Single-cell mutation analysis of clonal evolution in myeloid malignancies

Myeloid malignancies, including acute myeloid leukaemia (AML), arise from the expansion of haematopoietic stem and progenitor cells that acquire somatic mutations. Bulk molecular profiling has suggested that mutations are acquired in a stepwise fashion: mutant genes with high variant allele frequencies appear early in leukaemogenesis, and mutations with lower variant allele frequencies are thought to be acquired later1–3. Although bulk sequencing can provide information about leukaemia biology and prognosis, it cannot distinguish which mutations occur in the same clone(s), accurately measure clonal complexity, or definitively elucidate the order of mutations. To delineate the clonal framework of myeloid malignancies, we performed single-cell mutational profiling on 146 samples from 123 patients. Here we show that AML is dominated by a small number of clones, which frequently harbour co-occurring mutations in epigenetic regulators. Conversely, mutations in signalling genes often occur more than once in distinct subclones, consistent with increasing clonal diversity. We mapped clonal trajectories for each sample and uncovered combinations of mutations that synergized to promote clonal expansion and dominance. Finally, we combined protein expression with mutational analysis to map somatic genotype and clonal architecture with immunophenotype. Our findings provide insights into the pathogenesis of myeloid transformation and how clonal complexity evolves with disease progression.

The genomic landscape of myeloid malignancies has been well described, with a near-complete catalogue of putative driver mutations4–7. Although specific combinations of mutations have been investigated in preclinical models, there remains uncertainty about the co-occurrence and functional relevance of mutations at a clonal level. To analyse the clonal architecture of myeloid malignancies, we used a custom amplicon panel covering 31 frequently mutated genes to perform single-cell DNA sequencing8 (scDNA-seq; Supplementary Table 1). We sequenced 740,529 cells from 146 samples from 123 patients with myeloid malignancies: clonal haematopoiesis (CH), myeloproliferative neoplasms (MPN), or AML (Extended Data Fig. 1a). We queried samples from patients at diagnosis and/or relapse; the majority were from patients with relapsed or refractory disease (Extended Data Fig. 1b; Supplementary Table 2). The most common mutations identified with scDNA-seq were in DNMT3A (n = 62 patients), TET2 (n = 58 patients), NPM1 (n = 37 patients) and FLT3 (n = 32 patients), consistent with previous bulk sequencing studies9–11 (Extended Data Fig. 1c, d). Eighty per cent of patients had three or more mutations in or near exons (Extended Data Fig. 1e, f). Reconstructed variant allele frequencies (VAFs) from single-cell data correlated significantly with bulk sequencing (Pearson ρ = 0.84, P < 2.2 × 10−16; Extended Data Fig. 1g). scDNA-seq further identified rare mutations that were not present in bulk sequencing, which had significantly lower VAFs than mutations found in bulk sequencing (Extended Data Fig. 1h; P < 2.2 × 10−16).

Clonal architecture in myeloid malignancies

We next investigated disease subtypes, subdividing AML into samples with epigenetic mutations (DNMT3A, TET2, ASXL1 and/or IDH1 or IDH2) (here referred to as IDH1/2); collectively termed DTAI mutations, and samples with DTAI and co-mutated signalling effectors. The number of mutations per sample was significantly higher in AML than in MPN, and in MPN than in CH (Fig. 1a; false discovery rate (FDR) P ≤ 7.15 × 10−6 to P ≤ 0.067 for all indicated comparisons). The increase in mutations per sample was more pronounced in cases of...
AML with signalling effector mutations, specifically those in RAS and FLT3 (FDR \( P \leq 0.075 \)). We next explored clonal repertoire, with clones defined as cells with identical protein-encoding single-nucleotide variants (SNVs), and applied a bootstrapping approach to identify clones that included at least ten cells (Fig. 1b, Extended Data Fig. 2a). We observed a significant increase in clone number in samples from patients with AML compared to MPN or CH (Fig. 1c; FDR \( P = 1.37 \times 10^{-4} \) to \( P \leq 0.026 \) for all indicated comparisons), with the highest number of clones in FLT3-mutant AML samples (FDR \( P = 1.37 \times 10^{-4} \)). We assessed the diversity of clone size on a per-sample basis and observed a significant increase from CH or MPN to AML (FDR \( P \leq 0.008 \); Fig. 1d). Clonal diversity was higher in samples with mutations in RAS and FLT3 than in samples from patients with CH or MPN or in AML samples without mutations in RAS or FLT3 (FDR \( P \leq 0.039 \)). Despite the increase in complexity, most patients with AML had one (75.6%; 65/86) or two (10.5%; 9/86) clones that accounted for at least 30% of cells. We found a significant decrease in the relative size of the largest clone in different subtypes of AML, consistent with the presence of multiple clones with increased fitness (Fig. 1e; FDR \( P \leq 0.0102 \) to \( P \leq 0.074 \)). The increased clonal diversity in AML did not coincide with a difference in the number of mutations within the largest clone (Extended Data Fig. 2b, c). These data suggest that increased mutational burden within a clone is not the primary driver of clonal dominance.

