This is a pre print version of the following article:

Deciphering TP53 mutant Cancer Evolution with Single-Cell Multi-Omics / Rodriguez-Meira, A.; Norfo, R.; Wen, W. X.; Ch('e)deville, A. L.; Rahman, H.; O?sullivan, J.; Wang, G.; Louka, E.; Kretzschmar, W. W.; Paterson, A.; Brierley, C.; Martin, J. -E.; Demeule, C.; Bashton, M.; Sousos, N.; Hamblin, A.; Guermouche, H.; Pasquier, F.; Marzac, C.; Girodon, F.; Drummond, M.; Harrison, C.; Plo, I.; Jacobsen, S. E. W.; Psaila, B.; Thongjuea, S.; Antony-Debr('e), I.; Mead, A. J.. - (2022), pp. 1-3. [10.1101/2022.03.28.485984]

Terms of use:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

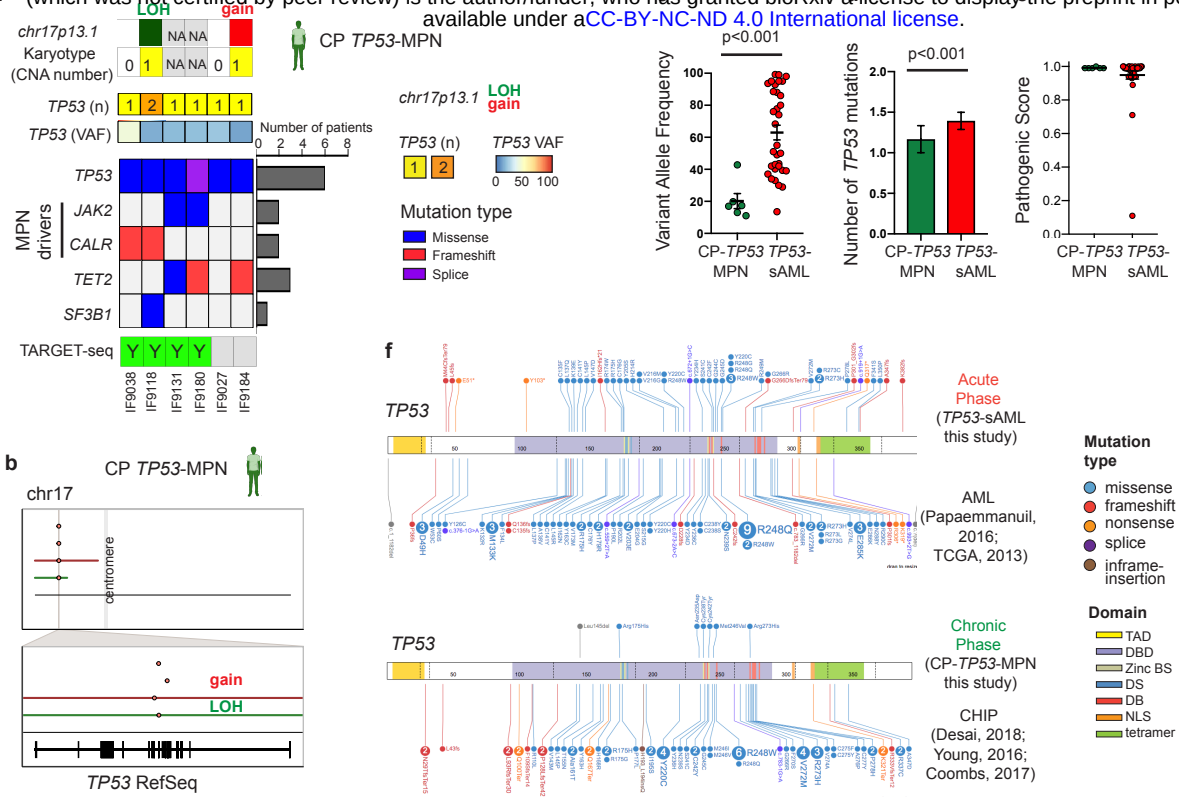
25/04/2024 15:13

(Article begins on next page)

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Extended Data Figure 8

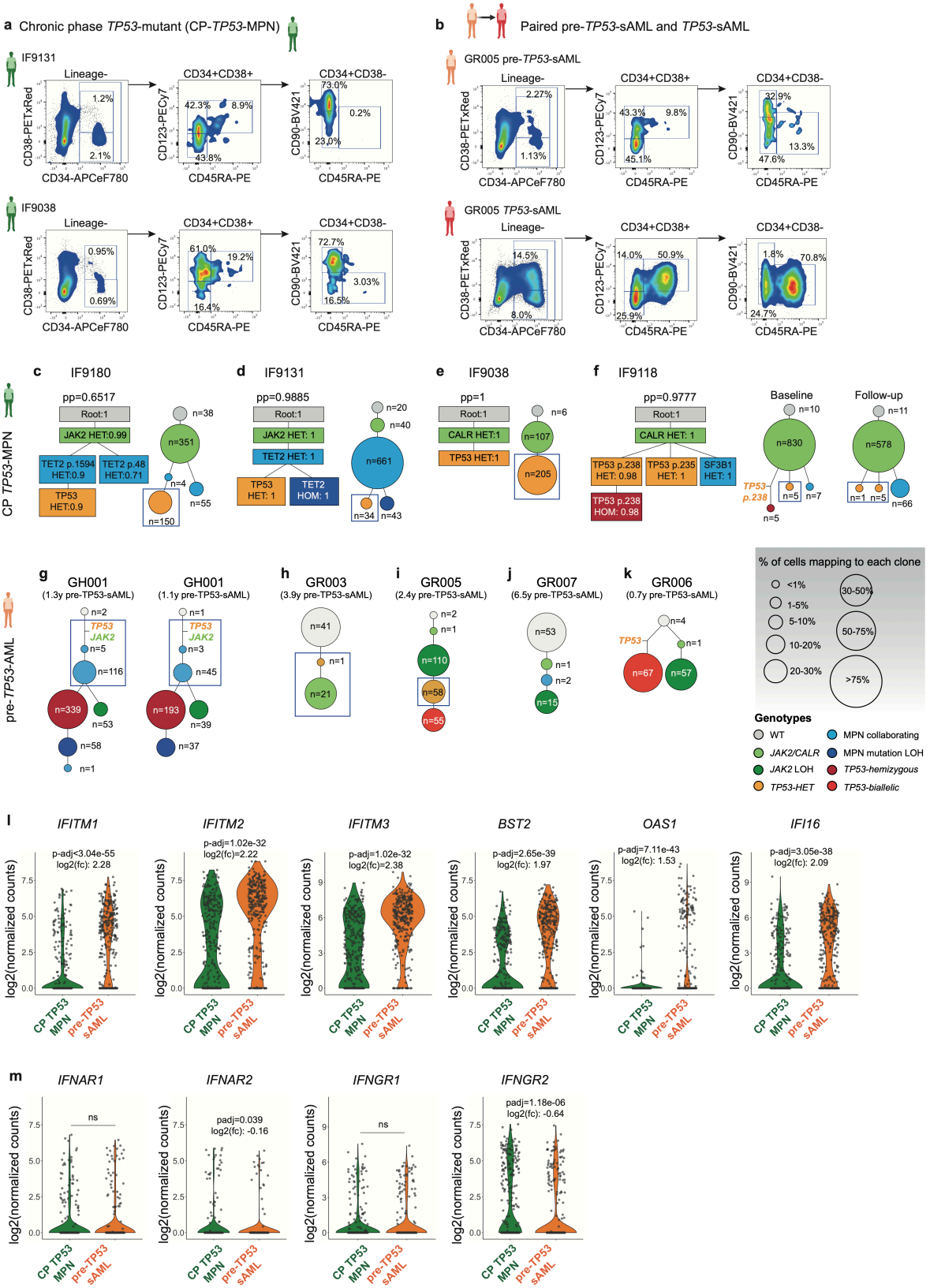
bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Extended Data Fig.8. Genetic landscape of chronic phase *TP53*-mutant MPN. **a,** Point mutations and cytogenetic abnormalities identified in a cohort of 6 CP *TP53*-MPN patients with no evidence of clinical transformation after 4.43 years [2.62-5.94] median follow-up. The number of patients in which each gene is mutated in shown on the barplot on the right and patients processed for TARGET-seq analysis are indicated below the heatmap. **b,** Summary of CNA events in chr17 and *TP53* gene in the 2 CP *TP53*-MPN patients with detectable CNAs. The top panel shows a whole chromosome view and the bottom one, the gene-level view and RefSeq track. Points indicate the location of each point mutation and solid lines indicate CNA status. **c-e,** Comparison of variant allele frequency (**c**), number of TP53 mutations (**d**) and pathogenic scores (**e**) of *TP53* variants identified in CP *TP53*-MPN (n=6) and *TP53*-sAML patients (n=33). Mean ± s.e.m. is shown; “p” indicates two-tailed Mann-Whitney test p-value. **f,** Location and mutation type stratified by patient group (chronic/acute phase) as compared to previously published CHIP and AML patient cohorts.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Extended Data Figure 9

