x	AGCATGTACTACAATGGATGTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGCAATGAATTTGGTCTTTTGA
TET2 Exon 7	TET2_EX7_0511	1x	GGTTTGTTTACTTCCTGATGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTCTCTTTTGGTTGTTCATGG
TET2 Exon 7	TET2_EX7_0515	1x	GGATCTTGCTTCTGGCAAACNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCATCTAAGCTAATGAATTCTCT
TET2 Exon 8	TET2_EX8_0516	50x	AGAAACTTGCACCTGATGCANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTGGACTTAGAATTTAATATGT
TET2 Exon 8	TET2_EX8_0519	1x	AGGTTTTGCAAATGAGACTCCAGTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNAGTTGTTACAATTGCTGC
TET2 Exon 9	TET2_EX9_0520	1x	CTGCATGTTTGGACTTCTGTGCTCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCATTCACACACACTTTT
TET2 Exon 9	TET2_EX9_0521	1x	GAGGACAGCTTAGCAGCTGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGCCGTCCATTCTCAGGGGTCA
TP53 Exon 1	TP53_1_0102	1x	AGCAGGGAGGAGAGATGACATCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTGTCAGTCTGAGTCAGGCC
TP53 Exon 1	TP53_1_0103	1x	AGCCTCCCACCCCCATCTCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGTCCAAAAAGGGTCAGTCTAC
TP53 Exon 10	TP53_10_0144	1x	GGGGCTGGTGCAGGGGCCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGTCACAGACTTGGCTGTCCCAG
TP53 Exon 10	TP53_10_0146_SNP_a	1x	GGGACAGCATCAAATCATCCANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGAGCTGCTGGTGCAGGGGC
TP53 Exon 10	TP53_10_0146_SNP_b	1x	AGGACAGCATCAAATCATCCANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGAGCTGCTGGTGCAGGGGC
TP53 Exon 10	TP53_10_0147	50x	AGGTCCTCAGCCCCCCAGCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGTGAACCATTGTTCAATATCGTC

Supplemental Table 1(continued)

Target	Probe name	RC	Probe sequence
TP53 Exon 10	TP53_10_0148	1x	AGACTTCAATGCCTGGCCGTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCTACCAGGGCAGCTACGGTT
TP53 Exon 10	TP53_10_16882	1x	AGCCTCTGGCATTCTGGGAGCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGGTTTTCTGGGAAGGGACAG
TP53 Exon 11	TP53_11_0154	1x	GGGACCTGGAGGGCTGGGGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCAGAGACCTGTGGGAAGCGAA
TP53 Exon 12	TP53_12_0156	1x	GGGATCCAGCATGAGACACTTCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGTTTCCTGACTCAGAGGG
TP53 Exon 12	TP53_12_0026	1x	CAGAGGGGGCTCGACGCTAGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTAGGGGGCTGGGGTTGGGGT
TP53 Exon 12	TP53_12_0027	1x	GTGAGTGGATCCATTGGAAGGGCAGGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGGTCACTGCCATGGAGGAG
TP53 Exon 12	TP53_12_618883	1x	GCTCGACGCTAGGATCTGACTGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTAGGGGGCTGGGGTTGGGG
TP53 Exon 2	TP53_2_0105	1x	TCCTTGAGTTCCAAGGCCTCATTCAGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNACCTAGGAAGGCAGGG
TP53 Exon 2	TP53_2_0107	1x	GTTCAAGTTACAATTGTTTGACTTTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCTCTCGGAACATCTCG
TP53 Exon 2	TP53_2_0109	1x	GGAGCAGGGCTCACTCCAGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTCTGTTGCTGCAGATCCGTGG
TP53 Exon 3_4	TP53_4_0111	25x	AGAACCATTTTCATGCTCTCTTTAACNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCCATTTTCAACTTACAA
TP53 Exon 3_4	TP53_4_0112	50x	AGTTGGTGTTCTGAAGTTAGTTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTCTGTATCAGGCAAAGTCA
TP53 Exon 5	TP53_5_0117	1x	GCACCCTTGGTCTCCTCCACNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCCAGTGGTTTCTTCTTTGGCT
TP53 Exon 5	TP53_5_0118	1x	AGTTTCCAGTCTAACACTCANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCAACAACACCAGCTCCTCTCC
TP53 Exon 6	TP53_6_0120	1x	AGGCTCCCCTTTCTTGCGGAGANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNAAGTGAATCTGAGGCATAACT
TP53 Exon 6	TP53_6_0122	25x	AGGAAATCAGGTCCTACCTGTCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGGTCTCTCCCAGGACAGG
TP53 Exon 6	TP53_6_0124	50x	AGCTGCCCCCAGGGAGCACNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNACTGGGACGGAACAGCTTTGAG

Supplemental Table 1(continued)

Target	Probe name	RC	Probe sequence
TP53 Exon 7	TP53_7_0126	1x	GGTTCATGCCGCCCATGCANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNAGAGGCAAGCAGAGGCTGGGG
TP53 Exon 7	TP53_7_0130	1x	GGAGGCCCATCCTCACCATCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGCCTCATCTTGGGCCTGTGT
TP53 Exon 8	TP53_8_0131	50x	TCCACACGCAAATTTCCTTCCACTCGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCCTCCTCCCAGAGACC
TP53 Exon 8	TP53_8_0132	50x	AGAGGCCTGGGGACCCTGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCGAAAAGTGTTTCTGTCATCCA
TP53 Exon 9	TP53_9_0136	1x	GTGACTGCTTGTAGATGGCCATGGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNACCAGCCCTGTCGTCT
TP53 Exon 9	TP53_9_0138	1x	AGGAAGGAGACAGAGTTGAAAGTCANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCGGGCGGGGGTGTGGA
TP53 Exon 9	TP53_9_0140	50x	AGGCGCTGCCCCCACCATGANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCTGCCCTGTGCAGCTGTGGGT
U2AF1 (c34 hotspot)	U2AF1_c34_0383	50x	GTCTCCATGACGACATGCTCCANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNACTGTCTTTGAAAAGAACA
U2AF1 (c83 hotspot)	U2AF1_c83_0382	50x	GTGACGTGACTGAGCACAGTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGGTTTAATGGACAGCCGATC
WT1 Exon 1	WT1_EX1_6591	50x	AGCGCCCCCTACGCGCGGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNAGGCGTCATCCGGCCAGGCCAGG
WT1 Exon 1	WT1_EX1_6640	50x	AGGCTCCGGCTGTGCCAGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGGCTCTCGAGGCAGCTGGGCAG
WT1 Exon 1	WT1_EX1_6943	50x	GGGCCCTTCGGTCCTCCTCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCGAGCTGGGGCGGCGCGGAGC
WT1 Exon 1	WT1_EX1_6990	50x	GTTTGATGAAGGAGTGAGGCGGCGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCTCCGGCTGTGCCAGT
WT1 Exon 1	WT1_EX1_270319	50x	GGCGGCGGAGCCGGTGGCGGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGTGGACAGTGAAGGCGCTCA
WT1 Exon 10B	WT1_EX10B_0019	1x	AGTTTTACACTGGAATGGTTTCACACCTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNNTAAATGTGAAGAAAAGT
WT1 Exon 10B	WT1_EX10B_0022	1x	ACCCACACCAGGACTCATACAGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGTTAGGGCCGAGGCTAGAC
WT1 Exon 11	WT1_EX11_0016	1x	GCCAGTCAGAGACACTTGCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGATGCATGTTGTGATGGCGGA

Supplemental Table 1(continued)

