Tracking the Evolution of Therapy-Related Myeloid Neoplasms Using Chemotherapy Signatures

SUPPLEMENTAL METHODS

Study cohort

The WGS cohort was compiled with both newly sequenced and publicly available data (**Supplemental Table 1**). Clinical records at Memorial Sloan Kettering Cancer Center (MSKCC) were screened to identify adult patients who had developed tMN following exposure to either of high-dose melphalan- or platinum-containing anti-neoplastic regimens. 18 tMN were eligible for sequencing (**Supplemental Tables 1-2**). Eight patients were exposed to melphalan as the sole cytotoxic therapy. The remaining 22 tMN genomes (from 21 patients) were imported from public datasets[1,2]. 21 *de novo* AML whole genomes (including three relapse samples) were imported from the TCGA[3] (dbGaP 000178) as comparators. 298 *de novo* AML and 22 tMN whole exomes were imported from the Beat AML dataset (dbGaP: phs001657). Additionally, non-myeloid secondary malignancies were sequenced: clinical records at MSKCC were screened to identify patients with hematologic and solid tumors that developed following exposure to melphalan. Specifically, we sequenced 5 patients with B-ALL following melphalan/ASCT and one patient with a secondary bladder tumor following exposure to melphalan/ASCT. The latter tumor was chosen for sequencing because unchanged melphalan is partially excreted in urine[4], thus putting urothelium in direct contact with the mutagen and here providing evidence of clonal expansion of a single melphalan-exposed urothelial cell eight years following exposure

In addition, we searched the MSKCC biobank for patients with secondary malignancies with multiple chromosomal gains following treatment with platinum-based chemotherapy.

All samples and data were obtained and managed in accordance with the Declaration of Helsinki and the Institutional Review Board of MSKCC under protocols 14-276 and 15-017. The study involved the use of human samples, which had been collected after written informed consent had been obtained.

Sequencing and Analytical Methods

A detailed description of the sequencing and analytical methods is provided in the **Supplemental Methods**. Briefly, for newly identified samples, following sample purity optimization and quality control, tumor samples and matched normal had WGS performed with an Illumina NovaSeq 6000 with target depth of 70x for tumor and 40x for normal. A variety of published tools were used for alignment, somatic mutations, indels, copy number, structural variation, phylogeny, and mutational signatures as previously described (**Supplemental Methods**)[5,6]. Target capture-based sequencing (MSK-IMPACT) was performed on available peripheral blood or CD34+ selected autograft

products collected prior to tMN diagnosis for identification of antecedent CH. Mutation calling followed stringent criteria using both variant callers and direct visualization of mutations with Integrated Genome Viewer[7]. Cell culture, transfection, and cytokine independence assay were performed as detailed by the Taylor Lab at the University of Miami.

Sample preparation. Tumor samples for in-house sequencing were curated from three sources. Tumors from high purity samples (and solid tumors) that had previously had clinical molecular testing performed and had leftover cDNA directly sequenced. The two secondary multiple myeloma samples were sequenced directly from leftover DNA extracted from CD138-sorted plasma cell used for clinical SNP-array. For the remaining, frozen bone marrow mononuclear cell aliquots were obtained from an in-house biobank and were either sequenced directly following DNA extraction or were first sorted via fluorescence-activated cell sorting to improve sample purity. Matched normal were selected from frozen peripheral blood mononuclear cells, peripheral blood granulocytes, or CD34-selected autograft products were used as normal match (**Supplemental Table 1**). The remaining 22 tMN genomes (from 21 patients) were imported from public datasets[1,2] (dbGaP: phs000159 and EGAD00001005028). The imported genomes were analyzed via the same pipeline as the in-house samples as follows.

Whole-genome sequencing. Following quantification via PicoGreen and quality control by Agilent Bioanalyzer, ~500 ng of genomic DNA was sheared (LE220-plus Focused-ultrasonicator; Covaris, catalog no., 500569) and sequencing libraries were prepared using a modified KAPA Hyper Prep Kit (Kapa Biosystems, KK8504). Briefly, libraries were subjected to a 0.5 × size select using aMPure XP Beads (Beckman Coulter, catalog no., A63882) after post-ligation cleanup. Libraries that were not amplified by PCR (07652_C) were pooled equivolume. Libraries amplified with five cycles of PCR (07652_D, 07652_F, and 07652_G) were pooled equimolar. Samples were run on a NovaSeq 6000 in a 150 bp/150 bp paired-end run, using the NovaSeq 6000 SBS v1 kit and an S4 Flow Cell (Illumina), as described previously[8]. Target coverage depth was 70x for tumor and 40x for normal.

Whole-genome analysis pipeline. Coverage for tumor and normal samples are reported in **Extended Data Table 1**. Short insert paired-end reads were aligned to the reference genome (GRCh37) using the Burrows–Wheeler Aligner (v0.5.9; ref. 17). All samples were uniformly analyzed by the following bioinformatic tools: somatic mutations were identified by CaVEman[8]; copy number analysis and tumor purity (i.e., cancer cell fraction) were evaluated using Battenberg (https://github.com/Wedge-Oxford/battenberg); structural variants were defined by BRASS (https://github.com/cancerit/BRASS) via discordant mapping of paired-end reads, passed through additional quality filters, and were manually curated to define complex events (i.e., templated insertions, chromothripsis, and chromoplexy) as described previously[9]. The phylogenetic tree of each case was reconstructed using Pyclone-VI (https://github.com/Roth-Lab/pyclone-vi) to determine clonal and subclonal variants.

The exomes data downloaded from the public repository were aligned to the reference human genome (GRCh37) using Burrows-Wheeler Aligner, BWA (v0.7.17). Deduplicated aligned BAM files were analyzed using FACETS (v0.5.6, https://github.com/mskcc/facets) for copy number variants, CaVEMan (v1.13.14, https://github.com/cancerit/dockstore-cgpwxs) for single nucleotide variants (SNVs) and Pindel (v3.2.0, https://github.com/cancerit/dockstore-cgpwxs) for small insertions-deletions.

The genome regions that were significantly modified in our cohort were identified by using GISTIC2.0 (v2.0.23, https://www.genepattern.org). To improve the test's statistical power, we ran our cohort of myeloid whole genomes (n=57; not including relapse cases) with Beat-AML samples (n=320). In this way we were able to detect the anomalous peaks and arms shared among all the sample. The analysis was executed using Gene Pattern web interface (http://genepattern.broadinstitute.org) and setting a q value threshold of 0.01. For further comparison, samples were split by status as *de novo AML*, tMN with chemotherapy mutational signature (i.e., chemotherapy-induced mutagenesis), and tMN without chemotherapy mutational signatures. Because mutational signatures were not run for exomes, Beat-AML tMN samples were excluded from this final comparison (n=22).

We applied the dN/dScv method to detect genes under positive selection in our cohort[10]. To increase the statistical power, we included 320 *de novo* AML and tMN samples from the Beat-AML study. Importantly, the original Beat AML study called mutations with Varscan and MuTect2, which are not as specific as CaVEMan[11]. For mutational signatures from genomic data, we retained our pipeline as described. For the comparison of the driver mutation landscape, we rescued mutations that had been filtered out by CaVEMan that had been called in the original Beat AML study[12] by Varscan and MuTect2.