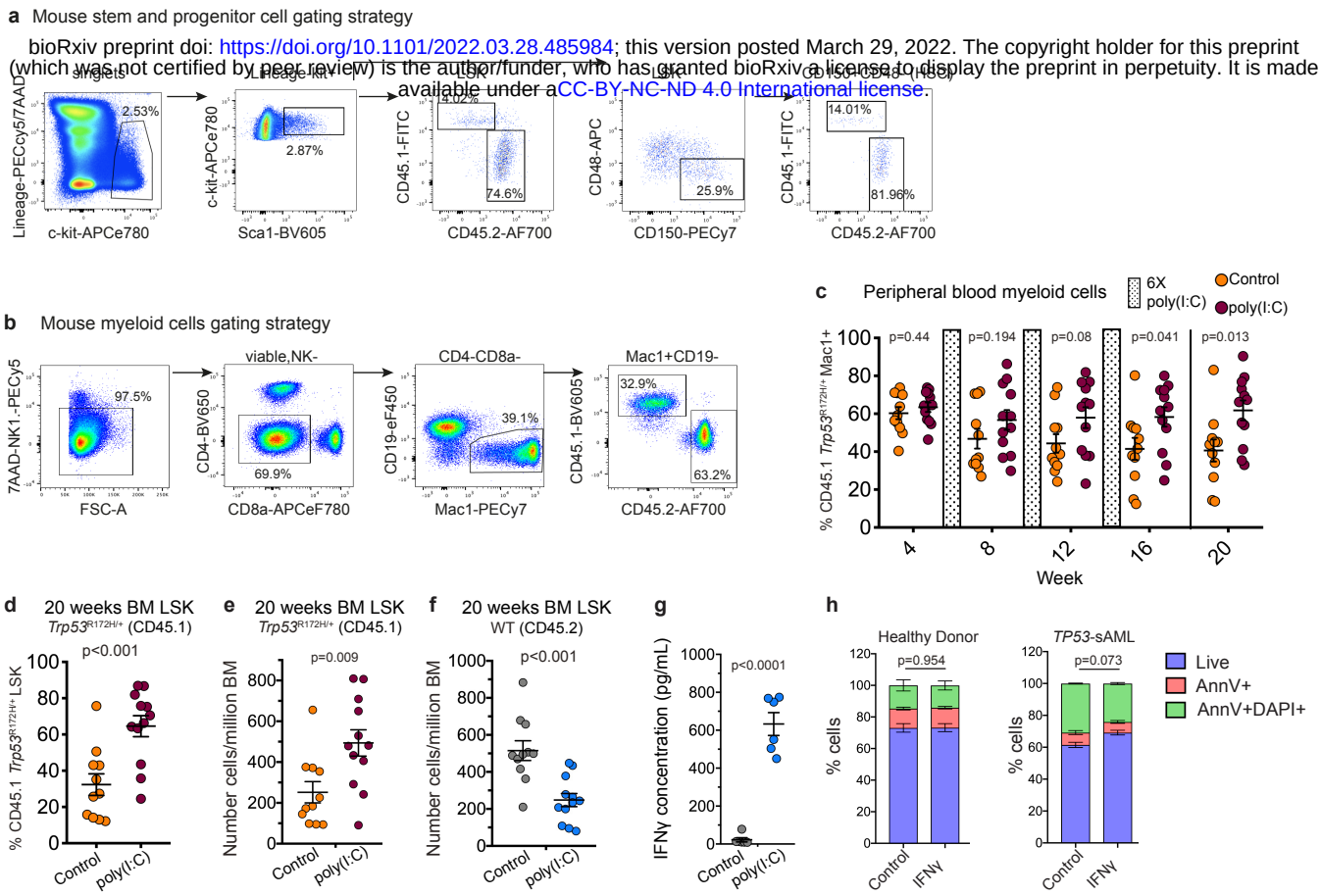


bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Extended Data Fig.9. Clonal evolution and molecular signatures of *TP53*-mutant patients at chronic phase. a-b, Flow cytometry profiles of the $Lin^{-}CD34^{+}$ HSPC compartment in two CP *TP53*-MPN patients without evidence of clinical transformation (**a**) and in a representative paired chronic phase (**b**, up; pre-*TP53*-sAML) and acute phase (**b**, bottom; *TP53*-sAML) sample (Related to Fig.4a). **c-f,** Phylogenetic reconstruction of clonal hierarchies in CP *TP53*-MPN patients from single-cell TARGET-seq genotyping data. In each panel, the phylogenetic tree computed using SCITE is shown on the left, and the number of cells mapping to each clone for each patient, on the right. “pp” indicates posterior probability or each consensus mutation tree, and the probability of each genotype transition is indicated in the square for each mutation. The size of the circles is proportional to the size of each clone and is coloured according to the genotype indicated in the genotype key. For patient IF9118 **(f**), baseline (left) and 4 years of follow-up (right) samples are shown separately. **g-k,** Phylogenetic reconstruction of clonal hierarchies in pre-*TP53*-AML patients from single-cell TARGET-seq genotyping data (related to Extended Data Fig.2). The size of the circles is proportional to each clone’s size, and is coloured according to the genotype indicated in the genotype key. In panels (c-k), blue boxes indicate *TP53*-heterozygous clones used for the analysis presented in Fig.4c. **l-m,** Expression of key interferon-response genes **(l)** and interferon receptors **(m)** in *TP53*-heterozygous cells from CP *TP53*-MPN (n=296 cells) and pre-*TP53*-sAML patients (n=314 cells). “p-adj” indicates adjusted p-value from combined Fisher’s exact test and Wilcoxon tests, calculated using Fisher’s method and adjusted using Benjamini & Hochberg procedure; “fc” indicates fold-change (related to Fig.4c). Violin plots indicate log2(counts) distributions and each point represents the expression value of a single-cell.

Extended Data Figure 10

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Extended Data Fig.10. *TP53*-mutant cells display an aberrant inflammatory response. **a-b,** Gating strategy for mouse chimera experiments (Related to Fig.4d-h) used to quantify CD45.1+ LSK and HSCs populations in the BM (**a**) and myeloid cells in the peripheral blood (PB) (**b**). **c-f,** Analysis of WT:*Trp53*$^{R172H/+}$ chimera mice treated with 3 regimes of 6 poly(I:C) injections with serial readouts of CD45.1 *Trp53*$^{R172H/+}$ Mac1+ PB cells (**c**), percentage of CD45.1 *Trp53*$^{R172H/+}$ BM LSK (Lin-Sca-1+c-Kit+) (**d**), number of CD45.1 *Trp53*$^{R172H/+}$ BM LSK (**e**) and CD45.2 WT BM LSK per million BM cells (**f**) 20 weeks post transplantation. n=11-12 mice per group from 3 biological replicates in 2 independent experiments. Bars indicate mean ± s.e.m. and "p", two-tailed unpaired t-test p-value. **g,** IFNγ level in spleen serum 4h after poly(I:C) injection. n=6 mice per group from 2 independent experiments. Lines indicate mean ± s.e.m. and "p", two-tailed unpaired t-test p-value. **h,** Percentage of apoptotic cells from healthy donor (n=3) or *TP53*-sAML patients (n=2) determined by Annexin-V/DAPI staining 24h after IFNγ treatment of HSPCs. "p" indicates two-tailed unpaired t-test p-value.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods

Banking and processing of human samples

Primary human samples (peripheral blood or bone marrow, described in Table S1) were analysed with approvals from the Inserm Institutional Review Board Ethical Committee (project C19-73, agreement 21-794, CODECOH n°DC-2020-4324); and from the INForMeD Study (REC: 199833, 26 July 2016, University of Oxford). Patients and normal donors provided written informed consent in accordance with the Declaration of Helsinki for sample collection and use in research. For secondary AML patients, we specifically selected samples from patients with known *TP53*-mutation.

Cells were subjected to Ficoll gradient centrifugation and for some samples, CD34 enrichment was performed using immunomagnetic beads (Miltenyi). Total mononuclear cells (MNCs) or $CD34^+$ cells were frozen in FBS supplemented with 10% DMSO for further analysis.

Targeted bulk sequencing

Bulk genomic DNA from patient samples' mononuclear or $CD34^+$ cells was isolated using DNeasy Blood & Tissue Kit (Qiagen) or QIAamp DNA Mini Kit (Qiagen) as per manufacturer's instructions. Targeted sequencing was performed using a TruSeq Custom Amplicon panel (Illumina) or a Haloplex Target Enrichment System (Agilent technologies) with amplicons designed around 32, 44 or 77 genes[46]. Targets were chosen based on the genes/exons most frequently mutated and/or likely to alter clinical practice (diagnostic, prognostic, predictive or monitoring capacity) across a range of myeloid malignancies (e.g. MDS/AML/MPN). Targets covered in all panels include *ASXL1, CALR, CBL, CEBPA, CSF3R DNMT3A, EZH2, FLT3, HRAS, IDH1, IDH2, JAK2, KIT, KRAS, MPL, NPM1, NRAS, PHF6, RUNX1, SETBP1, SF3B1, SRSF2, TET2, TP53, U2AF1, WT1, ZRSR2*. Sequencing was performed with a MiSeq sequencer (Illumina), according to the manufacturer's protocols. Results were analysed after alignment of the reads using two dedicated pipelines, SOPHiA DDM® (Sophia Genetics) and an in-house software GRIO-Dx®. For all samples, an average depth exceeding 200X for > 90% of the target regions was required, or as previously described[16]. All pathogenic variants were manually checked

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

using Integrative Genomics Viewer software. Analysis is presented in Extended Data Fig.1a and Extended Data Fig.8a.