Target	Probe name	RC	Probe sequence
WT1 Exon 11	WT1_EX11_0017	1x	GTTCAGTGTCCCAGGCAGCACNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNAGTTTGCCCGGTCAGATGA
WT1 Exon 3	WT1_EX3_0047	1x	GGTGCGAGGGCGTGTGACCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGGGGAGAAGGACTCCACTTGG
WT1 Exon 3	WT1_EX3_0050	50x	AGTTCCCCAACCACTCATTCANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCTGACACCTGTGCTTCTCTC
WT1 Exon 4	WT1_EX4_0044	50x	AGTGCGCCCCAAGGGCTCGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGGCTGCCGGTGCAGCTGTCGG
WT1 Exon 4	WT1_EX4_0045	50x	GTCTGGGTCCTTGGGACCCCGGANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGGTCTATGGCTGCCACAC
WT1 Exon 4	WT1_EX4_0010	1x	GCTGTCGGTGGGGGTGTGGCAGCCATAGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNNTCCCAAGGACCCAGAC
WT1 Exon 4	WT1_EX4_0011	1x	GCTGCTCACCTGCAGAGAGAACCGAAGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCCTTCCTACCTGCTGT
WT1 Exon 4	WT1_EX4_0015_SNP_a	1x	AGGCGCTGGGCTCTGCGTCTGGGTCCTTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCGGTCTATGGCTGCCA
WT1 Exon 4	WT1_EX4_0015_SNP_b	1x	AGGCGCTGGACTCTGCGTCTGGGTCCTTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCGGTCTATGGCTGCCA
WT1 Exon 5	WT1_EX5_0040	50x	CCACAAAATAATACACAACTGTTTCTTCANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNNTCATCTGATTCCAGGT
WT1 Exon 5	WT1_EX5_0041	1x	AGGTTTTCTCTATTCCATTGCCTTTCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGTTTTTCTAACAGTGAC
WT1 Exon 6	WT1_EX6_0037	1x	AGTGGGGAATGGAGCATGCATGGATCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTCTCCGCATTGTCCAC
WT1 Exon 7	WT1_EX7_0006	1x	GGGGGCCGCACTGCCTCTTCCTACTTANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNATAACCACACAACGCCCA
WT1 Exon 7	WT1_EX7_0008	1x	GCGTTGTGTGGTTATCGCTCTCGTNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCCCTTCCCGCTGGGGC
WT1 Exon 7	WT1_EX7_0033	1x	GTACCCTGTGCTGTGGCTGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGGGCCTGTCTGTGTGCTCACCT
WT1 Exon 7	WT1_EX7_0036	50x	AGAATACACACGCACGGTGTCTTCANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNACACTGAGCCTTTTTC
WT1 Exon 8	WT1_EX8_0004	1x	TTCAAGAGCTCCTTTTCCAGGCTCTCNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNATATTTTAAGCTGTCCCAC

Supplemental Table 1(continued)

Target	Probe name	RC	Probe sequence
WT1 Exon 8	WT1_EX8_0027	1x	GCGTTTCTCACTGGTCTCANNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNGGACAGCGGGCACACTTACCA
WT1 Exon 8	WT1_EX8_0029	1x	GTAGGTCTTGAGGGAGAGTGAGCACNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCGTACAAGAGTCGGGG
WT1 Exon 8	WT1_EX8_0031	1x	AGATATTTTAAGCTGTCCCACTTACAGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCGTGTGCCTGGAGTAG
WT1 Exon 9	WT1_EX9_0024	1x	AGGCAACCTCTCCTACTAGGACNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNTCGTTCACAGTCCTTGAA
WT1 Exon 9	WT1_EX9_0025	1x	AGGTTCACTTCTCATTGCTGNNNNNAGATCGGAAGAGCGTGTGTATAAGAGACAGNNNNNCAAGGTGAGAAACCATACCA

Supplemental Table 2: Index sequences for P5 adapter primers utilized for preparation of smMIP libraries, as well as barcode sequencing libraries. Index sequences present within the adapter-oligonucleotides are listed in 5'-3' orientation.