**Mutation patterns in clonal architecture**

We next investigated whether specific genes were more likely to be mutated in the dominant clone (Fig. 2a, Extended Data Fig. 2d). We identified gene-specific contributions to clonal expansion, with IDH2, NPM1 and JAK2 mutations nearly always being found in the dominant clone, whereas FLT3 and RAS mutations were found only in minor subclones in some patients, and in dominant clones in others. The presence of a mutation in the dominant clone could be inferred from the VAF in some cases (Extended Data Fig. 2e), especially for JAK2, which has a known relationship between VAF and clonal dominance in MPN\(^2\). Mutational inclusivity and exclusivity patterns on a per-sample basis were consistent with previously reported associations (Extended Data Fig. 2f). Of the 80 AML samples with DTAI mutations, 52.5% harboured mutations in more than one epigenetic modifier (Fig. 2b). In nearly all cases, epigenetic regulator mutations were in the same clone, and in 81% of cases the co-occurring mutations were within the dominant clone, suggesting cooperativity between DTAI mutations (Fig. 2c). DTAI mutations did not occur in the same clone in samples from patients with CH, suggesting that early clonal expansion is commonly mediated by individual mutations in epigenetic regulators (Extended Data Fig. 3a). By contrast, co-occurring signalling mutations in genes such as RAS and FLT3 very rarely occurred within the same clone, and almost never within the dominant clone (Fig. 2d).

We further identified distinct patterns of mutational cooperation in AML samples with mutations in DNMT3A and/or IDH1/2 (Fig. 2e). Similar patterns of co-occurring signalling effector mutations were observed in IDH1 mutant clones and in DNMT3A–IDH1 co-mutant clones, whereas single-mutant DNMT3A clones had fewer signalling mutations. DNMT3A–IDH2 co-mutant clones showed similar signalling co-mutation burdens and patterns to single-mutant DNMT3A clones, distinct from IDH2 single-mutant clones. IDH2 mutant clones had an increased frequency of JAK2 and NRAS co-mutations and fewer FLT3 co-mutations than single-mutant DNMT3A clones or clones with DNMT3A–IDH2 co-mutations. We focused on six patients who harboured mutations in both DNMT3A and IDH1/2 and had concurrent...
Initiating mutations and clonal dominance

These data provided an opportunity to delineate the sequence of somatic genetic events during myeloid transformation, and to map these events on clonal expansion. We adapted a zygosity-sensitive (Extended Data Fig. 3c) Markov decision process with reinforcement learning to generate evolutionary trajectories. We identified optimal evolutionary trajectories, with progressive clonal dominance and subclonal propagation (Fig. 3b).