Pathogenic scores for each *TP53* variant (Extended Data Fig.8e) were derived from COSMIC (Catalogue Of Somatic Mutations In Cancer) using the FATHMM-MKL algorithm. The FATHMM-MKL algorithm integrates functional annotations from ENCODE with nucleotide-based hidden Markov models to predict whether a somatic mutation is likely to have functional, molecular and phenotypic consequences. Scores greater than 0.7 indicate that a somatic mutation is likely pathogenic, whilst scores less than 0.5 indicate a neutral classification.

The type and location of *TP53* mutations from this study, *de novo* AML patients and CHIP individuals represented in Extended Data Fig.8f were generated using Pecan Portal[47]. *De novo* AML *TP53* mutations were downloaded from Papaemmanuil, *et al.*[48] and Ley, *et al.*[27]; CHIP associated *TP53* mutations were obtained from Coombs, *et al.,* Desai, *et al.*, Young, *et al.* [49-51]

Sanger sequencing of patient-associated mutations in PDX models

Genomic DNA from PDX sorted populations (LMPP: $hCD45^+Lin^-CD34^+CD38^-CD45RA^+CD90^-$ and GMP: $hCD45^+Lin^-CD34^+CD38^+CD45RA^+CD123^+$) was extracted using QIAamp DNA Mini Kit (Qiagen). Sanger sequencing was performed with forward or reverse primers (TableS6a) targeting mutations identified by targeted bulk sequencing in the corresponding primary samples using Mix2seq kit (Eurofins Genomics) and sequences were analysed with the ApE editor.

Single Nucleotide Polymorphism Array sample preparation, Copy Number Variant and Loss of Heterozygosity Analysis

Bulk genomic DNA from patients' mononuclear cells was isolated using DNeasy Blood & Tissue Kit (Qiagen) as per manufacturer's instructions. 250 ng of gDNA were used for hybridization on an Illumina Infinium OmniExpress v1.3 BeadChips platform.

To call mosaic copy number events in primary patient samples, genotyping intensity data generated was analysed using the Illumina Infinium OmniExpress v1.3 BeadChips platform. Haplotype phasing, calculation of log R ratio (LRR) and B-allele frequency (BAF)

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and calling of mosaic events was performed using Mocha (Mocha: A BCFtools extension to call mosaic chromosomal alterations starting from phased VCF files with either B Allele Frequency (BAF) and Log R Ratio (LRR) or allelic depth (AD)), as previously described[52,53]. In brief, Mocha comprises the following steps: (1) filtering of constitutional duplications; (2) use of a parameterized hidden Markov model to evaluate the phased BAF for variants on a per-chromosome basis; (3) deploying a likelihood ratio test to call events; (4) defining event boundaries; (5) calling copy number; (6) estimating the cell fraction of mosaic events. A series of stringent filtering steps was applied to reduce the rate of false positive calls. To eliminate possible constitutional and germline duplications, excluding calls with lod_baf_phase <10, those with length <500kbp and rel_cov>2.5, and any gains with estimated cell fraction >80%, logR>0.5 or length <24Mb. Given that interstitial LOH are rare and likely artefactual, all LOH events <8Mb were filtered[52]. Events on genomic regions reported to be prone to recurrent artefact[52] (chr6<58Mb, chr7>61Mb, and chr2 >50Mb) were also filtered, and those where manual inspection demonstrated noise or sparsity in the array.

To find common genomic lesions on a focal and arm level, Infinium OmniExpress arrays were initially processed with Illumina Genome Studio v2.0.4. Following this, Log R Ratio (LRR) data was extracted for all probes and array annotation obtained from Illumina (InfiniumOmniExpress-24v1-3_A1). LRR data was then smoothed and segmentation called using the CBS algorithm from the DNACopy[54,55] v1.60.0 package in R. A minimum number of 5 probes was required to call a segment, and segments where analysed using GenomicRanges[56] v1.38.0. Definitions of amplification, gain, loss and deletion events where as outlined in Bashton, *et al.*[57]. Segmentation data was then analysed in GISTIC[58] v2.023.

For PDX models, genomic DNA from sorted populations (LMPP: hCD45$^+$Lin$^-$CD34$^+$CD38$^-$ CD45RA$^+$CD90$^-$ and GMP: hCD45$^+$Lin$^-$CD34$^+$CD38$^+$CD45RA$^+$CD123$^+$) was extracted using QIAamp DNA Mini Kit (Qiagen). SNP-CGH array hybridization was performed using the Affymetrix Cytoscan® HD (Thermo Fisher Scientific) according to the manufacturer's recommendations. DNA amplification was checked using BioSpec-nano^{TM} spectrophotometer (Shimadzu) with expected concentrations between 2,500 and

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3,400ng/μL. DNA length distribution post-fragmentation was checked using D1000 ScreenTapes on Tapestation 4200 instrument (Agilent Technologies). Cytoscan HD array includes 2.6 million markers combining SNP and non-polymorphic probes for copy number evaluation. Raw data CEL files were analysed using the Chromosome Analysis Suite software package (v4.1, Affymetrix) with genome version GRCh37 (hg19) only if achieving the manufacturer's quality cut-offs. Only CNAs > 10kb were reported in the analysis presented in Extended Data Fig.3k,l.

Single-molecule cloning and sequencing of patient-derived cDNA

To experimentally verify the biallelic nature of *TP53* mutations in *TP53*-sAML patients, cDNA from a selected patient with putative *TP53* biallelic status (Patient ID GR004) was PCR-amplified using cDNA-specific primers spanning both *TP53* mutations (Fwd: 5'-GACCCTTTTTGGACTTCAGGTG-3', Rev: 5'-CCATGAGCGCTGCTCAGATAG-3'). PCR amplification was performed with KAPA 2X Ready Mix (Roche), a Taq-derived enzyme with A-tailing activity, for direct cloning into a TA vector (pCR2.1 TOPO vector, TOPO® TA Cloning® Kit, Invitrogen) as per manufacturer's instructions. Sanger sequencing for 10 different colonies was performed using M13 forward and reverse primers; a representative example is shown in Extended Data Fig.1h.

Fluorescent activated cell sorting (FACS) and single-cell isolation

Single cell FACS-sorting was performed as previously described[16], using BD Fusion I and BD Fusion II instruments (Becton Dickinson) for 96-well plate experiments or bulk sorting experiments, and SH800S or MA900 (SONY) for 384-well plate experiments. Experiments involving isolation of human haematopoietic stem and progenitor cells (HSPCs) included single colour stained controls (CompBeads, BD Biosciences) and Fluorescence Minus One controls (FMOs). Antibodies used for HSPC staining are detailed in TableS7a (Panel A or B).

Briefly, single cells directly sorted into 384-well plates containing 2.07 μL of TARGET-seq lysis buffer[59]. Lineage$^-$CD34$^+$ cells were indexed for CD38, CD90, CD45RA, CD123 and CD117 markers, which allowed us to record the fluorescence levels of each marker for each single cell. 7- aminoactinomycin D (7-AAD) was used for dead cell exclusion. Flow

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

cytometry profiles of the HSPC compartment (Extended Data Fig.2, Fig.9) were analysed using FlowJo software (version 10.1, BD Biosciences).

Single-cell TARGET-seq cDNA synthesis.

RT and PCR steps were performed as previously described[59], using 24 cycles of PCR amplification. Target-specific primers spanning patient-specific mutations were added to RT and PCR steps (TableS6a). After cDNA synthesis, cDNA from up to 384 single-cell libraries was pooled, purified using Ampure XP Beads (0.6:1 beads to cDNA ratio; Beckman Coulter) and resuspended in a final volume of 50 µL of EB buffer (Qiagen). The quality of cDNA traces was checked using a High Sensitivity DNA Kit in a Bioanalyzer instrument (Agilent Technologies).

Whole transcriptome library preparation and sequencing

Pooled and bead-purified cDNA libraries were diluted to 0.2 ng/µL and used for tagmentation-based library preparation using a custom P5 primer and 14 cycles of PCR amplification[59]. Each indexed library was purified twice with Ampure XP beads (0.7:1 beads to cDNA ratio), quantified using Qubit dsDNA HS Assay Kit (Invitrogen, Cat# Q32854) and diluted to 4 nM. Libraries were sequenced on a HiSeq4000, HiSeqX or NextSeq instrument using a custom sequencing primer for read1 (P5_seq: GCCTGTCCGCGGAAGCAGT GGTATCAACGCAGAGTTGC*T, PAGE purified) with the following sequencing configuration: 15 bp R1; 8 bp index read; 69 bp R2 (NextSeq) or 150 bp R1; 8 bp index read; 150 bp R2 (HiSeq).