Adapter Name	Index sequence	Adapter Name	Index sequence
TruSeq_P5_xGen_001	ATATGCGC	TruSeq_P5_xGen_048	GAGCAGTA
TruSeq_P5_xGen_002	TGGTACAG	TruSeq_P5_xGen_050	GATCGAGT
TruSeq_P5_xGen_003	AACCGTTC	TruSeq_P5_xGen_051	AGCGTGTT
TruSeq_P5_xGen_004	TAACCGGT	TruSeq_P5_xGen_052	GTTACGCA
TruSeq_P5_xGen_005	GAACATCG	TruSeq_P5_xGen_053	TGAAGACG
TruSeq_P5_xGen_006	CCTTGTAG	TruSeq_P5_xGen_055	CGGTTGTT
TruSeq_P5_xGen_007	TCAGGCTT	TruSeq_P5_xGen_056	GTTGTTCG
TruSeq_P5_xGen_008	GTTCTCGT	TruSeq_P5_xGen_057	GAAGGAAG
TruSeq_P5_xGen_009	AGAACGAG	TruSeq_P5_xGen_058	AGCACTTC
TruSeq_P5_xGen_010	TGCTTCCA	TruSeq_P5_xGen_059	GTCATCGA
TruSeq_P5_xGen_011	CTTCGACT	TruSeq_P5_xGen_060	TGTGACTG
TruSeq_P5_xGen_012	CACCTGTT	TruSeq_P5_xGen_061	CAACACCT
TruSeq_P5_xGen_013	ATCACACG	TruSeq_P5_xGen_062	ATGCCTGT
TruSeq_P5_xGen_014	CCGTAAGA	TruSeq_P5_xGen_063	CATGGCTA
TruSeq_P5_xGen_015	TACGCCTT	TruSeq_P5_xGen_064	GTGAAGTG
TruSeq_P5_xGen_016	CGACGTTA	TruSeq_P5_xGen_065	CGTTGCAA
TruSeq_P5_xGen_017	ATGCACGA	TruSeq_P5_xGen_066	ATCCGGTA
TruSeq_P5_xGen_018	CCTGATTG	TruSeq_P5_xGen_067	GCGTCATT
TruSeq_P5_xGen_019	GTAGGAGT	TruSeq_P5_xGen_068	GCACAACT
TruSeq_P5_xGen_021	CACTAGCT	TruSeq_P5_xGen_069	GATTACCG
TruSeq_P5_xGen_023	CGTGTGTA	TruSeq_P5_xGen_071	GTCGAAGA
TruSeq_P5_xGen_024	GTTGACCT	TruSeq_P5_xGen_072	CCTTGATC
TruSeq_P5_xGen_026	CAATGTGG	TruSeq_P5_xGen_073	AAGCACTG
TruSeq_P5_xGen_027	TTGCAGAC	TruSeq_P5_xGen_074	TTCGTTGG
TruSeq_P5_xGen_028	CAGTCCAA	TruSeq_P5_xGen_075	TCGCTGTT
TruSeq_P5_xGen_030	AACGTCTG	TruSeq_P5_xGen_076	GAATCCGA
TruSeq_P5_xGen_031	TATCGGTC	TruSeq_P5_xGen_077	GTGCCATA
TruSeq_P5_xGen_032	CGCTCTAT	TruSeq_P5_xGen_078	CTTAGGAC
TruSeq_P5_xGen_033	GATTGCTC	TruSeq_P5_xGen_079	AACTGAGC
TruSeq_P5_xGen_034	GATGTGTG	TruSeq_P5_xGen_080	GACGATCT
TruSeq_P5_xGen_035	CGCAATCT	TruSeq_P5_xGen_081	ATCCAGAG
TruSeq_P5_xGen_036	TGGTAGCT	TruSeq_P5_xGen_082	AGAGTAGC
TruSeq_P5_xGen_037	GATAGGCT	TruSeq_P5_xGen_083	TGGACTCT
TruSeq_P5_xGen_038	AGTGGATC	TruSeq_P5_xGen_084	TACGCTAC
TruSeq_P5_xGen_039	TTGGACGT	TruSeq_P5_xGen_085	GCTATCCT
TruSeq_P5_xGen_040	ATGACGTC	TruSeq_P5_xGen_086	GCAAGATC
TruSeq_P5_xGen_041	GAAGTTGG	TruSeq_P5_xGen_087	ATCGATCG
TruSeq_P5_xGen_042	CATACCAC	TruSeq_P5_xGen_088	CGGCTAAT
TruSeq_P5_xGen_043	CTGTTGAC	TruSeq_P5_xGen_090	CGCATGAT
TruSeq_P5_xGen_044	TGGCATGT	TruSeq_P5_xGen_091	TTCCAAGG
TruSeq_P5_xGen_045	ATCGCCAT	TruSeq_P5_xGen_092	CTTGTCGA
TruSeq_P5_xGen_046	TTGCGAAG	TruSeq_P5_xGen_093	GAGACGAT
TruSeq_P5_xGen_047	AGTTCGTC	TruSeq_P5_xGen_094	TGAGCTAG