We next assessed the fraction of the clonal architecture explained by a particular genetic trajectory to predict the disease-initiating mutation in samples with DTAI mutations (Fig. 3c). The majority of states were reconstructed when mutations in epigenetic modifiers such as DNMT3A and/or IDH1/2 were the initiating mutation(s). Conversely, very little of the clonal trajectory could be formed if the first mutation occurred in a signalling gene such as NRAS or FLT3. This observation was highly correlated to the computed VAF from scDNA sequencing (Spearman’s ρ = 0.93; P ≤ 2.2 × 10−16; Extended Data Fig. 3d). The notable exception was TET2, which could serve as the disease-initiating mutation or as an acquired mutation during clonal progression, consistent with studies in patients with MPN or post-MPN AML that have suggested a context-specific effect of TET2 loss-of-function during myeloid transformation and clonal evolution12,13. We next examined which gene mutations were observed as initiating, single-mutant clones and found that single-mutant clones with a DTAI mutation were commonly identified, confirming these as likely clone-initiating mutations (Extended Data Fig. 3e).

Fig. 2 | Elucidation of clonal dominance and co-mutation by single-cell DNA sequencing. a. Top, box plot (centre line, median; box, IQR; whiskers, 1.5×IQR) indicating the fraction of cells from a sample in the largest mutant clone. Red dots, dominant clones; black dots, subclones. Bottom, proportion of mutant clones for which the indicated gene is mutated in the most dominant identified clone (red bars; n = 485 clones, n = 111 samples). b, Upset plot of co-occurring mutations for AML samples with mutations in DTAI genes. Bar graph depicts number of samples with each mutant gene(s). Presence in dominant clone (red) or subclones (grey) is indicated. Grid (bottom) indicates combinations of mutations in each corresponding bar plot. c, d, Co-occurrence spectra of DTAI mutations (c) and signalling mutations (d). Size of vertex represents number of samples mutated for given gene. Edge colour denotes dominant clones (red) and subclones (grey), with edge width representative of clone size. e, Fractions of clones with co-occurring signalling mutations in DNMT3A, IDH1, and IDH2 in patients with AML. Different signalling mutations are coloured as indicated.
We next focused on discerning the order of subsequent mutations during clonal evolution and their contribution to clonal expansion and dominance. For the majority of samples (n = 19/23) with co-occurring DNMT3A–IDH1 and DNMT3A–IDH2 mutations, we observed a significant increase in relative clone size compared to single-mutant DNMT3A clones (IDH1 P = 0.00023; IDH2 P ≤ 2.16 × 10−4; Fig. 3d; Extended Data Fig. 3h) or single-mutant IDH1 or IDH2 clones (P ≥ 0.0016 and P = 1.37 × 10−8, respectively). In NPM1 mutant samples with co-occurring FLT3 mutations, the clone size of FLT3–NPM1 double-mutant clones was significantly greater than that of FLT3 mutant clones (P = 0.0097) or NPM1 mutant clones (P = 0.0089) single-mutant clones (Fig. 3d). By contrast, for RAS–NPM1 co-mutant clones we observed substantial variability in clone size and less evidence of cooperativity compared to single-mutant NPM1 clones (P = 0.462), whereas the double-mutant clone was significantly larger than RAS mutant-only clones (P ≤ 0.0009). This finding suggests that different combinations of co-occurring gene mutations vary in their capacity to promote clonal dominance in AML, even for genotypes that are commonly observed in bulk sequencing.

Clonal evolution in myeloid malignancies

We next sought to determine whether clonal architecture is altered during disease transformation and response to therapy. In four out of six patients in whom the disease transformed from MPN to AML, we observed a significant alteration in clonal architecture, or a ‘clonal sweep’, with emergence of new dominant clone(s) (Extended Data Fig. 4a). In samples MSK7/76, the dominant CALR–ASXL1 mutant clone in the MPN (sample A) was replaced by a CALR–ASXL1–IDH1 dominant clone at transformation (sample B), which was a minor clone in the MPN phase (Fig. 3e). We next tested pre- and post-therapy samples from three patients bearing FLT3 mutations who were treated with the FLT3 inhibitor gilteritinib. All three patients showed a decrease in FLT3-mutant clones in response to gilteritinib, with significant clonal sweeps (Extended Data Fig. 4b, c). In two of the three patients, we observed outgrowth of clones with RAS mutations, which has previously been described as a potential resistance mechanism to FLT3 inhibitor therapy, often with RAS mutations acquired in the FLT3 mutant clone5. In samples MSK82/83, we observed diminution of clones with FLT3 internal tandem duplication (ITD) mutations and expansion of FLT3 wild-type RAS mutant clones (Extended Data Fig. 4b). In another patient (samples MSK95/96), two FLT3 wild-type U2AF1–RAS mutant clones (KRASG12D and NRASG12D) achieved clonal dominance during FLT3 inhibitor therapy (Extended Data Fig. 4c). Meanwhile, a FLT3–KRAS mutant clone was suppressed following therapy. These results indicate that transformation and therapeutic perturbations can alter clonal architecture in both a linear and branched manner.