Mutational Signatures. Mutational signatures were analyzed across all whole genomes. To estimate the activity of mutational signatures, we first employed a three step process of de novo extraction, assignment, and fitting[1]. For the first step, we ran SigProfiler for SBS, DBS, and ID signatures[13]. All extracted signatures were then compared with the latest Catalogue of Somatic Mutations in Cancer (COSMIC) reference (https://cancer.sanger.ac.uk/cosmic/signatures/SBS) to identify the known mutational processes active in the cohort. In the case of ID signatures, the deconvolution and fitting solution was accepted outright. For SBS and DBS signatures, we required the addition of signatures not currently included in the most recent version (3.2) of the COSMIC catalogue (**Supplemental Table 4**). These are SBS-MM1 (melphalan)[6], SBS-HSC (clock-like signature in hematopoietic cells)[14], E-DBS3 and E-DBS9 (platinum)[15]. We performed an adjusted deconvolution with the respective SBS and DBS COSMIC catalogues with the addition of these four signatures using a bespoke algorithm (https://github.com/UM-Myeloma-Genomics/Signature-Assignment)[1]. The code generates a pairwise fitting contribution of user-supplied reference mutational signatures to *de novo* extracted signatures and is particularly useful for the addition and evaluation of signatures not included in the COSMIC reference. The top deconvolution combination with biologic rationale reflective of signatures known to be active in included tumor histologies was chosen for each *de novo* signature extraction unless the SigProfiler solution was more

appropriate. Deconvolution revisions are marked with an asterisk in **Supplemental Figures 2-4** and reported in **Supplemental Tables 6 and 7**[1]. For SBS, we applied mmsig (https://github.com/UM-Myeloma-Genomics/mmsig)[16], a fitting algorithm, to confirm the presence and estimate the contribution of each mutational signature in each sample guided by the catalog of signatures extracted for each individual sample by SigProfiler's de novo refit, our revised deconvolution, and with revisions for processes known or not known to be active in disease histologies (i.e., addition of SBS8 for multiple myeloma samples, removal of flat SBS40 signature in all cases in favor of SBS5 and SBS-HSC). mmsig confidence intervals were generated by bootstrapping 1,000 mutational profiles from the multinomial distribution each time repeating the signature fitting procedure, and finally taking the 2.5th and 97.5th percentile for each signature. At least 40 mutations were required per sample for this analysis. For DBS and indels, we similarly used a modified version of mmsig capable of fitting our catalogue of DBS and indel signatures to each sample. Given the small number of DBS and indels, each sample's process catalogue was based on Sigprofiler's de novo refit and our pairwise deconvolution, and then fit with expectation maximization using mmsig.

Chemotherapy-Related Mutational Signatures in Transitional Cell Carcinoma and in tMN Treated with Oral Melphalan. SBS signature de novo extraction for a single sample is technically ill-advised. We imported mutational signatures for the PCAWG cohort of Transitional Cell Carcinoma (n=23)[13] and then we used MutationalPatterns plot_compare_profiles and cos_sim functions (https://github.com/UMCUGenetics/MutationalPatterns) to compare the case 96-profile with de novo signatures extracted by Sigprofiler. The difference in mutations was quantified and then compared directly to the SBS-MM1 mutational signature using cosine similarity (**Supplemental Figure 7b**). A similar approach was applied for tMN exposed only to oral melphalan in the absence of ASCT (**Supplemental Figure 7a**), in which SBS-MM1 was not previously detected[12,18]. Mutational signatures were first extracted in the tumor and then for 18 *de novo* AML in the cohort. The difference in mutational profiles was then ascertained and compared to SBS-MM1 using cosine similarity.

Chemotherapy-Related Mutational Signatures in Driver Genes of Platinum-Exposed Individuals. Non-synonymous SNV in leukemic drivers from WGS of platinum-exposed tMN (6 mutations, 5 patients) in our cohort were pooled with CH mutations from 944 cancer patients exposed to platinum chemotherapies from Bolton et al.[17] After excluding driver genes with less than 10 non-synonymous SNV[18] and including only one mutation at a given chromosomal position to remove any hotspot bias, post-platinum SNV in driver genes totaled n= 749 for 19 driver genes defined by *dndscv* (see above). Signatures were then fit with mmsig expectation–maximization function for each gene to reveal the contribution of platinum mutational signatures to their respective mutational profiles. To avoid overfitting and false positivity calling, we sacrificed some sensitivity by fitting only the chemotherapy mutational signature peaks associated with strand bias to ensure specificity in this gene-level analysis[16].

Target capture-based sequencing. Peripheral blood samples and/or CD34-selected autograft products collected prior to chemotherapy exposure were collected from eleven patients with matched whole genomes (**Supplemental Table 1, 14**). Where possible, apheresis product was prioritized. Samples were sequenced using MSK-IMPACT, a Food and Drug Administration-authorized hybridization capture-based next-generation sequencing assay of protein-coding exons from 505 known cancer-associated genes (**Supplemental Table 14**)[17,19]. Matched normal was obtained by pooling IMPACT-505 data from 8 healthy individuals.

Target capture-based sequencing variant calling and filtering. Given the potential for extremely low VAF (i.e., small clone size) for CH mutations in normal samples preceding tumor expansion by many years, population-based CH screening techniques[17] would fail to capture many low allelic frequency variants of interest and so directed mutation query was employed. Our approach consisted of two distinct strategies to characterize antecedent CH variants. Our first approach was targeted-sequencing-centric and was a modified workflow from Bolton et al[17]. First, stringent quality filters were applied to calls from a triple caller pipeline of Mutect, Strelka, and Caveman[8,20,21]. We required a SNV to be called by two or more callers, have a VAF of >0.02, have passed default quality flags, and not result in synonymous substitution. We further required at least 10 supporting forward and reverse reads in Mutect, and to further filter any germline SNPs, we removed any variants reported for any population in the gnomAD database at a frequency greater than 0.005. Indels were called with Pindel [22] and considered if they passed all quality flags. Further postprocessing filters to remove sequencing artifacts were employed as per Bolton et al.[17] However, as we also had WGS data, we were also able to work in the reverse direction: for the list of all nonsynonynous mutations and indels identified in driver genes, each individual variant was queried directly in targeted sequencing bam files using Integrated Genome Viewer[7] and following pileup at target panel loci. If at least one read for the mutation was identified, the mutation was considered present in the target sequencing sample. Variants were further confirmed and VAF was calculated by generating a pileup of reads for the targeted sequencing regions with SAMtools[23] for both reference and alternate alleles.

Germline Susceptibility Variants. To ascertain whether single nucleotide polymorphisms in germline susceptibility variants may have played a role in tMN development, we first compile a list of vetted variants from a comprehensive review performed by Takahashi[24]. A pileup was performed on bam files from matched normal and further downstream filtering was applied to limit our investigation to high quality variant calls. We removed calls with more than one alternate allele, required that calls be supported by both forward and reverse reads, removed synonymous variants, and removed calls with an alternate allele with unclear consequence. We also filtered out variants with high RPB, MQB, BQBm and MQOF. We further calculated strand bias for each variant with a Fisher test; with those displaying strand bias more likely to be false positive calls. Finally, we selected only variants with VAF>0.25 to ensure SNP status. Although single nucleotide polymorphisms were identified in genes involved in DNA damage response pathways, acknowledging samples size constraints, there was no

enrichment for samples with chemotherapy-induced mutagenesis (**Supplemental Table 16**).

Cell culture, *SMARCA4* transfection and cytokine independence assay. Ba/F3 cells were gifted to Dr. Taylor from Dr. Omar Abdel-Wahab (Memorial Sloan Kettering Cancer Center). The cells were cultured in RPMI media supplemented with 10% FBS, 1% Penicillin/Streptomycin and 10 µg/mL of mouse IL3 (mIL3) and were maintained at 37ºC and 5% CO_2. pQCXIH *BRG1* was a gift from Joan Massague (Addgene plasmid # 19148; RRID:Addgene_19148). The *BRG1/SMARCA4* plasmid was linearized with SalI and transfected into Ba/F3 cells via electroporation (Neon transfection system, ThermoFisher, Waltham, MA). Hygromycin was utilized to select for *SMARCA4* overexpressing Ba/F3 cells and overexpression was confirmed by immunoblotting with BRGI/SMARCA4 antibody (Cell Signaling E906E, 1:500 dilution). *SMARCA4* or vector expressing Ba/F3 cells were then cultured in media without IL3. Cells were seeded in triplicates at a starting concentration of 100,000 cells/mL and were counted daily using Vi-Cell BLU automated cell counter (Beckman Coulter, Indianapolis, IN) and plotted using GraphPad Prism Version 9 Software.