TARGET-seq single-cell genotyping

After RT-PCR, cDNA+amplicon mix was diluted 1:2 by adding 6.25 µL of DNAse/RNAse free water to each well of each 384-well plate. Subsequently, a 1.5 µL aliquot from each single cell derived library was used as input to generate a targeted and Illumina-compatible library for single cell genotyping[59]. In the first PCR step, target-specific primers containing a plate-specific barcode (TableS6b) were used to amplify the target regions of interest. In a subsequent PCR step, Illumina compatible adaptors (PE1/PE2) containing single-direction indexes (Access Array™ Barcode Library for Illumina® Sequencers-384, Single Direction, Fluidigm) were attached to pre-amplified amplicons, generating single-

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

cell barcoded libraries. Amplicons from up to 3,072 libraries were pooled and purified with Ampure XP beads (0.8:1 ratio beads to product; Beckman Coulter). These steps were performed using Biomek FxP (Beckman Coulter), Mosquito (TTP Labtech) and VIAFLO 96/384 (INTEGRA Biosciences) liquid handling platforms. Purified pools were quantified using Qubit dsDNA HS Assay Kit (Invitrogen, Cat# Q32854) and diluted to a final concentration of 4 nM. Libraries were sequenced on a MiSeq or NextSeq instrument using custom sequencing primers as previously described[59] with the following sequencing configuration: 150 bp R1; 10 bp index read; 150 bp R2.

Targeted single-cell genotyping analysis

Data pre-processing

For each cell, the FASTQ file containing both targeted gDNA and cDNA-derived sequencing reads was aligned to the human reference genome (GRCh37/hg19) using Burrow-Wheeler Aligner (BWA v0.7.17)[31] and STAR (v2.6.1d)[60]. Custom perl scripts were used to demultiplex the gDNA and mRNA reads in the BAM file into separate SAM files based on targeted-sequencing primer coordinates (https://github.com/albarmeira/TARGET-seq). Next, Samtools (v1.9)[61] was used to concatenate the BAM header to the resulting SAM files before re-converting the SAM file to BAM format, which was subsequently sorted by genomic coordinates and indexed. Both gDNA and mRNA reads were tagged with the cell's unique identifier using Picard (v2.3.0) "*AddOrReplaceReadGroups"* and duplicate reads were subsequently marked using Picard "*MarkDuplicates"*. The sequencing reads overhanging into intronic regions in the mRNA reads were additionally hard-clipped using GATK (v4.1.2.0) *SplitNCigarReads*[62,63].

Variant calling

Variants were called from the processed BAM files using GATK *Mutect2* with the options [*--tumor-lod-to-emit 2.0 --disable-read-filter NotDuplicateReadFilter --max-reads-per-alignment-start*] to increase the sensitivity of detecting low-frequency variants. The frequency of each nucleotide (A, C, G, T) and indels at each pre-defined variant site were

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

also called using a Samtools *mpileup* as previously described[16]. Lastly, the coverage at each pre-defined variant site were computed using Bedtools (v2.27.1)[64].

To determine the coverage threshold of detection for each variant site, the coverage for "blank" controls (empty wells) were first tabulated. A cut-off coverage outlier value was computed as having a coverage exceeding 1.5 times the length of the interquartile range from the 75th percentile. Next, a value of 30 was added to this outlier value to yield the final coverage threshold to be used for genotype assignment.

Genotype assignment

For each pre-defined variant site, the number of reads representing the reference and alternative (variant) alleles for indels (insertion and deletions) and SNVs (single nucleotide variants) were tabulated from the outputs of GATK *Mutect2* and Samtools *mpileup*, respectively.

Here, a genotype scoring system was introduced to assign each variant site into one of three possible genotypes: wildtype, heterozygous, or homozygous mutant. Chi-square (χ^2) test was first used to compare the observed frequency of reference and alternative alleles against the expected fraction of reference and alternative alleles corresponding to the three genotypes. The expected fraction of the reference alleles was 0.999, 0.5, and 0.001, and the expected fraction of the alternative alleles was 0.001, 0.5, and 0.999 for wildtype, heterozygous, and homozygous mutant genotype, respectively. The χ^2 statistics were then tabulated for each fitted model and converted to genotype scores using the following formula:

$$Score_{genotype} = \frac{1}{log10(\chi^2 + 1)}$$

The genotype assigned to the variant site was based on the genotype model with the highest score.

Next, the variant (alternative) allele frequency (VAF) was computed and variant sites with $2 < VAF < 4$ and $96 < VAF < 98$ were reassigned as "ambiguous". For cells with no variants

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

detected at the specific variant sites by the mutation callers (either due to the absence of the variants, i.e. wild-type genotype, or that such variants were present below the detection limit), a "wild-type" genotype was assigned to those cells with a coverage above the specific threshold and "low coverage" to those cells with coverage below such threshold.

Taken together, each variant site was assigned one of the five following genotypes: wildtype, heterozygous, homozygous mutant, ambiguous, or low coverage. Variants with ambiguous or low coverage assignments for a particular cell were excluded from the analysis.

Computational reconstruction of clonal hierarchies

Genotypes for each single cell were recoded for input to SCITE in a manner inspired by Morita *et al* [65]: each mutation in each gene was coded as two loci, representing two different alleles. In the first recorded loci, all homozygous calls from each mutation where coded as heterozygous genotype calls. In the second recorded loci, all heterozygous and homozygous genotype calls in the original mutation matrix were coded as homozygous reference (i.e. WT) and heterozygous, respectively. For example, if for a certain mutation 0 represents WT status, 1 encodes heterozygous and 2 refers to homozygous status, these would be encoded as (0,0), (1,0) and (1,1) respectively, where the first term in the parenthesis corresponds to the first loci and the subsequent, to the second loci.

Then, SCITE was used (git revision `2016b31`, downloaded from https://github.com/cbg-ethz/SCITE.git[66]) to sample 1000 mutation trees from the posterior for every single-cell genotype matrix corresponding to a particular patient, where all possible mutation trees are equally likely *a priori*. For patients in which several disease timepoints were available, all timepoints were merged for SCITE analysis. As parameters for every SCITE run "`–fd` 0.01" (corresponding to the allelic dropout rate of reference allele in our adapted SCITE model), "`-ad` 0.01" (corresponding to the allelic dropout of the alternate allele), a chain length (`-l`) of 1e6, and a thinning interval of 1 while marginalizing out cell attachments (`-p 1 -s`) were used.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To summarize the posterior tree sample distribution, the number of times a particular sample matched each tree was computed. For each patient, the most common tree topology in the posterior tree samples is reported (Extended Data Fig.2b-o, Fig.9c-k), where "pp" is the proportion of samples that match this tree. For each clade in the most common posterior tree, clade probabilities were estimated as the proportion of trees in the posterior that contained the clade. These are indicated in each square for each mutation in (Extended Data Fig.2b-o, Fig.9c-k).

Clone assignment

For every patient's most common posterior tree, we assigned every cell to the tree node that matches the genotype of that particular cell. If an exact match was not found, then for every tree node the loss of assigning a cell to that node was calculated using the following loss function:

$$\begin{aligned} l(m) = &\log(\mathrm{ADO})(m[1,2] + m[3,2]) \\ &+ \log(\mathrm{FD})(m[2,1] + m[2,3]) \\ &+ \log(\mathrm{ADO}^2\mathrm{FD})(m[1,3] + m[3,1]) \end{aligned}$$

where m is a confusion matrix generated across all loci of a cell in which the first index represents the genotype that was measured for that particular cell (1 = homozygous reference, 2 = heterozygous, 3 = homozygous alternate), and the second index represents the genotype implied by the tree node. ADO $= 0.01$ and FD $= 0.001$ were used. Every cell was assigned to the node with the lowest loss l. For the trees presented in Extended Data Fig.2b-o and Extended Data Fig.9c-k only the numbers of cells with exact genotype matches were reported.

Testing for evidence of homozygous genotypes

Due to the nature of our loci-specific mutation encoding (each gene is encoded as two loci), homozygous mutations are placed in the clonal hierarchy independently of their accuracy. Therefore, for every patient and at every locus with observed homozygous alternate genotype calls, the tested null hypothesis was that all homozygous alternate genotype calls are due to allelic dropout at a level not exceeding 0.05 using a one-tailed binomial test. The total number of draws for the test is the number of heterozygous and

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

homozygous alternate genotype calls at the locus, the number of successful draws is the number of homozygous alternate calls, and the success rate is 0.05. Only homozygous alternate genotype calls below this 0.05 cut-off were reported in Extended Data Fig.2b-o and Extended Data Fig.9c-k; the results of the binomial test are reported for each patient and mutation in TableS8.

Computational validation of *TP53* biallelic status from single-cell targeted genotyping datasets

To further validate the biallelic status of *TP53* mutations in our dataset, the patterns of allelic dropout in TARGET-seq single-cell genotyping data from patient carrying at least 2 different *TP53* mutations were investigated (n=6; Extended Data Fig.1j).

To test the hypothesis that the observed *TP53*-WT/*TP53*-homozygous (*TP53*-WT/HOM; or (0,2)) cells are the result of a chromosomal loss (and therefore, in different alleles), the following null hypothesis (H_0) was formulated: observed *TP53*-WT/HOM cells are double allelic dropout events. Under H_0, every *TP53*-WT/HOM cell (0,2), *TP53*-HOM/WT cell (2,0), *TP53*-HOM/HOM (2,2) as well as an unknown number of *TP53*-WT/WT (0,0) are the result of a *TP53*-HET/HET (1,1) cell undergoing allelic dropout (ADO) at both sites. The following assumptions were made: (a) ADO is unbiased towards HOM or WT and (b) ADO events at each *TP53* site are independent. The null hypothesis was then tested with a binomial test, where the number of (2,2) events should be half the sum of (0,2) + (2,0) events (Extended Data Fig.1j). (0,0) events were disregarded.