Supplemental Table 2 (continued)

Adapter Name	Index sequence	Adapter Name	Index sequence
TruSeq_P5_xGen_096	CTGATCGT	TruSeq_P5_xGen_143	CGTACGAA
TruSeq_P5_xGen_097	CGACCATT	TruSeq_P5_xGen_144	GACTTAGG
TruSeq_P5_xGen_098	GATAGCGA	TruSeq_P5_xGen_145	AGTGCAGT
TruSeq_P5_xGen_099	AATGGACG	TruSeq_P5_xGen_146	TTGATCCG
TruSeq_P5_xGen_100	CGCTAGTA	TruSeq_P5_xGen_147	TGCCATTC
TruSeq_P5_xGen_101	TCTCTAGG	TruSeq_P5_xGen_148	CTTGCTGT
TruSeq_P5_xGen_103	TGAGGTGT	TruSeq_P5_xGen_149	CCTACTGA
TruSeq_P5_xGen_104	AATGCCTC	TruSeq_P5_xGen_150	CCAAGTTG
TruSeq_P5_xGen_105	CTGGAGTA	TruSeq_P5_xGen_152	TAGTTGCG
TruSeq_P5_xGen_106	GTATGCTG	TruSeq_P5_xGen_153	GTCTGATC
TruSeq_P5_xGen_107	TGGAGAGT	TruSeq_P5_xGen_154	CGTTATGC
TruSeq_P5_xGen_108	CGATAGAG	TruSeq_P5_xGen_155	GCTCTGTA
TruSeq_P5_xGen_109	CTCATTGC	TruSeq_P5_xGen_156	TTACCGAG
TruSeq_P5_xGen_111	GAATCGTG	TruSeq_P5_xGen_157	GCCATAAC
TruSeq_P5_xGen_112	AGGCTTCT	TruSeq_P5_xGen_158	CTCAGAGT
TruSeq_P5_xGen_113	CAGTTCTG	TruSeq_P5_xGen_159	CGAGACTA
TruSeq_P5_xGen_114	TTGGTGAG	TruSeq_P5_xGen_160	TGTGCGTT
TruSeq_P5_xGen_115	CATTCGGT	TruSeq_P5_xGen_161	TTCAGGAG
TruSeq_P5_xGen_116	TGTGAAGC	TruSeq_P5_xGen_162	GACTATGC
TruSeq_P5_xGen_117	TAAGTGGC	TruSeq_P5_xGen_163	AGGTTCGA
TruSeq_P5_xGen_119	GTAGAGCA	TruSeq_P5_xGen_164	AGTCTGTG
TruSeq_P5_xGen_120	GTCAGTTG	TruSeq_P5_xGen_166	TGCAGGTA
TruSeq_P5_xGen_121	ATTCGAGG	TruSeq_P5_xGen_167	AAGGACAC
TruSeq_P5_xGen_122	GATACTGG	TruSeq_P5_xGen_168	CAACCTAG
TruSeq_P5_xGen_123	GCCTTGTT	TruSeq_P5_xGen_169	CTGACACA
TruSeq_P5_xGen_124	TTGGTCTC	TruSeq_P5_xGen_171	AGCTCCTA
TruSeq_P5_xGen_125	CCGACTAT	TruSeq_P5_xGen_172	TACATCGG
TruSeq_P5_xGen_126	GTCCTAAG	TruSeq_P5_xGen_173	CACAAGTC
TruSeq_P5_xGen_128	GATGCACT	TruSeq_P5_xGen_174	CGGATTGA
TruSeq_P5_xGen_129	GCTGGATT	TruSeq_P5_xGen_175	AGTCGACA
TruSeq_P5_xGen_130	ATGGTTGC	TruSeq_P5_xGen_176	GTCTCCTT
TruSeq_P5_xGen_131	CAGAATCG	TruSeq_P5_xGen_177	GAGATACG
TruSeq_P5_xGen_132	GAACGCTT	TruSeq_P5_xGen_178	ATCGGTGT
TruSeq_P5_xGen_133	TCGAACCA	TruSeq_P5_xGen_179	TCTCGCAA
TruSeq_P5_xGen_134	CTATCGCA	TruSeq_P5_xGen_180	TCTAACGC
TruSeq_P5_xGen_135	TACGGTTG	TruSeq_P5_xGen_181	CAATCGAC
TruSeq_P5_xGen_136	GAGATGTC	TruSeq_P5_xGen_182	GAGGACTT
TruSeq_P5_xGen_137	CTTACAGC	TruSeq_P5_xGen_183	TGGAGTTG
TruSeq_P5_xGen_138	AGGAGGAA	TruSeq_P5_xGen_184	CTAGGCAT
TruSeq_P5_xGen_139	GACGAATG	TruSeq_P5_xGen_185	CTCTACTC
TruSeq_P5_xGen_140	GAAGAGGT	TruSeq_P5_xGen_186	AGAAGCGT
TruSeq_P5_xGen_141	CGTCAATG	TruSeq_P5_xGen_187	TCGAAGGT
TruSeq_P5_xGen_142	TACCAGGA	TruSeq_P5_xGen_188	GTCGGTAA