Molecular Time and Absolute Timing of Gains. The relative timing of large chromosomal duplications (multi-gain events) was estimated with the mol_time function (https://github.com/UM-Myeloma-Genomics/mol_time). As previously described, this approach allows for the relative timing of gains of large chromosomal segments[6,25] by using the corrected ratio of clonal single nucleotide variants duplicated or non-duplicated across the gains: VAF 66% if duplicated and found on two alleles (pre-gain mutation) and VAF 33% if non-duplicated and found on one allele (post-gain mutation). VAFs were corrected for sample purity (i.e., cancer cell fraction) by combining Battenberg's estimation of tumor purity and the density and distribution of SNV VAF within clonal diploid regions of each sample genome. We required that gains have a minimum of 40 clonal mutations to be included in analysis for accuracy concerns[6]. For the purposes of increasing power to accurately detect mutational signatures, gains occurring in an overlapping time window estimate were collapsed (i.e, multi-gain event within the same molecular time). For myeloid cases, this provided evidence of relative timing of gain with respect to tumor diagnosis and corroborated results of the following duplicated mutation analysis.

For two multiple myeloma cases, mutational signatures were quantified for pre- and post-gain mutations. The mutational burden of clock-like mutations (SBS5) was then used to ascertain each patient's individual SBS5 mutation rate. This was accomplished by pooling the two secondary multiple myeloma tumors and the multiple myeloma longitudinal cohort from Rustad et al., 2020 to use a linear mixed effects model to estimate SBS5 accumulation over time[6,26]. Then, to convert molecular time estimates into absolute time: i) the most recent common ancestor was estimated dividing clonal SBS5 mutations by the individual SBS5 mutation rate with interval of confidence derived from the upper and lower bounds of standard deviation for the patient-specific mutation rate; ii) the multi-gain event was estimated by dividing the number of clonal SBS5 pre-gain mutations by

the mutation rate corrected for size of the gains and the interval of confidence derived from the mutation rate as above; iii) the multi-gain event was additionally estimated by multiplying the MRCA by the molecular time estimate with interval of confidence generated from bootstrapping the molecular time estimate 1000 times. Only gains with at least 40 clonal mutations were included for the purpose of accuracy.

Duplicated Mutational Burden and Chemotherapy Barcoding. In an approach similar to that used for estimated timing of gains using duplicated clock-like mutations[6,25], we leveraged mutations that were duplicated across gains (i.e., VAF 66%) to determine variants that predated the copy number gain. VAFs were corrected for sample purity (i.e., cancer cell fraction) by combining Battenberg's estimation of tumor purity and the density and distribution of SNV VAF within clonal diploid regions of each sample genome. Mutations from within different gains were collapsed if molecular time (above) supported the gains as occurring in the same time window. For IID_H198325, IID_198331, and IID_H198328 the molecular time was not possible to be estimated because of either low mutational burden or excess of unassigned mutations. Given the patterns observed in the other tMN, the gains from these samples were treated as having occurred in the same molecular time for the purposes of this approach.

Tumor phylogeny from Pyclone, above, revealed clonal and subclonal mutations and all SNV were subsequently grouped into clonal duplicated and non-duplicated groups, and subclonal (non-duplicated) groups. Mutational signature analysis was run on each group of intra-gain mutations, as above, to determine pre- and post-gain mutational processes active before and after the gain. At least 40 total mutations per multi-gain were required in a group to confidently ascertain mutational signature contribution.

A similar approach was applied to amplifications that were caused by chromothripsis events[27]. After correction of VAF for sample purity, mutations across gains deemed part of chromothripsis events[9] were pooled and separated into duplicated and non-duplicated groups. Clustered mutations (i.e., kataegis) were filtered out so as not to skew results towards distinct mutational signatures (e.g., APOBEC)[1]. Mutational signature analysis was performed for events with 40 or more clonal mutations.

Clinical correlation with timing of chemotherapy and associated signatures was then compared to molecular time estimates and according to duplication status to reconstruct evolutionary timelines relative to chemotherapy exposure.

The code used for this chemotherapy barcoding approach is available at: https://github.com/UM-Myeloma-Genomics/Timing-of-Gains-with-Chemotherapy-Barcoding/tree/main

Statistical Analysis

Each specific statistical test is annotated in the text. Fisher's Exact Test and the Wilcoxon Ranked Sum test were used to compare differences between groups. A false discovery rate was used to correct for multiple hypothesis testing. P-values <0.05 and q values <0.1 were considered statistically significant. Survival data were analyzed and visualized with Kaplan-Meier methods.

SUPPLEMENTAL METHODS REFERENCES

1. Maura F, Degasperi A, Nadeu F, et al. A practical guide for mutational signature analysis in hematological malignancies. *Nature communications*. 2019;10(1):1-12.
2. Wong TN, Ramsingh G, Young AL, et al. Role of TP53 mutations in the origin and evolution of therapy-related acute myeloid leukaemia. *Nature*. 2015;518(7540):552-555.
3. Network CGAR. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *New England Journal of Medicine*. 2013;368(22):2059-2074.
4. Gouyette A, Hartmann O, Pico J-L. Pharmacokinetics of high-dose melphalan in children and adults. *Cancer chemotherapy and pharmacology*. 1986;16(2):184-189.
5. Landau HJ, Yellapantula V, Diamond BT, et al. Accelerated single cell seeding in relapsed multiple myeloma. *Nature communications*. 2020;11(1):1-10.
6. Rustad EH, Yellapantula V, Leongamornlert D, et al. Timing the initiation of multiple myeloma. *Nature communications*. 2020;11(1):1-14.
7. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nature biotechnology*. 2011;29(1):24-26.
8. Jones D, Raine KM, Davies H, et al. cgpCaVEManWrapper: simple execution of CaVEMan in order to detect somatic single nucleotide variants in NGS data. *Current protocols in bioinformatics*. 2016;56(1):15.10. 11-15.10. 18.
9. Rustad EH, Yellapantula VD, Glodzik D, et al. Revealing the impact of structural variants in multiple myeloma. *Blood cancer discovery*. 2020;1(3):258.
10. Martincorena I, Raine KM, Gerstung M, et al. Universal patterns of selection in cancer and somatic tissues. *Cell*. 2017;171(5):1029-1041. e1021.
11. Nik-Zainal S, Van Loo P, Wedge DC, et al. The life history of 21 breast cancers. *Cell*. 2012;149(5):994-1007.
12. Tyner JW, Tognon CE, Bottomly D, et al. Functional genomic landscape of acute myeloid leukaemia. *Nature*. 2018;562(7728):526-531.
13. Alexandrov LB, Kim J, Haradhvala NJ, et al. The repertoire of mutational signatures in human cancer. *Nature*. 2020;578(7793):94-101.
14. Pich O, Cortes-Bullich A, Muiños F, Pratcorona M, Gonzalez-Perez A, Lopez-Bigas N. The evolution of hematopoietic cells under cancer therapy. *Nature communications*. 2021;12(1):1-11.
15. Pich O, Muiños F, Lolkema MP, Steeghs N, Gonzalez-Perez A, Lopez-Bigas N. The mutational footprints of cancer therapies. *Nature genetics*. 2019;51(12):1732-1740.
16. Rustad EH, Nadeu F, Angelopoulos N, et al. mmsig: a fitting approach to accurately identify somatic mutational signatures in hematological malignancies. *Communications biology*. 2021;4(1):1-12.
17. Bolton KL, Ptashkin RN, Gao T, et al. Cancer therapy shapes the fitness landscape of clonal hematopoiesis. *Nature genetics*. 2020;52(11):1219-1226.
18. Chapuy B, Stewart C, Dunford AJ, et al. Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. *Nature medicine*. 2018;24(5):679-690.
19. Cheng DT, Mitchell TN, Zehir A, et al. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. *The Journal of molecular diagnostics*. 2015;17(3):251-264.