If *TP53* mutations are biallelic, the expected number of WT/HOM and HOM/WT would be higher than HOM/HOM cells taking into account TARGET-seq expected allelic dropout rates (1-5%).

Single cell 3'-biased RNA-sequencing data pre-processing

FASTQ files for each single cell were generated using bcl2fastq (version 2.20) with default parameters and the following read configuration: Y8N*, I8, Y63N*. Read 1 corresponds to a cell-specific barcode, index read correspond to an i7 index sequence from each cDNA pool, and read 2 corresponds to the cDNA molecule. PolyA tails were trimmed from

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

demultiplexed FASTQ files with TrimGalore (version 0.4.1). Reads were then aligned to the human genome (hg19) using STAR (version 2.4.2a) and counts for each gene were obtained with FeatureCounts (version 1.4.5-p1; options --primary). Counts were then normalized by dividing each gene count by the total library size of each cell and multiplying this value by the median library size of all cells processed, as implemented in the "*normalize_UMIs*" function from the SingCellaR package[67] (https://github.com/supatt-lab/SingCellaR). A summary of the pre-processing pipeline can be found in https://github.com/albarmeira/TARGET-seq-WTA.

Quality control was performed using the following parameters: number of genes detected>500, percentage of ERCC derived reads<35%, percentage of mitochondrial reads<0.25%, percentage of unmapped reads<75%. Cells with less than 2000 reads in batch1, 5000 reads in batch2 and 10000 reads in batch3 were further excluded. This QC step was performed independently for each sequencing batch owing to differences in sequencing depth (mean library size: 42949 batch 1, 93580 batch 2 and 173145 batch3). After these QC steps, 7200 cells passed QC for batch1, 5838 for batch2 and 6490 for batch 3 (78.5%, 75.0% and 82.4% of cells processed, respectively). Then, 2733 cells from a previously published study[16] corresponding to 8 myelofibrosis patients and 2 normal donor controls were further integrated, encompassing a final dataset of 22261 cells in total.

Identification of highly variable genes

Highly variable genes above technical noise were identified by fitting a gamma generalized linear model (GLM) model of the log2(mean expression level) and coefficient of variation for each gene, using the "*get_variable_genes_by_fitting_GLM_model*" from SingCellaR package and the following options: *mean_expr_cutoff* = 1, *disp_zscore_cutoff* = 0.1, *quantile_genes_expr_for_fitting* = 0.6, *quantile_genes_cv2_for_fitting* = 0.2. Those genes with a coefficient of variation above the fitted model and expression cut-off were selected for further analysis, excluding those annotated as ribosomal or mitochondrial genes.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

CNA inference from single cell transcriptomes

InferCNV was used to identify CNAs in single-cell transcriptomes[68] (https://github.com/broadinstitute/inferCNV/wiki). Briefly, inferCNV creates genomic bins from gene expression matrices and computes the average level of expression for each of these bins. The expression across each bin is then compared to a set of normal control cells, and CNAs are predicted using a hidden markov model. For each patient, inferCNV was performed with the following parameters: "*cutoff=0.1, denoise=T, HMM=T*", compared to the same set of normal donor control cells (n=992). To identify CNA subclones, inferCNV in *analysis_mode='subclusters'* was used. CNAs identified by inferCNV were manually curated by removing those with size<10kb, merging adjacent CNA calls with identical CNA status into larger CNA intervals and comparing them to SNP-Array bulk CNA calls. Finally, to generate combined TARGET-seq single-cell genotyping and CNA-based clonal hierarchies, the CNA status from each inferCNV cluster was assigned to its predominant genotype.

Dimensionality reduction, data integration and clustering

PCA was performed using "*runPCA*" function from the *SingCellaR* R package, and Force-directed graph analysis was subsequently performed using the "*runFA2_ForceDirectedGraph*" with the top 30 PCA dimensions and the following options: *n.neighbors*=5, *fa2_n_iter*=1000 to generate the plots in Extended Data Fig.4a.

For the analysis of patient IF0131 presented in Extended Data Fig.3m, PCA was performed using "*runPCA*" function from the *SingCellaR* R package and then UMAP was performed using the "*runUMAP*" function with the top 10 PCA dimensions and the following options: *n.neighbors*=20, *uwot.metric* = "correlation", *uwot.min.dist*=0.30, *n.seed* = 1.

Integration of TARGET-seq single-cell transcriptomes from 10538 cells corresponding to 14 *TP53*-sAML samples was performed using "*runHarmony*" function implemented in the SingCellaR package, using the patient ID as covariate and the following options: *n.dims.use*=20, *harmony.theta* = 1, *n.seed* = 1. Diffusion map analysis was performed using "*runDiffusionMap*" with the integrative Harmony embeddings and the following

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

parameters: *n.dims.use*=20, *n.neighbors*=5, *distance*="euclidean". Signature scores were calculated using "*plot_diffusionmap_label_by_gene_set*" to generate the plots in Fig.2a and Fig.3a.

Pseudotime trajectory analysis

Monocle3[69] (https://cole-trapnell-lab.github.io/monocle3/) was used to infer differentiation trajectories from single cell transcriptomes. Raw UMI count matrix and clustering annotations were extracted from the SingCellaR object to build a Monocle3 'cds' object. '*learn_graph'* function was then used calculate the trajectory, using *TP53*-WT preleukemic cell cluster as the root node. Pseudotime was calculated using '*order_cells'* function and overlayed on the diffusion map embeddings to generate the plot in Fig.2b.

Differential expression analysis

Differentially expressed genes from TARGET-seq datasets were identified using a combination of non-parametric Wilcoxon test, to compare the expression values for each group, and Fisher's exact test, to compare the frequency of expression for each group, as previously described[17]. Logged normalized counts were used as input for this comparison, including genes expressed in at least 2 cells. Combined p-values were calculated using Fisher's method and adjusted p-values were derived using Benjamini & Hochberg procedure. Significance level was set at p-adjusted<0.05. For the analysis presented in Extended Data Fig.4b and TableS2, the top 100 differentially expressed genes with log2(fold-change)>0.3 and at least 20% expressing cells are shown. For the analysis presented in Fig.2k,l, only genes overexpressed in *TP53* multi-hit cells and log2(fold-change)>0.75 were included; for Fig.4c, only those with log2(fold-change)>1 were considered. Violin plots (Extended Data Fig.9l,m) from selected differentially expressed genes were generated using "ggplot2" package in R.

Gene-Set Enrichment analysis

For analysis involving <500 cells per group (Fig.4c, TableS5) GSEA was performed using GSEA software (https:/www.gsea-msigdb.org/gsea/index.jsp) with default parameters and 1000 permutations on the phenotype, using log2(normalized counts).

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

For analysis involving >500 cells per group (Fig.3k and Extended Data Fig.4c), GSEA was performed with "*identifyGSEAPrerankedGene"* function from *SingCellaR* R package with default options. Briefly, differential expression analysis was performed between two cell populations using Wilcoxon rank sum test and the resulting p-values were adjusted for multiple testing using the Benjamini-Hochberg approach. Prior to the differential expression analysis, down-sampling was performed so that both cell populations had the same number of cells. Next, -log10(p-value) transformation was performed and the resulting p-values were multiplied by +1 or -1 if the corresponding log2FC was>0.1 or <-0.1, respectively. The genelist was ranked using this statistic in ascending order and used as input for GSEA analysis using "*fgsea"* function from the *fgsea* R package with default options.

MSigDB HALLMARK v7.4 50-gene sets or previously published signatures (https://www.gsea-msigdb.org/gsea/msigdb/cards/GENTLES_LEUKEMIC_STEM_CELL_UP) were used for all analysis. Normalised enrichment scores (NES) were displayed in a heatmap using *pheatmap* R package. Gene sets with False Discovery Rate (FDR) q-value lower than 0.25 were considered significant.

Projection of single cell transcriptomes

A previously published human haematopoietic atlas was downloaded from https://github.com/GreenleafLab/MPAL-Single-Cell-2019 and used as a normal haematopoietic reference to project *TP53*-sAML and *de novo* AML transcriptions using Latent Semantic Index Projection (LSI)[70]. Common genes to all datasets were selected and then, *TP53*-sAML or previously published *de novo* AML cells[25] were projected using "*projectLSI'* function for the analysis presented in Fig.2c,d. A previously published human myelofibrosis atlas[71] was used as a reference to project *TP53*-sAML multi-hit cells in the analysis presented in Extended Data Fig.5a,b, using previously defined force-directed graph embeddings.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Velocyto analysis

Loom files were generated for each single cell using velocyto (v0.17.13) with options *-c* and *-U,* to indicate that each BAM represents an independent cell and reads are counted instead of molecules (UMIs), respectively[72]. The individual loom files were subsequently merged using the *combine* function from the *loompy* python module.