Supplemental table 2 (continued)

Adapter Name	Index sequence	Adapter Name	Index sequence
TruSeq_P5_xGen_190	TCCGTATG	TruSeq_P5_xGen_202	TATCAGCG
TruSeq_P5_xGen_191	CTAGGTGA	TruSeq_P5_xGen_203	AGCAGATG
TruSeq_P5_xGen_192	CATTGCCT	TruSeq_P5_xGen_204	AACGGTCA
TruSeq_P5_xGen_194	TCGTGGAT	TruSeq_P5_xGen_205	CGAACTGT
TruSeq_P5_xGen_195	GTTCATGG	TruSeq_P5_xGen_206	TCCGAGTT
TruSeq_P5_xGen_196	TAGGATGC	TruSeq_P5_xGen_207	TTCTCTCG
TruSeq_P5_xGen_197	CATGGAAC	TruSeq_P5_xGen_208	ATTCTGGC
TruSeq_P5_xGen_198	GCTTAGCT	TruSeq_P5_xGen_210	CATAACGG
TruSeq_P5_xGen_199	CTAACTCG	TruSeq_P5_xGen_211	CAGTCTTC
TruSeq_P5_xGen_201	TCAGACGA	TruSeq_P5_xGen_212	TGCCTCTT

Supplemental Table 3: Index sequences of P7 adapter primers utilized for preparation of smMIP libraries. Index sequences present within the adapter-oligos are listed in 5'-3' orientation.