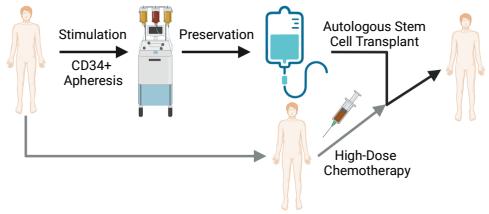
20. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*. 2013;31(3):213-219.
21. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics*. 2012;28(14):1811-1817.
22. Raine KM, Hinton J, Butler AP, et al. cgpPindel: identifying somatically acquired insertion and deletion events from paired end sequencing. *Current protocols in bioinformatics*. 2015;52(1):15.17. 11-15.17. 12.
23. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079.
24. Takahashi K. Germline polymorphisms and the risk of therapy-related myeloid neoplasms. *Best Practice & Research Clinical Haematology*. 2019;32(1):24-30.
25. Gerstung M, Jolly C, Leshchiner I, et al. The evolutionary history of 2,658 cancers. *Nature*. 2020;578(7793):122-128.
26. Oben B, Froyen G, Maclachlan KH, et al. Whole-genome sequencing reveals progressive versus stable myeloma precursor conditions as two distinct entities. *Nature communications*. 2021;12(1):1-11.
27. Shoshani O, Brunner SF, Yaeger R, et al. Chromothripsis drives the evolution of gene amplification in cancer. *Nature*. 2021;591(7848):137-141.

Tracking the Evolution of Therapy-Related Myeloid Neoplasm Using Chemotherapy Signatures

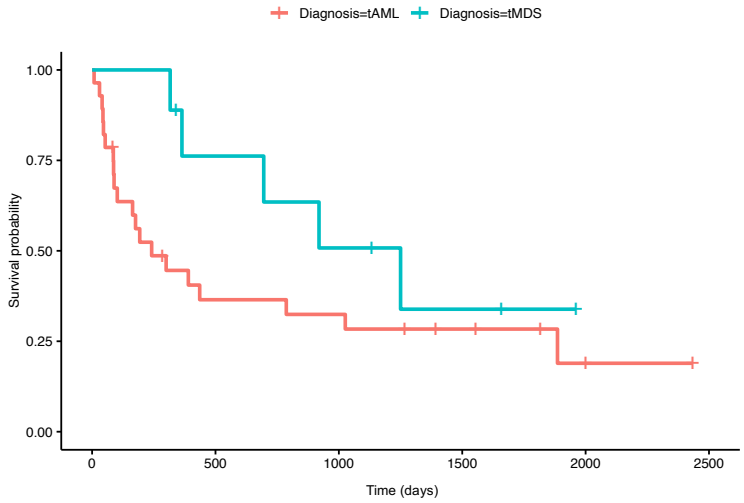
Supplemental Figures

Supplemental Figure 1. Clinical summary of therapy-related neoplasms included in this study. a) Depiction of autologous stem cell transplantation. **b)** Kaplan-Meier curves for therapy-related AML and MDS from the whole genome cohort.