Simultaneous scDNA-seq and immunophenotyping

We next investigated whether specific mutations or combinations influenced immunophenotypes. We performed simultaneous scDNA-seq and cell-surface protein expression analysis17 in patients with CH with one or more mutation(s), to investigate the contribution of CH mutations to mature haematopoietic lineages. We observed differential contributions to the B and T cell lineages, depending on which CH gene was mutated. In four out of four patients with DNMT3AIDH1 mutations, the mutant clone contributed minimally to the B (CD19 high) and T cell (CD3 high) lineages, consistent with restricted myeloid (CD11b high) bias (Fig. 4a, Extended Data Fig. 5). Conversely, non-R882 DNMT3A mutations (for example, DNMT3ADNMT3AIDH1 mutations) were more strongly represented in the T cell lineage, but less in the myeloid and B cell lineages. Allele-specific lineage skew was even observed in patients with more than one CH clone; we observed differential lineage contributions of DNMT3AIDH1 and DNMT3AIDH1 mutant clones within the
same patient. By contrast, TET2 mutations offered less consistent results, with some mutants (2/4) showing myeloid lineage bias and others (2/4) making a balanced contribution to all mature lineages (Extended Data Fig. 5).

Recent single-cell RNA-sequencing results have highlighted a continuum of differentiation states in AML. To assess clonal architecture and differentiation state, we analysed simultaneous scDNA-seq and immunophenotype in samples (n = 17) from patients with AML. We observed significant differences in protein expression between wild-type and mutant cells, with wild-type cells expressing high levels of CD34 (P ≤ 2.2 × 10−16) and low levels of CD34 (P ≤ 2.2 × 10−16) compared to mutant cells (Extended Data Fig. 6a, b). With respect to different mutations, cells with mutations in TET2 (P ≤ 1.38 × 10−6), RUNX1 (P ≤ 9.48 × 10−10), IDH1 (P ≤ 2.2 × 10−16), or JAK2 (P ≤ 2.2 × 10−16) were enriched for high CD34 surface expression (Fig. 4b), whereas cells bearing mutations in the MAPK/ERK signalling pathway (NRAS P ≤ 0.04, KRAS P ≤ 2.2 × 10−16 and PTPN11 P ≤ 2.2 × 10−16) had higher expression of CD11b compared to cells with other mutant genes. Moreover, NPM1 mutant cells harboured lower expression of CD34 than all other mutant cells (P ≤ 2.2 × 10−16), consistent with previous flow cytometric data. The high CD34 expression seen in IDH1 mutant cells decreased in cells with co-mutations in signalling effectors (P ≤ 2.2 × 10−16), and CD11b expression was increased in IDH1–RAS co-mutant subclones (P ≤ 2.2 × 10−16).

We expanded this analysis to assess how immunophenotypes differed across distinct genetic clones. We observed co-occurring mutation-specific changes in expression, including increased CD11b expression in RAS mutant subclones compared to RAS wild-type subclones (Fig. 4c, Extended Data Fig. 7; sample MSK71). We also observed reduced CD34 expression in DNMT3A–FLT3 co-mutant clones compared to DNMT3A single-mutant clones, with a concomitant increase in CD38 and CD45RA expression, consistent with a myeloid progenitor phenotype (Extended Data Fig. 7; sample MSK71). To summarize combinatorial differences in immunophenotype, we clustered cells into eight communities present in MSK71, coloured as in d. Heat map depicts CLR normalized expression of indicated proteins for each community from d.
which was marked by the highest level of CD11b expression. Meanwhile, a KRAS mutant clone showed increased representation in community 4, associated with high CD19 expression (Extended Data Fig. 8a).