Adapter Name	Index (5'-3')	Adapter Name	Index (5'-3')
NT+RS_P7_xGen_2	TCGAGAGT	NT+RS_P7_xGen_36	AGGTGTTG
NT+RS_P7_xGen_3	CTAGCTCA	NT+RS_P7_xGen_37	CAGTCACA
NT+RS_P7_xGen_5	TCGACAAG	NT+RS_P7_xGen_38	TCGATGAC
NT+RS_P7_xGen_6	CCTTGGAA	NT+RS_P7_xGen_39	GAAGTGCT
NT+RS_P7_xGen_7	ATCATGCG	NT+RS_P7_xGen_40	CTTCCTTC
NT+RS_P7_xGen_8	TGTTCCGT	NT+RS_P7_xGen_41	CGAACAAC
NT+RS_P7_xGen_9	ATTAGCCG	NT+RS_P7_xGen_44	CGTCTTCA
NT+RS_P7_xGen_10	CGATCGAT	NT+RS_P7_xGen_45	TGCGTAAC
NT+RS_P7_xGen_11	GATCTTGC	NT+RS_P7_xGen_47	ACTCGATC
NT+RS_P7_xGen_12	AGGATAGC	NT+RS_P7_xGen_48	TGAGCTGT
NT+RS_P7_xGen_13	GTAGCGTA	NT+RS_P7_xGen_50	GACGAACT
NT+RS_P7_xGen_14	AGAGTCCA	NT+RS_P7_xGen_51	CTTCGCAA
NT+RS_P7_xGen_15	GCTACTCT	NT+RS_P7_xGen_52	ATGGCGAT
NT+RS_P7_xGen_16	CTCTGGAT	NT+RS_P7_xGen_53	ACATGCCA
NT+RS_P7_xGen_17	AGATCGTC	NT+RS_P7_xGen_54	GTCAACAG
NT+RS_P7_xGen_18	GCTCAGTT	NT+RS_P7_xGen_55	GTGGTATG
NT+RS_P7_xGen_19	GTCCTAAG	NT+RS_P7_xGen_57	GACGTCAT
NT+RS_P7_xGen_21	TCGGATTC	NT+RS_P7_xGen_59	GATCCACT
NT+RS_P7_xGen_23	CCAACGAA	NT+RS_P7_xGen_60	AGCCTATC
NT+RS_P7_xGen_24	CAGTGCTT	NT+RS_P7_xGen_61	AGCTACCA
NT+RS_P7_xGen_25	GATCAAGG	NT+RS_P7_xGen_62	AGATTGCG
NT+RS_P7_xGen_26	TCTTCGAC	NT+RS_P7_xGen_63	CACACATC
NT+RS_P7_xGen_28	CGGTAATC	NT+RS_P7_xGen_64	GAGCAATC
NT+RS_P7_xGen_29	AGTTGTGC	NT+RS_P7_xGen_65	ATAGAGCG
NT+RS_P7_xGen_30	AATGACGC	NT+RS_P7_xGen_66	GACCGATA
NT+RS_P7_xGen_32	TTGCAACG	NT+RS_P7_xGen_67	CAGACGTT
NT+RS_P7_xGen_33	CACTTCAC	NT+RS_P7_xGen_68	CTGAACGT
NT+RS_P7_xGen_34	TAGCCATG	NT+RS_P7_xGen_69	TTGGACTG
NT+RS_P7_xGen_35	ACAGGCAT	NT+RS_P7_xGen_70	GTCTGCAA

Supplemental Table 3 (continued)

Adapter Name	Index (5'-3')	Adapter Name	Index (5'-3')
NT+RS_P7_xGen_71	CCACATTG	NT+RS_P7_xGen_110	GTATCGAG
NT+RS_P7_xGen_72	GATGGAGT	NT+RS_P7_xGen_113	GTCAGTCA
NT+RS_P7_xGen_73	AGGTCAAC	NT+RS_P7_xGen_114	CACGTCTA
NT+RS_P7_xGen_74	TACACACG	NT+RS_P7_xGen_119	TATGACCG
NT+RS_P7_xGen_75	CAAGTCGT	NT+RS_P7_xGen_127	GTGATCCA
NT+RS_P7_xGen_76	AGCTAGTG	NT+RS_P7_xGen_128	ACTGGTGT
NT+RS_P7_xGen_78	ACTCCTAC	NT+RS_P7_xGen_144	GTTCTTCG
NT+RS_P7_xGen_80	TCGTGCAT	NT+RS_P7_xGen_160	TGAGACGA
NT+RS_P7_xGen_81	TAACGTCG	NT+RS_P7_xGen_200	TTACGTGC
NT+RS_P7_xGen_83	TCTTACGG	NT+RS_P7_xGen_242	TCACTCGA
NT+RS_P7_xGen_84	CGTGTGAT	NT+RS_P7_xGen_267	GCCAATAC
NT+RS_P7_xGen_85	AACAGGTG	NT+RS_P7_xGen_287	ATCCACGA
NT+RS_P7_xGen_87	TGGAAGCA	NT+RS_P7_xGen_295	ACGCTTCT
NT+RS_P7_xGen_90	AAGCCTGA	NT+RS_P7_xGen_324	GTTATGGC
NT+RS_P7_xGen_92	CGATGTTC	NT+RS_P7_xGen_330	TCCGATCA
NT+RS_P7_xGen_94	GAACGGTT	NT+RS_P7_xGen_332	TCAGTAGG
NT+RS_P7_xGen_97	TGATAGGC	NT+RS_P7_xGen_364	GCCACTTA
NT+RS_P7_xGen_100	CTGATGAG	NT+RS_P7_xGen_379	CGCAATGT
NT+RS_P7_xGen_104	TGCTGTGA	NT+RS_P7_xGen_380	CCTAGAGA