Healthy donors with at least 300 cells with RNA-sequencing data and patients with at least 300 cells consisting of >50 preleukemic (*TP53* wildtype) cells and > 50 *TP53* multi-hit cells were included for analysis. For each individual, Seurat object was created from the merged loom file and processed for downstream RNA-velocity analysis[73]. Specifically, for each patient, the spliced RNA counts were normalised using regularised negative binomial regression with the *SCTransform* function[74]. Next, linear dimension reduction was performed using *RunPCA* function and the first 30 principal components were further used to perform non-linear dimension reduction using the *RunUMAP* function. Ninety-six multiple rate kinetics (MURK) genes previously shown to possess coordinated step-change in transcription and hence violate the assumptions behind scVelo were removed [75]. The processed and MURK gene-filtered Seurat object was then saved as h5Seurat format using the *SaveH5Seurat* function and finally converted to h5ad format using the *Convert* function.

AnnData object was created from the h5ad file using the *scvelo* python module for RNA velocity analysis[76]. Highly variable genes were identified and the corresponding spliced and unspliced RNA counts were normalized and log2-transformed using the *scvelo.pp.filter_and_normalize* function. Next, the 1st and 2nd order moments were computed for velocity estimation using the *scvelo.pp.moments* function. The velocities (directionalities) were computed based on the stochastic model as defined in the *scvelo.t1.velocity* function, and the velocities was subsequently projected on the UMAP embeddings generated from Seurat above. Finally, the UMAP embeddings were annotated using the HSPC and erythroid lineage signature scores [67], and *TP53* mutation status. For each cell, the cell lineage signature score was computed using the average *SCTransform* expression values of the individual cell lineage genes.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Analysis of bulk BeatAML and TCGA gene expression datasets

Data retrieval and pre-processing

Two publicly available AML cohorts with genetic mutation and RNA-sequencing data available were used to validate findings from our single-cell analysis, namely BeatAML[26] and The Cancer Genome Atlas (TCGA)[27]. Gene expression values in FPKM (fragments per kilobase of transcript per million mapped reads) were retrieved from the National Cancer Institute (NIH) Genomic Data Commons (GDC)[77]. Gene expression values were then offset by 1 and log2-transformed. *TP53* point mutation status was retrieved from the cBio Cancer Genomics Portal (cBioPortal)[78]. Clinical data including survival data for BeatAML and TCGA was retrieved from the BeatAML data viewer (Vizome) and NIH GDC, respectively.

We selected samples from the BeatAML cohort with an AML diagnosis (540 *de novo* AML and 96 secondary AML) collected within 1 month of the patient's enrolment in the study, and with both *TP53* mutation status and RNA-sequencing data available. For patients in which multiple samples were available, samples were collapsed to obtain patient-level data. Specifically, the mean gene expression value for each gene from multiple samples was used to represent patient-level gene expression value. Furthermore, patients with at least one sample with a *TP53* mutation were considered *TP53*-mutant. Analysis of *TP53* variant allele frequency and reported karyotypic abnormalities indicated that the vast majority of patients could be classified as "multi-hit", and therefore patients were classified as *TP53*-mutant or WT without taking into account *TP53* allelic status. In total, 360 patients with *TP53* mutation status (329 *TP53*-WT and 31 *TP53*-mutant) and RNA-sequencing data available were included for analysis. Of these, 322 patients had concomitant survival data available (294 *TP53*-WT and 28 *TP53*-mutant).

The TCGA cohort consisted for 200 *de novo* AML patients represented by one sample each, out of which 151 patients had *TP53* mutation status (140 *TP53*-WT and 11 *TP53*-mutant) and RNA-sequencing data available, and were included for analysis. Of these, 132 patients had concomitant survival data available (124 *TP53*-WT and 8 *TP53*-mutant).

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Cell lineage gene signature scores

For each sample, a given cell lineage gene signature score was computed as the mean expression values of the individual genes belonging to the cell lineage gene signature. Here, the gene signature scores for two cell lineages were computed, namely myeloid and erythroid populations. Two gene sets for each cell lineage were compiled. The first gene set was based on cell lineage markers previously reported in the literature whereas the second gene set was based on cell lineage markers derived from analysing a published single-cell dataset[70]. Genes from each score are described in TableS3.

For the former approach, six erythroid genes (*KLF1, GATA1, ZFPM1, GATA2, GYPA, TFRC;* Fig.2e, Extended Data Fig.5h) and seven myeloid genes (*FLI1, SFPI1, CEBPA, CEBPB, CD33, MPO, IRF8;* Fig.2f) were identified. For the latter approach, the expression values of erythroid and myeloid cell clusters were first compared separately against all other cell clusters using Wilcoxon ranked sum test. The erythroid cluster consisted of the early and late erythroid populations while the myeloid cluster consisted of granulocyte, monocyte, and dendritic cell populations. Erythroid and myeloid-specific gene signatures were defined as genes having FDR values < 0.05 and log2 fold change > 0.5 in >=20 and 17 comparisons, respectively. In total, 100 erythroid genes and 135 myeloid genes were identified from this single-cell dataset (TableS3), and were used to compute the scores presented in Extended Data Fig.5c-f.

Prognostic signatures and Cox-regression survival models

Leukaemic stem cell (LSC) signature score

The 17-gene leukaemic stem cell (LSC17) gene set was retrieved from Ng *et al* [31]. For each sample, the LSC17 score was defined as the linear combination of gene expression values weighted by their respective regression coefficients.

To identify *TP53*-sAML leukaemic stem cell signatures from our TARGET single-cell dataset, two different approaches were used. First, differentially expressed genes were identified as overexpressed in all $Lin^{-}CD34^{+}$ *TP53* multi-hit cells regardless of their transcriptional classification (“p53-all-cells”) versus myelofibrosis, healthy donor and *TP53*-WT preleukaemic cells; this gene-set consists of 30 genes (TableS4a). For the

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

second approach, the same analysis was performed, but *TP53* multi-hit cells transcriptionally defined as leukaemic stem cells (falling in the leukaemic stem cell-like cluster, Fig.2a, middle) were specifically selected; this gene-set is comprised of 102 genes ("p53LSC"; TableS4a).

Next, lasso cox regression with 10-fold cross-validation implemented in the *glmnet* R package was used to identify p53-all-cells and p53-LSC genes that were associated with survival and to estimate their respective regression coefficients[79]. Specifically, Harrel's concordance measure (C-index) was used to assess the performance of each fitted model during cross-validation. The best model was defined as the fitted model with the highest C-index. Subsequently, the coefficient for each gene estimated using the best model was used to compute the gene signature scores. Only genes with non-zero coefficient values were included in the final gene set. In total, 27 and 51 genes were retained from the p53-all-cells and p53-LSC gene sets, respectively. For each sample, the gene signature score for each gene set was defined as the linear combination of gene expression values weighted by their respective regression coefficient[31,79]. The list of p53-LSC and p53-all-cells gene signatures is provided in TableS4b.

Survival analysis

For each gene expression signature, patients were first split using the median gene expression signature score. This resulted in two groups of patients, namely patients with high expression scores (greater than or equal to the median) and patients with low expression scores (lower than the median).

The Cox proportional hazards regression model implemented by the *survival* R package was fitted to estimate the hazard ratio associated with each feature. Log-rank test was used to test the differences between survival curves. The features analysed here were LSC17, p53-all-cells and p53-LSC signatures. Patients with low gene expression signature scores (below median) and patients with *TP53* wildtype status were specified as the reference groups in the model. Kaplan-Meier curves were plotted using the *survminer* R package to visualize the probability of survival and sample size at a respective time interval.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In vitro assays

Short-term liquid culture experiments and interferon treatment

For short-term liquid culture differentiation experiments (Fig.3j, Extended Data Fig.7g,h), 1, 5 or 10 cells from different Lineage$^-$CD34$^+$ HSPC populations (HSC CD34$^+$CD38$^-$CD45RA$^-$CD90$^+$, MPP CD34$^+$CD38$^-$CD45RA$^-$CD90$^-$, LMPP CD34$^+$CD38$^-$CD45RA$^+$, more committed progenitors CD34$^+$CD38$^+$) were directly sorted into a 96-well tissue culture plate containing 100 µL of differentiation media: StemSpan (Catalog #09650, StemCell Technologies), 1% Penicillin+Streptomycin, 20 % BIT9500 (Cat# 9500, StemCell Technologies), 10 ng/mL SCF (Cat #300-07, Peprotech), 10 ng/mL FLT3L (Cat# 300-19, Peprotech), 10 ng/mL TPO (Cat# 300-18-10, Peprotech), 5 ng/mL IL3 (Cat # 200-03, Peprotech), 10 ng/mL G-CSF (Cat# 300-23, Peprotech), 10 ng/mL GM-CSF (Cat# 300-03, Peprotech), 1 IU/mL EPO (Janssen, erythropoietin alpha, clinical grade) and 10 ng/mL IL6 (Cat# 200-06, Peprotech).

For differentiation experiments involving recombinant IFNγ (R&D Systems, 285-IF-100) and IFNα (rhIFN-alpha-2a, PBL Assay Science; 11100-1) treatment (Fig.4i), 100-500 Lin$^-$CD34$^+$ cells were directly sorted into a 96-well tissue culture plate containing 50 µL of 2X differentiation media as described above, and incubated for 1 hour at 37°C 5% CO2. Then, an additional 50 µL of media containing 2X recombinant interferon was added to each well and mixed carefully, to generate a 1X IFNα dilution (final concentration 50 IU/µL) and 1X IFNγ dilution (final concentration 2 ng/µL).

For all liquid culture experiments, 50 µL of fresh 1X differentiation media was added at day 4. Readout was performed by flow cytometry after 12 days of culture using the antibodies detailed in TableS7.c (Panel D).