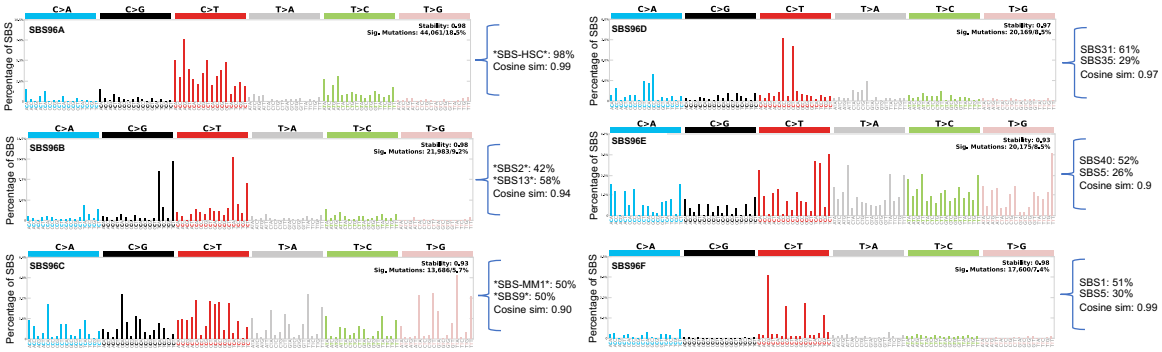
a



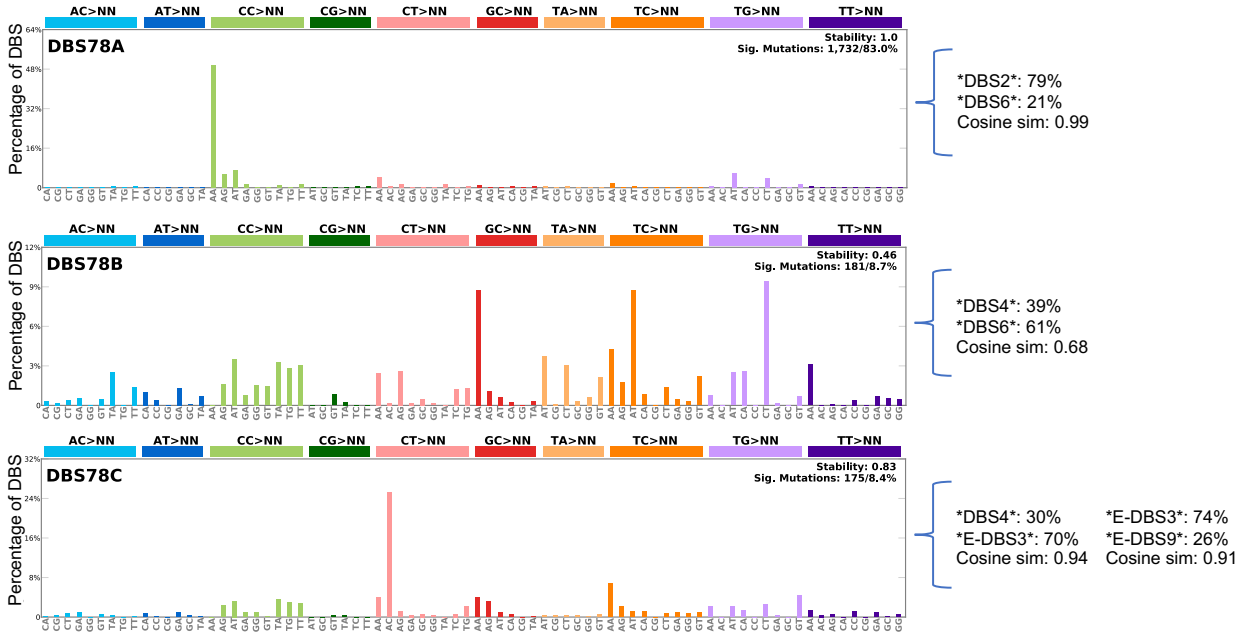
b



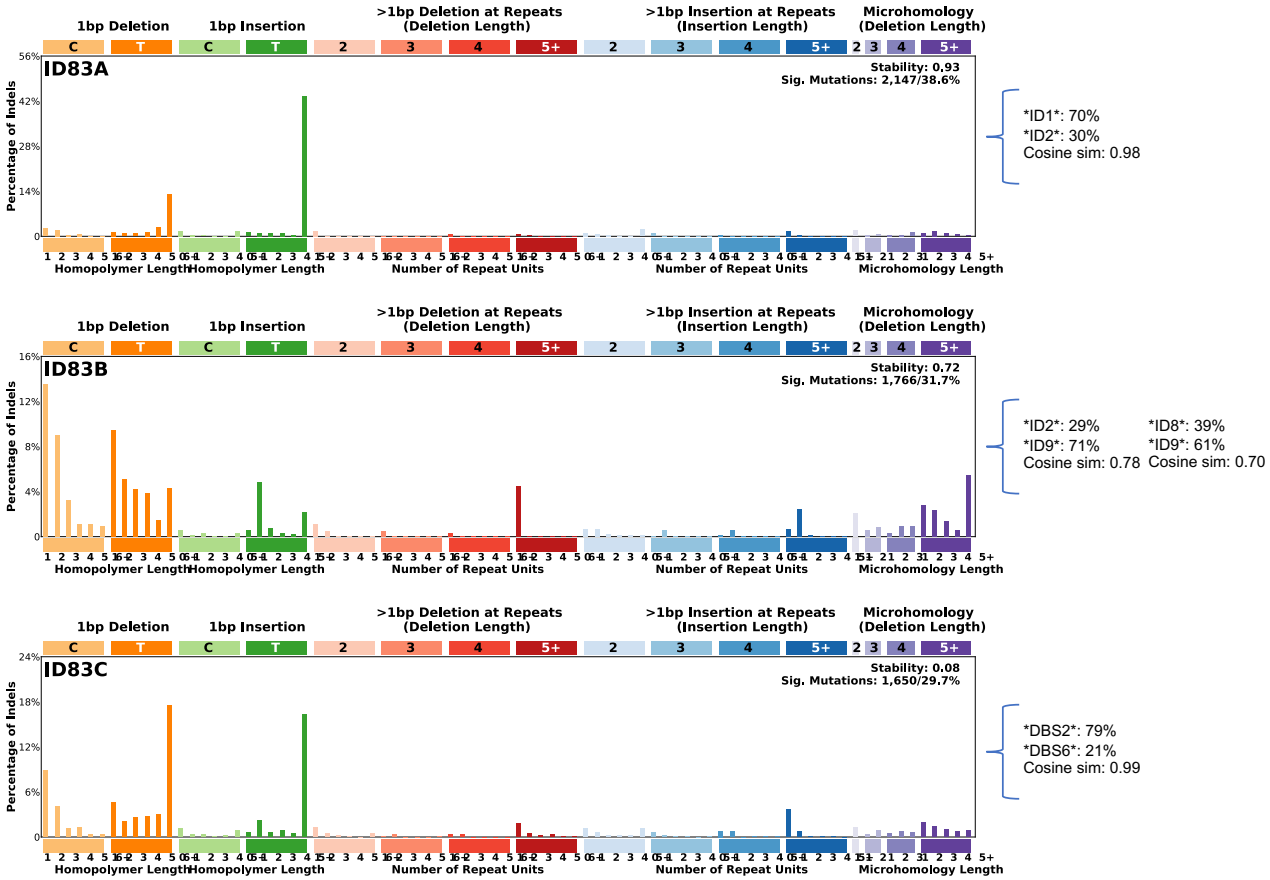
Supplemental Figure 2. Single Base Substitution (SBS) signatures. Output from *SigProfiler* SBS96 signatures de novo extraction. SBS signatures annotated with an asterisk have been revised using a pairwise deconvolution solution containing non-COSMIC mutational signatures (**Methods**). The remaining are original *SigProfiler* suggested deconvolution solutions.



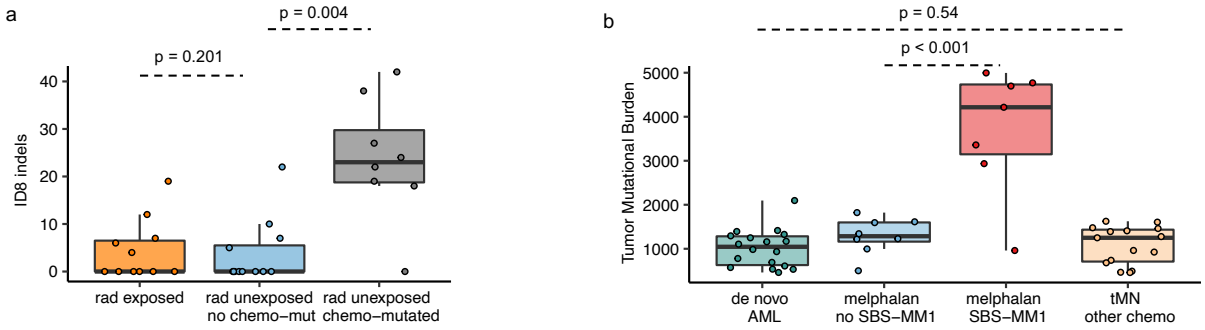
Supplemental Figure 3. Double Base Substitution (DBS) signatures. Output from *SigProfiler* DBS78 signatures interpreted using a pairwise deconvolution solution containing non-COSMIC mutational signatures (**Methods**).



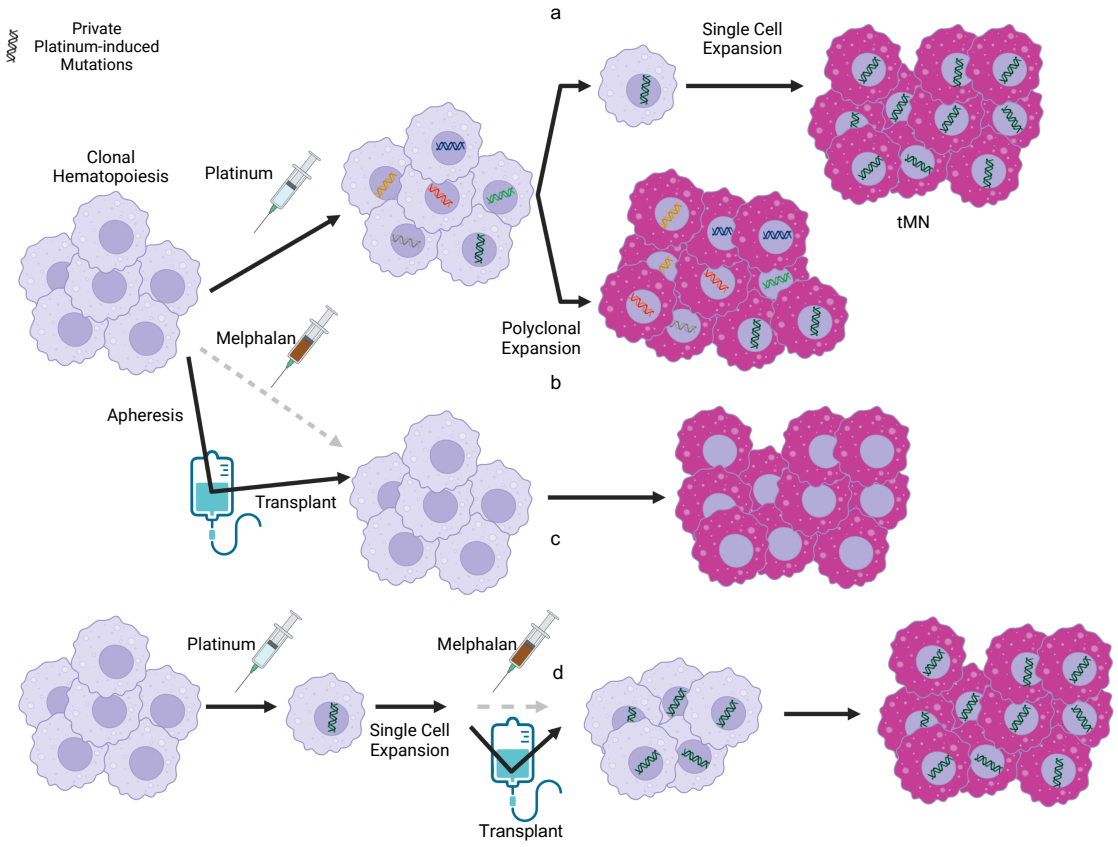
Supplemental Figure 4. Indel (ID) signatures. Output from *SigProfiler* ID83 signatures revised using a pairwise deconvolution solution (**Methods**).