Supplemental Table 4: Reagents costs for preparation of smMIP libraries. The current workflow results in costs of about 3.30€ per library an 6.59€ per sample, if libraries are prepared in technical duplicates. As the ordered amounts of smMIPs oligonucleotides are sufficient for millions of reactions, the minimal costs per reaction are below 0.01€ and hence not included in this table. With initial costs of about 3,400€ for synthesis of all smMIP probes, the additional costs per reaction are only at 3€ per library or 6€ per sample already after the first pilot experiment.

Reagent name	Vendor	Cat.No.	Official price	Costs per library	Costs per sample
Hemo KlenTaq	NEB	M0332L	452.00 €	0.07 €	0.14 €
Ampligase DNA Ligase + Buffer	Biozym	111250	301.00 €	0.12 €	0.24 €
Exonuclease I	NEB	M0293L	294.00 €	0.20 €	0.39 €
Exonuclease III	NEB	M0206L	260.00 €	0.52 €	1.04 €
Q5 Hot Start High-Fidelity DNA Polymerase	NEB	M0493L	534.00 €	2.14 €	4.27 €
Deoxynucleotide Solutions, Mix [large]	NEB	N0447L	255.00 €	0.36 €	0.71 €
Illumina Adapter Primers	SigmaAldrich	-	21.50€ (50nmol)	0.04 €	0.09 €
SPRI beads [homemade]	-	-	22.86€ (50ml)	0.07 €	0.15 €
			Total costs:	**3.51 €**	**7.04 €**

10 Acknowledgements

I want to thank Prof. Dr. Wolfgang Enard for the opportunity to work on this project and his dedicated support and supervision. A big thanks also goes to Dr. Ines Hellmann, Dr. Beate Vieth and Dr. Swati Parekh for their help with any computational issues.

I am also very grateful for the great time with my colleagues with who with me: Dr. Aleks Janjic, Johanna Geuder, Lucas Wange and Dr. Johannes Bagnoli. A special thanks goes out to Dr. Johannes Bagnoli for support, ideas and fruitful discussion on all kinds of wetlab and drylab related problems and mutual help in any tasks.

Similarly, I want to thank Dr. Christoph Ziegenhain for his support and productive feedback on any wetlab related issues!

Another special thanks befits Simon Krauss, who collaborated with me for establishing the smMIP assay within his research project, master thesis and beyond. I wish you all the best for your PhD in Leipzig and many interesting findings using our 'CHIPMIP' panel in future studies!

Additionally I like to thank the whole work group for a social and productive research environment encouraging cooperation and each other's support.

I also want to thank my collaborators at the Helmholtz Zentrum München. Prof. Dr. Irmela Jeremias for providing the excellent Pdx mouse model and support for planning experimental designs. Dr. Christina Zeller for conducting all cell culture and mouse related experiments of the cellular barcoding project. And last but not least, thank you Dr. Binje Vick for your dedicated support within the cellular barcoding project.

Likewise, I want to thank the collaborators at the Klinikum der Universität München Prof. Dr. Klaus Metzeler as well as his former work group members Dr. Maja Rothenburg-Thurley, Dr. Frank Ziemann, Dr. Luise Hartmann and Eva Telzerow for providing AML patient samples as well as samples from control individuals, cross-checking the smMIP variants with clinical data from the Haloplex panel and many productive meetings discussing the state and performance of the new assay as well as the detected variants.

Last but not least I want to thank my family as well as Larissa for their continuous support during the course of my PhD!