Long-term culture initiating-cell (LTC-IC) assay

50 cells from each Lin$^-$CD34$^+$ population (HSC; MPP; LMPP; CD38+) and donor type (HD, MF, *TP53*-sAML) were sorted in triplicate. Cells were resuspended in 100 µL of myelocult (Stem Cell Technologies, #H5150) + Hydrocortisone (10^{-6}M; Stem Cell Technologies, Cat#74142) and plated into an irradiated supportive stromal cell layer (5000 SI/SI cells

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and 5000 M2-10B4 cells per well) in a 96-well tissue-culture plate coated with Collagen type I (CORNING; Cat#354236).

Medium was changed weekly and after 6 weeks of culture, cells were washed in IMDM+20%FCS and plated into 1.4 mL of cytokine-rich methylcellulose (Methocult H4435, Stem Cell Technologies). Colonies were scored 14 days later under an inverted microscope, and each colony was classified according to its morphology as BFU-E (Burst-forming unit erythroid), CFU-G (granulocyte), CFU-GM (granulocyte-macrophage), CFU-M (macrophage) or CFU-GEMM (granulocyte, erythrocyte, macrophage, megakaryocyte). Selected colonies were used for cytospin and genotyping as outlined below.

LTC-IC colony genotyping

LTC-IC colonies were picked from methylcellulose media, washed, resuspended in 10 µL of PBS and transferred to individual wells in a 96-well PCR plate. 15 µL of lysis buffer (Triton X-100 0.4%, Qiagen Protease 0.1 AU/mL) were added to each well and samples were incubated at 56 °C for 10 minutes and 72 °C for 20 minutes. A 3 µL aliquot from each lysate was used as input to generate a targeted and Illumina-compatible library for colony genotyping. The preparation of single cell genotyping libraries involves 3 PCR steps. In the first PCR step, target-specific primers spanning each mutation of interest are used for amplification (TableS6a); in the second PCR step, nested target-specific primers (TableS6b) attached to universal CS1 / CS2 adaptors (Forward adaptor, CS1: ACACTGACGACATGGTTCTACA; Reverse adaptor, CS2: TACGGTAGCAGAGACTTGGTCT) further enrich for target regions and in the third PCR step, Illumina-compatible adaptors containing sample-specific barcodes are used to generate sequencing libraries.

Apoptosis experiments under IFNγ treatment

500 $Lin^{-}CD34^{+}$ cells were sorted into StemSpan (Catalog # 09650, StemCell Technologies) supplemented with 1% Penicillin+Streptomycin, 20 % BIT9500 (Cat# 9500, StemCell Technologies), 10 ng/mL SCF (Cat #300-07, Peprotech), 10 ng/mL FLT3L (Cat# 300-19, Peprotech), 10 ng/mL TPO (Cat# 300-18-10, Peprotech), 5 ng/mL IL3 (Cat # 200-03, Peprotech) and 2 ng/µL rhIFNγ (R&D Systems, 285-IF-100). Cell were incubated at

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

37 C 5% CO_2 and 24 hours later, washed with AnnexinV Binding buffer 1X, stained with 1:100 AnnexinV-PE (Biolegend, Cat# 640907), DAPI and analysed immediately by flow cytometry.

TP53 knockdown and differentiation of human CD34+ cells

shRNA sequence for p53 knockdown has been previously cloned into the lentiviral vector pRRLsin-PGK-eGFP-WPRE and validated[80]. Primary human $CD34^+$ cells from patients with MPN (Table S1) were infected twice with scramble (shCTL) or shTP53 with a MOI (Multiplicity of Infection) of 15 and sorted 48h later on CD34 and GFP expression. Cells were cultured in serum-free medium with a cocktail of human recombinant cytokines containing EPO (1 U/mL, Amgen), FLT3-L (10 ng/mL, Celldex Therapeutics, Inc.), G-CSF (20 ng/mL, Pfizer), IL-6 (10 ng/mL, Miltenyi), GM-CSF (5 ng/mL, Peprotech), IL-3 (10 ng/mL, Miltenyi), TPO (10 ng/mL, Kirin Brewery) and SCF (25 ng/mL, Biovitrum AB).

At day 12 of culture, cells were stained with the antibodies detailed in TableS7.c, Panel C. DAPI was used for dead cell exclusion before acquisition on a FACSCanto II (BD Biosciences) instrument. Analysis of FACS data was performed using Kaluza (Beckman Coulter) software.

Quantitative real time PCR in shRNA experiments

In p53 knockdown experiments, RNA from either $CD34^+$ cells sorted after transduction or bulk cells at day 12 of culture was extracted using Direct-Zol RNA MicroPrep Kit (Zymo Research) and reverse transcription was performed with SuperScript Vilo cDNA Synthesis Kit (Invitrogen). Quantitative RT-PCR was performed on a 7500 Real-Time PCR Machine using SYBR-Green PCR Master Mix (Applied Biosystems). Expression levels were normalized to *PPIA* (housekeeping gene). Primers used are listed in TableS6c.

Xenotransplantation

Purified $CD34^+$ cells from AML patients were transplanted via retroorbital vein injection in sublethally irradiated (1.5Gy) NOD.CB17-*Prkdcscid IL2rgtm1*/Bcgen mice (B-NDG, Envigo). All experiments were approved by the French National Ethical Committee on Animal Care (n° 2020-007-23589). Blood cell counts were performed monthly by submandibular sampling of mice with blood chimerism assessed by flow cytometry using

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

hCD34, hCD45 and mCD45 antibodies (TableS7.b). At sacrifice (27 weeks or 31 weeks post-transplant), human bone marrow HSPC fractions were sorted on an Influx Cell sorter (BD Biosciences) after staining with the antibodies detailed in TableS7.b.

Evaluation of cell morphology

Cell morphology from PDX models (Extended Data Fig.3d) and *in vitro* LTC-IC cultures (Extended Data Fig.7e) was assessed after cytospin of 50-100,000 cells onto a glass slide (5 min at 500 rpm) and May-Grünwald Giemsa staining, according to standard protocols. Images were obtained using an AxioPhot microscope (Zeiss).

Mouse Bone Marrow Chimaeras

Trp53^{tm2Tyj} *Commd10*$^{Tg(Vav1-icre)A2Kio}$ (hereafter referred to as Trp53$^{R172H/+}$) CD45.1 mice and CD45.2 wild-type mice used for BM chimera experiments and IFNγ ELISA assays were bred and maintained in accordance to UK Home Office regulations. All experiments carried out in the UK were performed under Project License P2FF90EE8 approved by the University of Oxford Animal Welfare and Ethical Review Body.*Trp53*^{tm2Tyj} [81] and *Commd10*$^{Tg(Vav1-icre)A2Kio}$ [82] (Jackson laboratory stock number #008610) have been previously described.

1 million bone marrow (BM) cells from Trp53$^{R172H/+}$ CD45.1 mice and 1 million BM CD45.2 wild-type competitor mice were transplanted intra-venously into lethally irradiated (10 Gy total body irradiation, split dose) congenic CD45.2 mice. In each cohort, a selection of mice were injected intra-peritoneally with 3 rounds of 6 injections each of 200µg poly(I:C) (GE Healthcare, #27-4732-01). Poly(I:C) was administered during weeks 6-7, 10-11, 14-15. Within each round, injections were spaced one or two days apart. Analysis of peripheral blood chimerism was performed every 4 weeks, while BM chimerism was analysed 20 weeks after transplantation. Chimerism was assessed by flow cytometry (using the antibodies detailed in TableS7.d. 7AAD (Sigma) was used for dead cell exclusion. FACS analyses were carried out on BD Fortessa or BD Fortessa X20 (BD Biosciences) and profiles were later analysed using FlowJo software (version 10.1, BD Biosciences).

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

IFNγ ELISA assay

Wild-type mice were injected intra-peritoneally with a single dose of 200 µg poly(I:C) and spleens were collected from injected mice and non-treated controls 4 hours after injection. Spleens were processed into a single-cell suspension in 200 µl PBS, spun down at 500g for 5 minutes and supernatant was collected and used as spleen serum. IFNγ levels were assessed using mouse IFNγ Quantikine ELISA assay (R&D Systems, cat MIF00) following the manufacturer's instructions. 450nm and 540nm optical densities were determined using Clariostar microplate reader (BMG Labtech).

Statistical analysis

Statistical analyses are detailed in Figure Legends and performed using GraphPad Prism software (7 or later version) or R (version 3.6.1) software. Number of independent experiments, donors and replicates for each experiment are detailed in Figure Legends.

Data and code availability

Scripts to reproduce all figures will be uploaded in GitHub (https://github.com/albarmeira/) upon publication. Raw sequencing data will be made available through GEO (GSEXXXXXX) and targeted single-cell genotyping data will be made publicly available through SRA (SRAXXXXXX).