To determine whether patterns of immunophenotype changes existed across multiple samples, we merged all samples, clustered cells according to cell-surface protein expression, and then identified communities of cells (Extended Data Fig. 8b, c). We found that multiple overlapping immunophenotypic states occur across samples with divergent genotypes; no community was exclusive to an individual sample, and six communities were observed in every sample (communities 7, 8, 9, 18, 32, and 42); these communities were intercorrelated with high expression of either CD90 or CD38 (Extended Data Fig. 9a). We observed significant shifts in community representation between the dominant clone and subclones in 8 out of 14 samples with more than one leukemic clone (Extended Data Fig. 9b). In contrast to the increases in CD11b expression, which were specific to NRAS mutant clones, we observed in sample MSK130 that an FLT3/FLT3 mutant dominant clone showed expansion of a community with high expression of CD34 ($P \leq 2.2 \times 10^{-15}$) and low expression of CD11b ($P \leq 2.2 \times 10^{-16}$; Extended Data Fig. 9c). Furthermore, a JAK2 mutant sample (MSK94) contained expanded communities with high CD38 and low CD11b expression in the dominant clone compared to subclones (CD38 and CD11b, both $P \leq 2.2 \times 10^{-15}$). These findings suggest that the acquisition of signaling effector mutations induces divergent clone-specific changes in immunophenotype.

**Discussion**

The identification of frequent, recurrent mutations in epigenetic regulators in patients with CH and the lower incidence of overt myeloid malignancies relative to CH suggest that the rate-limiting step in myeloid transformation is clonal evolution from disease-initiating clones to leukemic clones. Previous studies have used bulk sequencing analyses to predict important features of clonal evolution and; however, the molecular sequence of events that drive myeloid transformation has not been dissected at a single-cell, clonal level. Here we have used scDNA-seq to map clonal evolution in myeloid malignancies, and to achieve insights into the pathogenesis of myeloid transformation that were not previously discernible by bulk sequencing.

First, we found that clonal complexity increases from CH or MPN to AML and continues to evolve as AML clones acquire mutations in signalling effectors. By contrast, signalling effector mutations were often subclonal, and very rarely co-occurred in the same clone. Second, we observed significant differences in how mutational combinations contributed to clonal dominance; specific co-occurring disease alleles (for example, NPM1–FLT3/ITD or DNMT3A–IDH2) were associated with clonal dominance, whereas other mutational combinations (NPM1–RAS) did not promote clonal expansion. Analysis of paired samples in the context of the evolution of disease from MPN to AML and in the setting of therapeutic perturbation show that myeloid malignancies are characterized by clonal sweeps in the setting of specific stressors, and that the changes in clonal architecture are largely due to expansion of pre-existing minor clones. These data have biological and therapeutic relevance, as the clones that emerge with transformation (for example, expanding IDH2 mutant clones) or therapeutic selection (FLT3 wild-type, RAS mutant) can be detected using scDNA-seq and may inform the use of therapies that target these clones before they achieve clonal dominance. Last, we identified significant genotype-driven changes in cell-surface protein expression using simultaneous single-cell mutational profiling and immunophenotyping. Signalling effector mutations were particularly notable for altering cell-surface protein expression, with mutations in the MAPK/ERK pathway leading to increased CD11b expression.

Together, these data suggest that myeloid malignancies manifest as a complex ecosystem of clones that evolves over time, and that scDNA-seq gives a glimpse into this milieu that is not seen with conventional bulk sequencing. Our studies of clonal architecture at a single-cell level give us insights into how clonal complexity contributes to the pathogenesis of myeloid transformation (see Supplementary Discussion). Similar studies across different pre-malignant and malignant contexts will give new information about how malignancies initiate and progress and will lead to new therapeutic strategies aimed at intercepting clonal evolution and/or targeting cancer as a multi-clonal disease.
Methods