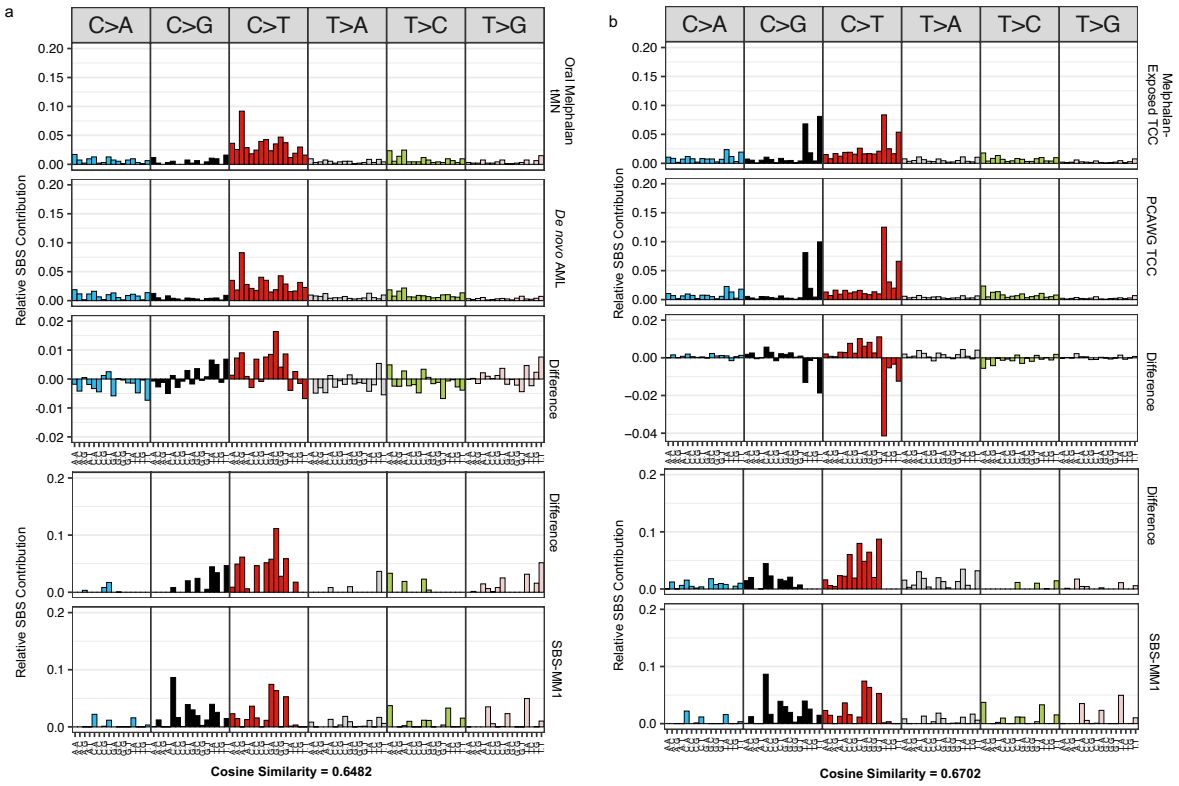
Supplemental Figure 5. Impact of exposure to chemotherapy on indels and SBS burden. a) Number of indels attributable to the ID8 indel signature between radiation (rad)-exposed patients with or without concurrent chemotherapy SBS mutational signature. **b)** Comparison of SBS mutational burden in post-melphalan therapy-related myeloid neoplasms with or without the SBS-MM1 signature. The p-values were estimated using Wilcoxon test.



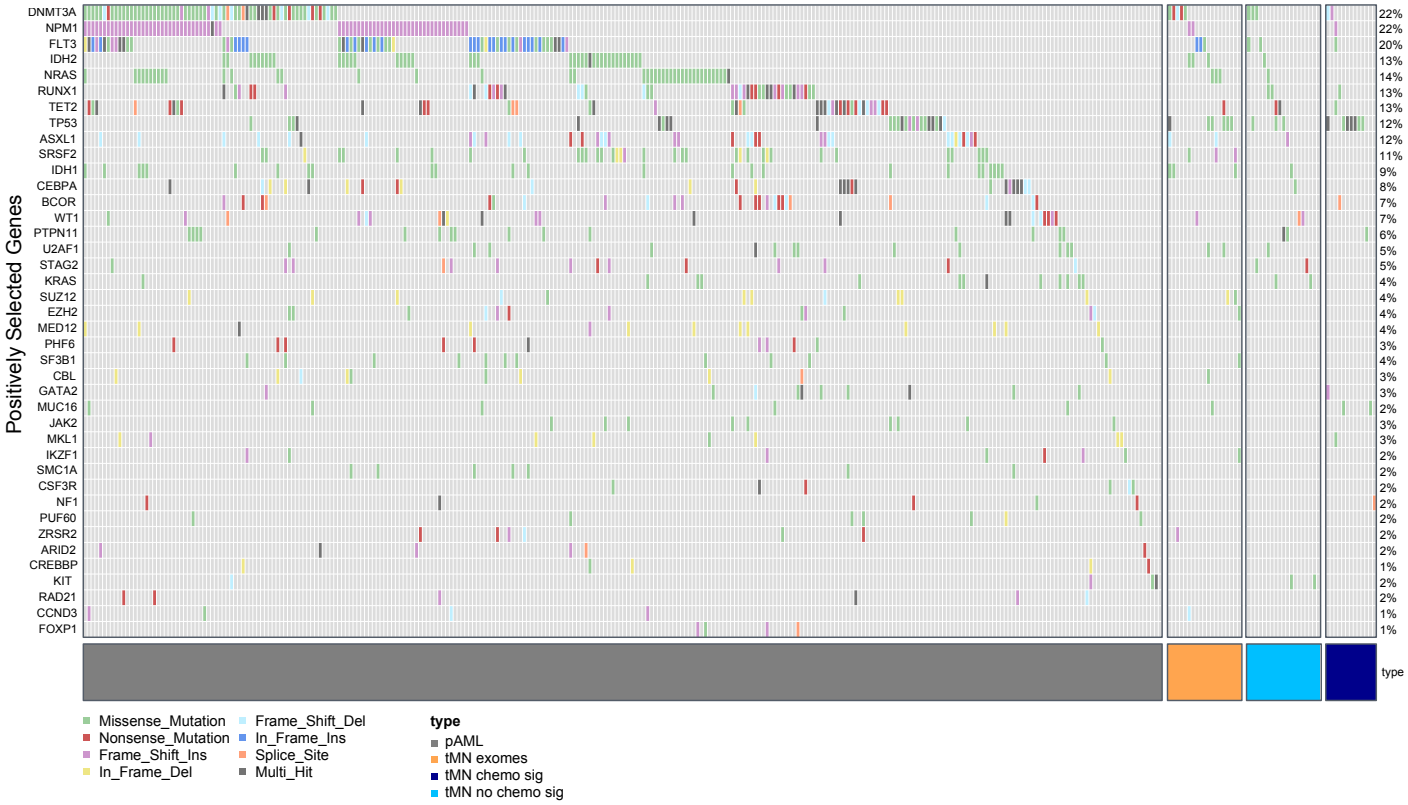
Supplemental Figure 6. Schema for the measurement of chemotherapy-related mutational signatures in bulk whole genome sequencing data. a) Measurement of chemotherapy-associated mutational signatures depends on a single cell, bearing a unique chemotherapy-associated mutational catalogue, to expand to clonal dominance (single cell expansion model). **b)** Polyclonal expansion following chemotherapy exposure does not yield a measurable chemotherapy-associated signature. **c)** Another explanation for lack of signature expression is escape from exposure entirely via leukapheresis and reinfusion. Rather than observing both platinum and melphalan signatures following sequential exposure in the same patient **(d)**, only platinum signatures are present suggesting escape to subsequent melphalan exposure via leukapheresis.