Methods References

46. Hamblin, A. *et al.* Development and Evaluation of the Clinical Utility of a Next Generation Sequencing (NGS) Tool for Myeloid Disorders. *Blood* **124**, 2373 (2014).
47. Zhou, X. *et al.* Exploring genomic alteration in pediatric cancer using ProteinPaint. *Nat Genet* **48**, 4-6 (2016).
48. Papaemmanuil, E. *et al.* Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N Engl J Med* **374**, 2209-2221 (2016).
49. Coombs, C.C. *et al.* Therapy-Related Clonal Hematopoiesis in Patients with Non-hematologic Cancers Is Common and Associated with Adverse Clinical Outcomes. *Cell Stem Cell* **21**, 374-382.e4 (2017).
50. Desai, P. *et al.* Somatic mutations precede acute myeloid leukemia years before diagnosis. *Nat Med* **24**, 1015-1023 (2018).
51. Young, A.L., Challen, G.A., Birmann, B.M. & Druley, T.E. Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. *Nat Commun* **7**, 12484 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

52. Loh, P.R. *et al.* Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* **559**, 350-355 (2018).
53. Loh, P.R., Genovese, G. & McCarroll, S.A. Monogenic and polygenic inheritance become instruments for clonal selection. *Nature* **584**, 136-141 (2020).
54. Olshen, A.B., Venkatraman, E.S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557-72 (2004).
55. Venkatraman, E.S. & Olshen, A.B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657-63 (2007).
56. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9**, e1003118 (2013).
57. Bashton, M. *et al.* Concordance of copy number abnormality detection using SNP arrays and Multiplex Ligation-dependent Probe Amplification (MLPA) in acute lymphoblastic leukaemia. *Sci Rep* **10**, 45 (2020).
58. Mermel, C.H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**, R41 (2011).
59. Rodriguez-Meira, A., O'Sullivan, J., Rahman, H. & Mead, A.J. TARGET-Seq: A Protocol for High-Sensitivity Single-Cell Mutational Analysis and Parallel RNA Sequencing. *STAR Protoc* **1**, 100125 (2020).
60. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
61. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
62. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-303 (2010).
63. Schischlik, F. *et al.* Mutational landscape of the transcriptome offers putative targets for immunotherapy of myeloproliferative neoplasms. *Blood* **134**, 199-210 (2019).
64. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-2 (2010).
65. Morita, K. *et al.* Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics. *Nature Communications* **11**, 5327 (2020).
66. Jahn, K., Kuipers, J. & Beerenwinkel, N. Tree inference for single-cell data. *Genome Biol* **17**, 86 (2016).
67. Roy, A. *et al.* Transitions in lineage specification and gene regulatory networks in hematopoietic stem/progenitor cells over human development. *Cell Reports* **36**, 109698 (2021).
68. Patel, A.P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396-401 (2014).
69. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods* **14**, 979-982 (2017).
70. Granja, J.M. *et al.* Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat Biotechnol* **37**, 1458-1465 (2019).
71. Psaila, B. *et al.* Single-Cell Analyses Reveal Megakaryocyte-Biased Hematopoiesis in Myelofibrosis and Identify Mutant Clone-Specific Targets. *Mol Cell* **78**, 477-492 e8 (2020).
72. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494-498 (2018).
73. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* **33**, 495-502 (2015).
74. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol* **20**, 296 (2019).
75. Barile, M. *et al.* Coordinated changes in gene expression kinetics underlie both mouse and human erythroid maturation. *Genome Biology* **22**, 197 (2021).
76. Bergen, V., Lange, M., Peidli, S., Wolf, F.A. & Theis, F.J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol* **38**, 1408-1414 (2020).

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

77. Heath, A.P. *et al.* The NCI Genomic Data Commons. *Nat Genet* **53**, 257-262 (2021).
78. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* **2**, 401-4 (2012).
79. Anande, G. *et al.* RNA Splicing Alterations Induce a Cellular Stress Response Associated with Poor Prognosis in Acute Myeloid Leukemia. *Clin Cancer Res* **26**, 3597-3607 (2020).
80. Mahfoudhi, E. *et al.* P53 activation inhibits all types of hematopoietic progenitors and all stages of megakaryopoiesis. *Oncotarget* **7**, 31980-92 (2016).
81. Olive, K.P. *et al.* Mutant p53 gain of function in two mouse models of Li-Fraumeni syndrome. *Cell* **119**, 847-60 (2004).
82. de Boer, J. *et al.* Transgenic mice with hematopoietic and lymphoid specific expression of Cre. *Eur J Immunol* **33**, 314-25 (2003).

Acknowledgments

We are grateful to patients and donors; without their generosity, this study would not have been possible. We also thank Steve Knapper, clinical study teams and other investigators involved in supporting sample collection, and King's Health Partners Biobank for providing access to samples. We thank William Vainchenker for his scientific input; Zemin Ren and Timothé Denaes for help with mouse experiments and Sally-Ann Clark for help with sorting. This work was funded by a Medical Research Council (MRC) Senior Clinical Fellowship (A.J.M.; MR/L006340/1), a CRUK Senior Cancer Research Fellowship (A.J.M.; C42639/A26988.), a Cancer Research UK (CRUK) DPhil Prize Studentship (C5255/A20936) to A.R-M, a British Spanish Society Scholarship to A.R-M., a MRC Confidence in Concept/MLSTF Grant to A.R-M. and A.J.M (MC_PC_19049), the MRC Molecular Haematology Unit core award (A.J.M and S.E.J. Eirik; MC_UU_12009/5), Emergence Cancéropôle Ile de France 2017 (I.A-D.), Association pour la Recherche contre le Cancer 2018 (I.A-D.), Siric-Socrate 2019 (I.A-D.), INCA-PLBIO 2020 (I.A-D.). A.L.C. was supported by Paris University (MENRT grant). The authors would like to acknowledge the flow cytometry facility at the MRC Weatherall Institute of Molecular Medicine (WIMM) which is supported by the MRC Human Immunology Unit; MRC Molecular Haematology Unit (MC_UU_12009); National Institute for Health Research (NIHR), Oxford Biomedical Research Centre (BRC); Kay Kendall Leukaemia Fund (KKL1057), John Fell Fund (131/030 and 101/517), the EPA fund (CF182 and CF170) and by the MRC WIMM Strategic Alliance awards G0902418 and MC_UU_12025. The authors acknowledge the contributions of Dr. Neil Ashley at the MRC Weatherall Institute of Molecular Medicine (WIMM) Single Cell Facility and MRC-funded Oxford Consortium for

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Single-Cell Biology (MR/M00919X/1). The authors would also like to acknowledge the contribution of the WIMM Sequencing Facility, supported by the MRC Human Immunology Unit and by the EPA fund (CF268), the Gustave Roussy flow cytometry platform and mouse facility. We also thank the Oxford Genomics Centre at the Wellcome Centre for Human Genetics (funded by Wellcome Trust grant reference 203141/Z/16/Z) for the generation and initial processing of the OmniExpress SNP array data. The results published here are in whole or part based upon data generated by the TCGA Research Network (https://www.cancer.gov/tcga) and the BeatAML team. The views expressed are those of the authors and not necessarily those of the National Health Service (NHS), the NIHR or the Department of Health.

Author contributions

A.R.M. conceived the project, designed and performed experiments, performed computational analysis, analysed data and wrote the manuscript. R.N., A.L.C., H.R., J.O.S., E.L. and A.P. designed, performed experiments and analysed data, W.W.W., G.W. and W.W.K. performed computational analysis. J.E.M. collected primary samples and clinical and bibliographic data. C.D. provided clinical data. C.B. and M.B. analysed SNP-Array data. J.O.S., C.B., N.S., F.G., F.P., I.P., M.D., C.H. provided patients samples, clinical data and scientific input. C.M., H.G. analysed and provided patients and PDX biological data (NGS and SNP-array). A.H. performed and analysed patient's targeted sequencing. S.E.J, B.P. and S.T provided scientific input and conceptualization. S.T. supervised computational analysis. I.A-D. conceived and supervised the project, analysed data and wrote the manuscript. A.J.M. conceived and supervised the project, provided clinical care and wrote the manuscript.

Competing Interests statement

A patent relating to the TARGET-seq technique is licensed to Alethiomics Ltd, a spin out company form the University of Oxford with equity owned by B.P. and A.J.M. The other authors declare no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.28.485984; this version posted March 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Materials & Correspondence. Requests for material(s) should be addressed and will be fulfilled by corresponding authors: Alba Rodriguez-Meira (albarmeira@gmail.com), Iléana Antony-Debré (ileana.antony-debre@gustaveroussy.fr) and Adam J. Mead (adam.mead@imm.ox.ac.uk).

Supplementary Tables

TableS1. Clinical and genetic details from healthy donors and patients included in the study.

TableS2. Differentially expressed genes between *TP53* multi-hit HSPCs and *TP53*-WT cells.

TableS3. Genesets used to calculate gene expression signature scores in TARGET-seq and publicly available bulk-transcriptomic datasets.

Table S4. Differentially upregulated genes in *TP53* multi-hit cells (globally or LSCs) and genes selected by lasso regression to derive p53-all-cells and p53-LSC signatures.

TableS5. Gene signatures from *TP53* mutant heterozygous HSPCs from CP-*TP53*-MPN and pre-*TP53*-AML patients.

TableS6. Primers used throughout the experiments presented in the manuscript.

TableS7. Antibodies used for all experiments presented throughout the manuscript.

TableS8. Summary of mutation-specific homozygous status statistical testing for *TP53*-sAML and CP *TP53*-MPN patients. Related to Fig.1b-f; Fig.4b; Extended Data Fig.2b-o; Extended Data Fig.9c-k.