Reagents
All antibodies for flow cytometry were purchased from Biolegend. These studies used the following antibodies: FITC–CD3 (clone UCHT1), FITC–CD19 (clone HIB19) and FITC–CD56 (clone HCD56). Human TruStain FcX was also purchased from Biolegend. The DNA+Protein oligo-conjugated antibodies were produced and provided by Biolegend as part of a collaboration with Mission Bio, Inc. The antibodies in the conjugate pool were the following: CD3 (clone SK7), CD11b (clone RCRF44), CD19 (clone HIB19), CD34 (clone SS1), CD38 (clone HIT2), CD45RA (clone HI100), and CD90 (clone SE10). Antibody conjugates were pooled in equimolar ratios. All Tapestry related reagents were included as part of a custom single-cell DNA sequencing kit purchased from Mission Bio, Inc. The custom ampiclon panel used in these studies covers 109 amplicons over 31 genes previously found to be frequently mutated in human myelodysplastic syndromes (MDS), myeloproliferative neoplasms (MPN), and acute myeloid leukaemia (AML)\(^\text{14,15}\) (Supplementary Table 1).

Patient samples
Patients with myeloid neoplasms or acute myeloid leukaemia between 2014 and 2019 were studied. Informed consent was obtained from patients according to protocols approved by the institutional IRBs and in accordance with the Declaration of Helsinki. This study was approved by MSKCC Institutional Review Board (protocol #15-017) and Thomas Jefferson University (TJU) Institutional Review Board (protocol #17D.083). Diagnosis and disease status were confirmed and assigned according to World Health Organization (WHO) classification criteria\(^\text{16}\). Patient characteristics are summarized in Supplementary Table 2 and Extended Data Fig. 1a, b. With the exception of four complex karyotype/TP53 mutant samples (denoted with asterisks), all samples were confirmed as normal karyotype. Bone marrow from healthy individuals was obtained with informed consent according to procedures approved by the institutional review boards of Memorial Sloan Kettering Cancer Center and Hospital for Special Surgery. Patient samples were collected and processed by the MSKCC Human Oncology Tissue Bank (HOTB) or TJU Haem Malignancy Repository. Mononuclear cells were obtained by centrifugation on Ficoll from peripheral blood or bone marrow and viably frozen. Patient samples from MSKCC underwent high-throughput genomic sequencing with a targeted deep sequencing assay of 685 genes (HemePACT) or by an NGS platform panel composed of 49 genes that are recurrently mutated in myeloid disorders (RainDance Technologies ThunderBolts Myeloid Panel). Single point variants were called using Mutect and short insertions and deletions using Pindel as described previously, comparing samples to a sample representing a pool of normal samples\(^\text{28}\). Mutations were excluded if found to be present in at least one database of known non-somatic variants (dbSNP and 1000 genomes) and absent from COSMIC. Samples with non-excluded mutations with variant allele frequency >2% were classified as clonal haematopoiesis. Samples were selected based on mutation coverage by the Mission Bio Custom amplicon panel, variant allele frequencies of all covered mutations (>5% VAF for each gene covered on panel), and number of cells collected (>5 x 10⁶ cells) per frozen aliquot. Specifically, samples were prioritized if they harboured 1) more than one mutation in epigenetic modifier genes DNMT3A, TET2, ASXL1, or IDH1/2; 2) an NPM1 mutation; 3) mutations in NRAS or KRAS; and/or 4) mutations in FLT3 (either internal tandem duplication (ITD) or tyrosine kinase domain (TKD) mutations).

Single-cell DNA sequencing library preparation and sequencing
Patient samples were thawed, washed with PBS buffer, and quantified using a Countess cell counter. Cells (1.0–4.0 x 10⁶ viable cells) were then resuspended in DPBS (Gibco) and incubated with TruStain FcX, dextran sulfate (100 μg/ml; Research Products International), and 1x Tapestry staining buffer for 3 min at room temperature. The pool of seven oligo-conjugated antibodies (CD3, CD11b, CD19, CD34, CD38, CD45RA, and CD90) was then added and incubated for 30 min at room temperature. Cells were then washed multiple times with DPBS supplemented with 5% fetal bovine serum (FBS; Gibco) followed by resuspension of the cells in Tapestry cell buffer, requantification, and loading of the cells into a Tapestry microfluidics cartridge. Single cells were encapsulated, lysed and barcoded as above with the exception of adding an additional forward primer mix (30 μM each) for the antibody tags before barcoding. DNA PCR products were then isolated from individual droplets and purified with Ampure XP beads. The DNA PCR products were then used as a PCR template for library generation as above and repurified using Ampure XP beads. Protein PCR products (supernatant from Ampure XP bead incubation) were incubated with Tapestry pullout oligo (5 μM) at 96 °C for 3 min, followed by incubation on ice for 5 min. Protein PCR products were then purified using Streptavidin C1 beads (Invitrogen) and beads were used as a PCR template for the incorporation of 15/17 Illumina indices followed by purification using Ampure XP beads. All libraries, both DNA and protein, were quantified using an Agilent Bioanalyzer and pooled for sequencing on an Illumina NovaSeq by the MSKCC Integrated Genomics Core.