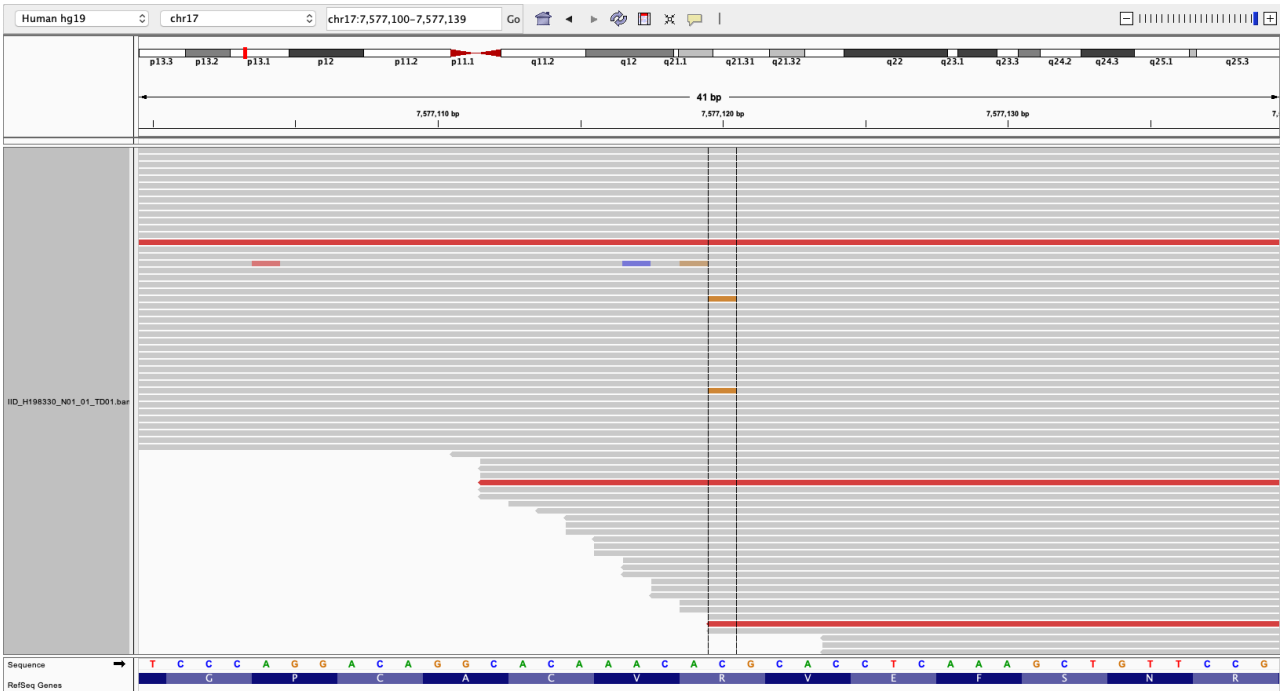
Supplemental Figure 7. Detecting the SBS-MM1 (melphalan) signature in two tumors without potential for transplant-mediated escape. **a)** SBS difference between mutational profile for a tMN developing after oral melphalan exposure (IID_H201267) and the cumulative profiles of n=18 *de novo* AML WGS (top). The difference (removing negative contributions) compared to the SBS-MM1 mutational signature profile (bottom). **b)** SBS difference between mutational profile for a melphalan-exposed transitional cell carcinoma (TCC) and the cumulative profiles of n=23 PCAWG TCC (top). The difference (removing negative contributions) compared to the SBS-MM1 mutational signature profile (bottom).



Supplemental Figure 8. Mutations in driver genes in therapy related myeloid neoplasms. Oncoplot of driver SNV defined using *dndscv* for 316 *de novo* AML and 61 tMN combining WGS and Beat AML exomes.

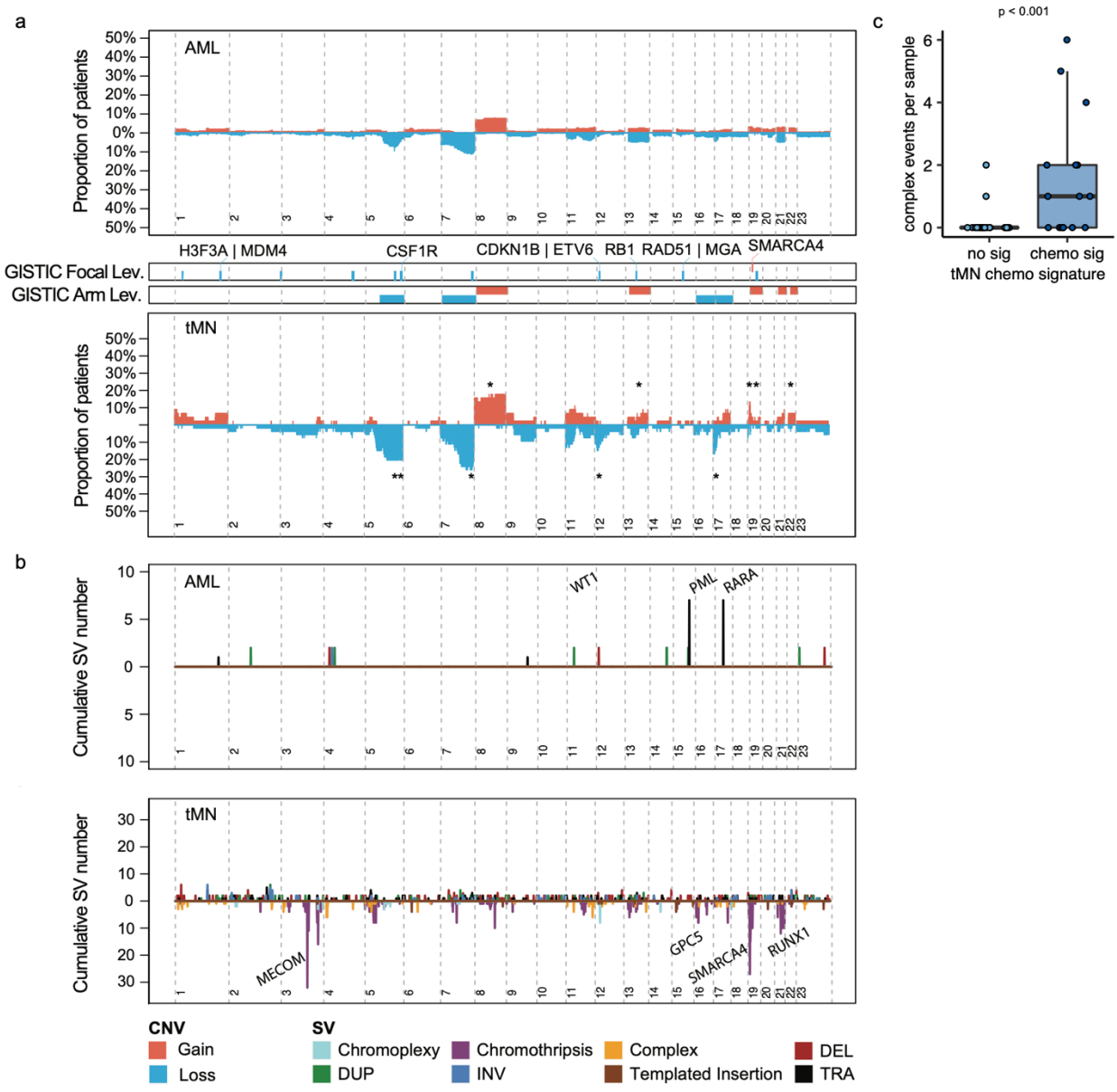


Supplemental Figure 9. Example of Clonal Hematopoiesis. Screenshot from Integrated Genome Viewer (IGV) showing a small *TP53* mutant clone in a sample taken from the leukapheresis product. This clone expanded into a tumor without a melphalan signature, having escaped direct exposure to chemotherapy and resultant mutagenesis.