Data analysis
Data processing. FASTQ files for single-cell DNA libraries were analysed through the Tapestry pipeline using Bluebee’s high-performance genomics platform. In brief, this pipeline trims adaptor sequences, aligns reads to the human genome (hg19), assigns sequence reads to cell barcodes, and performs genotype calling with GATK v3.7. Data are then consolidated into a multiple sample VCF file and output as a loom file for subsequent processing. Initial steps for filtering low-quality genotypes or cells were performed in Tapestry Insights with default parameters and R, where the minimum variant quality score was set to 30 with a minimum of 10 reads per variant per cell. We further removed variants present in <50% of cells and removed cells in which <50% of potential variants reported informative genotypes. Data were exported from Tapestry Insights and subsequent filtering was performed in R. For DNA analysis on the DNA+Protein platform, we used the Tapestry pipeline on Bluebee as described above. For the protein analysis, custom scripts in R were used by Mission Bio to enumerate the number of reads per antibody per cell. Subsequent normalization was performed using the Tapestry package in R. Variants were filtered through an empirically curated banned list of panel-specific mutations that were not identified in bulk sequencing nor present in COSMIC. We further removed variants constrained to one problematic
hyper-mutated amplicon (chr20:31,022,898–31,023,107) and focused all subsequent work on protein-encoding, non-splicing mutations. Variants were included if there were at least two cells that were heterozygous or homozygous. Samples were included if they harboured one or more protein-encoding, non-synonymous/insertion/deletion variants and more than 100 cells with definitive genotype for all protein-coding variants within the sample. We next sought to define genetic clones, which we identified as cells that possessed identical genotype calls for the protein-encoding variants of interest. In order to focus our analyses on reproducible clones, we performed a bootstrapping analysis over 10,000 samplings to calculate 95% CIs for the presence of each clone. Clonal analyses in Fig. 1b and onward focus on clones for which the lower 95% CI was > 10 cells. We further excluded rare variants that were identified only in clones that did not pass this threshold. Samples were included in clonal analyses if they included more than two protein-encoding variants and more than two clones. A flowchart of sample inclusion can be found in Extended Data Fig. 2a, and patient characteristics can be found in Supplementary Table 2. Dominant clones (as referred to in the text) were defined as the largest mutant clone in the sample, excluding cells that were wild-type for all variants of interest.

Genetic trajectory analysis. For the genetic trajectory analysis constructed in Fig. 3, we implemented a Markov decision process with reinforcement learning. Generally, this allowed us to model the optimal track of mutation acquisition if a cell were to acquire one mutation at a time and not revert that mutation to a wild-type state. Technically, for a given sample, we first constructed a reward matrix by enumerating all possible clones given the number of mutations present in a sample, and the maximum zygosity for a given mutant (that is, if we did not observe a homozygous state for a mutant, it was not considered in the reward matrix). After construction of the reward matrix, we set permissible decision processes with a value of 0, and impermissible decision processes with a value of −1 (that is, decisions where a mutation was reverted to wild-type or required more than one genetic alteration were penalized). Decisions were considered permissible if a clone was separated by a single genetic event, either a variant changing from wild-type to heterozygous or homozygous to homozygous. For observed clones, the frequency of the clone (ranging from 0 to 100% of cells) was used as the value in the reward matrix, while unobserved clones retained a value of 0. The matrix was then converted to long form and state transitions between clones were associated with the action/mutation causative to that state change. This was then used as input to the ReinforcementLearning package in R to generate a Q matrix through the experience replay algorithm32. Custom scripts in R were used to navigate this Q matrix to determine the optimal trajectory from the wild-type clone.