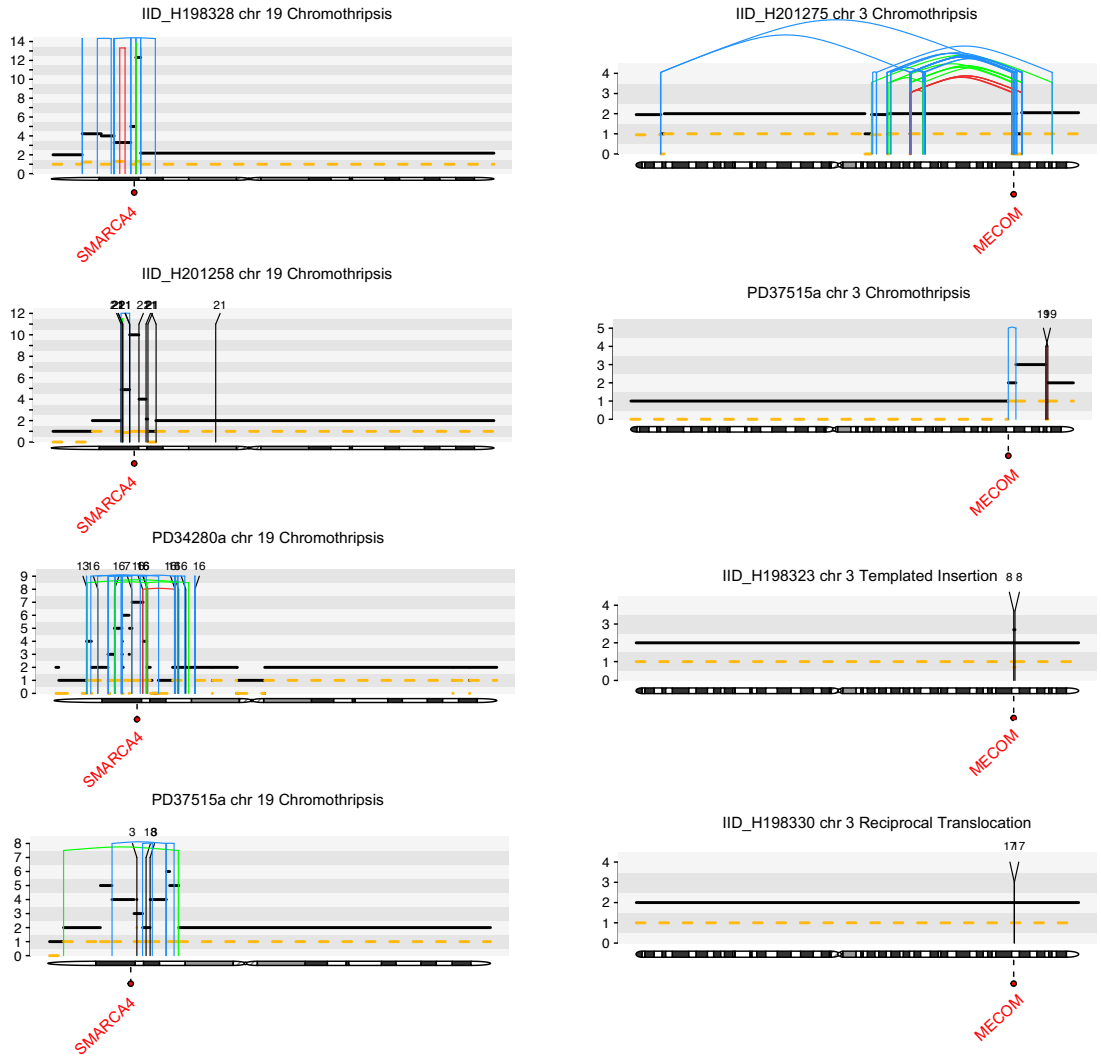


TP53

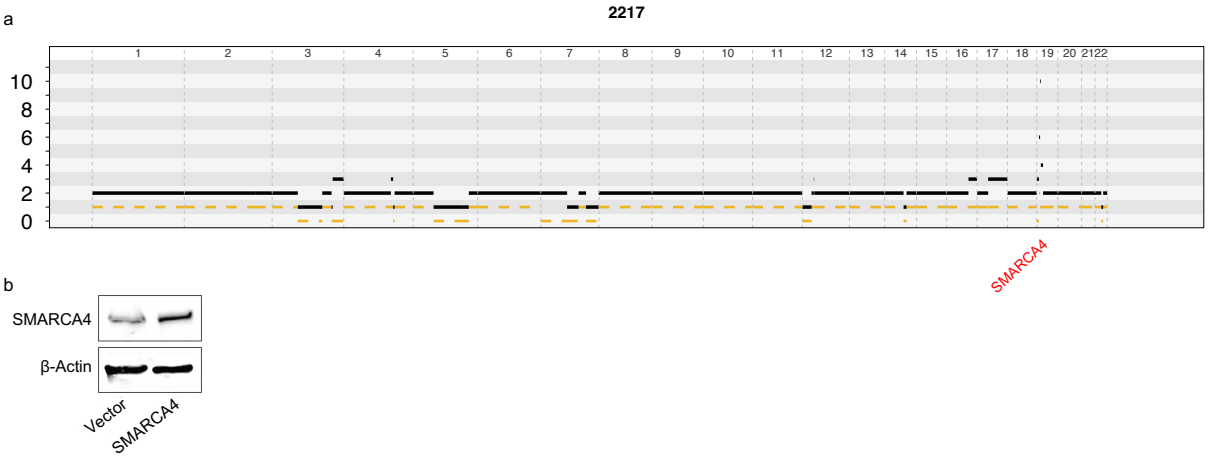
Supplemental Figure 10. Copy number and structural variant differences in all therapy-related myeloid neoplasms and *de novo* AML. a) Top. Cumulative copy number profile for all *de novo* AML samples (n = 316; 18 genomes from TCGA and 298 exomes from Beat AML). Bottom. Cumulative copy number for 39 tMN genomes and 22 tMN exomes. GISTIC peaks and CNV arm-level events enriched in tMN with chemotherapy signatures (p<0.05, FDR<0.1; Fisher test) are annotated with an asterisk for significance as compared to: tMN without chemotherapy signatures, grey; *de novo* AML, red; both, black. **b)** Structural variant landscape across all *de novo* AML (Top; n=18) and tMN genomes (Bottom; n=39). SV breakpoints are binned into 1 megabase segments. For visual purposes, simple events point upward from the x-axis and complex events (e.g. chromothripsis) point downward. **c)** Boxplot of number of complex structural variants by presence of chemotherapy signature among tMN cases. The p-value was estimated using Wilcoxon test.



Supplemental Figure 11. Simple and complex structural variants involving the *SMARCA4* and *MECOM* loci in tMN genomes. The horizontal black line indicates the total copy number; the dashed orange line indicates the minor copy number. The vertical lines represent SV breakpoints, color-coded based on SV class: blue = inversion, green = tandem-duplication; red = deletion; black = translocation.



Supplemental Figure 12. *SMARCA4* in *de novo* AML and Ba/F3 cell line. a) Copy number profile from the only *de novo* AML tumor (among n=298 cases in the BEAT-AML study) to contain a high copies amplification of the *SMARCA4* locus. **b)**: Western blot showing expression of *SMARCA4* in transfected cells compared to vector for cytokine independence assay.



Supplemental Figure 13. Molecular time and clock-like single base substitution mutation rate in multiple myeloma, *de novo* and therapy-related myeloid neoplasms. a) Linear regressions showing the (lack of) association between clock-like SBS mutations and age of tumor, regardless of presence or absence of chemotherapy-induced mutagenesis (i.e., chemotherapy-associated mutational signatures). P-values and R squares were estimated using *lm* R function. **b)** Molecular time estimates for large gains in eligible tumors. Events occurring closer to tumor sequencing (i.e., diagnosis) are later in molecular time (i.e., closer to 1). **c)** Individual SBS5 mutation rate estimate for multiple myeloma using linear mixed effect model. The SBS5 mutational burden was derived from the phylogenetic branches of each patient (dots). A total of 77 WGS (multiple myeloma and smoldering myeloma) from 47 patients cases from a prior study (Rustad et al. Nat Comm 2020) were included together with two newly sequenced post-platinum multiple myeloma tumors.