Statistical analysis. Statistical significance was evaluated using a two-sided Student’s t-test and two-sided Fisher’s exact test where indicated. Multiple test correction was implemented using the Benjamini–Hochberg/FDR approach as indicated. Shannon diversity index was assessed using the diversity function in the vegan package in R30. Genetic co-occurrence analysis was performed using the cooccur package in R. UMAP clustering was performed using the R package umap, with default parameters30. Subsequent community analysis was performed using phenograph implemented with the Rphenograph package31,32. The perplexity factor K was set to 50. For multiple comparisons, a range of significant P values or FDR values have been provided for clarity. Complete P values and measures of significance can be found with the publicly available code below.

Plotting and graphical representations. All bar plots, box plots, heat maps and scatter plots were produced using the ggplot2 package in R34. Error bars depict standard error of the mean. Box plots are depicted in Tukey’s style with boxes representing the median and quartile range, with whiskers representing ±1.5 x IQR. The oncoprint presented in Extended Data Fig. 1a was produced using the ComplexHeatmap package in R35. Upset plots shown in Fig. 2b and Extended Data Fig. 3a were produced using the UpsetR package36. Network plots in Figs. 2c, d, 3a, b were produced with the igraph package in R37. UMAP data were plotted using the ggplot2 package. Other packages used in data processing include tidyr, dplyr, RColorBrewer, pals, and cowplot.

Rigour and reproducibility
Sample inclusion criteria are described in detail above. In brief, high-quality variants were selected with a minimum GATK quality score of 30, and 10 reads supporting each variant. Variants and cells were filtered if incomplete genotype information was present for all variants of interest as described above. Variants on a subset of samples were visually inspected on IGV to ensure mutation caller fidelity. If fewer than 100 informative cells were present in a sample, it was removed from the analysis to filter out low quality. Rigorous evaluation of clonal abundance was estimated with a bootstrapping approach to establish 95% CIs. Clones with a lower CI (more than 10 cells) were retained for analysis. Duplicate aliquots from select samples were processed on different days to assess the replicability of the tapestrí platform. To enable reproducibility and transparency, all code and data are available as described below.

Reporting summary
Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability
Raw data are available on dbGAP (accession number phs002049.v1.p1) in the form of loom files and FASTQ files for each sample.

Code availability
All scripts and processed data files are available at https://github.com/bowman/scDNA_myeloid.

Acknowledgements
We acknowledge the use of the MSKCC Integrated Genomics Core for all library sequencing, which is funded by MSKCC Support Grant NIH P30 CA008748. We thank members of the Levine laboratory for their critique of our work and assistance with revisions, and M. Rosal and W. Xiao for their input regarding AML stem and progenitor cell protein expression. L.A.M. is supported by a Career Development Program Fellowship of the Leukemia and Lymphoma Society (5479-19) and a Postdoctoral Fellowship from the MSKCC Marie-Josée Kravis Women in Science Endeavor (WISE). A.L.B. is supported by the Sohn Foundation Fellowship of the Damon Runyon Cancer Research Foundation (DRG 22-17) and a National Cancer Institute grant (K99 CA248460). C.L.D. is supported by a Swiss National Science Foundation fellowship (grant no. 183853). K.L.B. is supported by grants including a National Cancer Institute grant (K99 CA248460).
Institute of Health grant (K08 CA241318), an American Society of Hematology (ASH) grant, and an EvansMDS grant. A.D.V. is supported by the William Raveis Charitable Fund Fellowship of the Damon Runyon Cancer Research Foundation (DRG 117-15), an EvansMDS Young Investigator grant from the Edward P. Evans Foundation, and a National Cancer Institute career development grant (K08 CA215317). This work is supported by grants to S.E.M. including National Cancer Institute R37 CA226433, a Conquer Cancer New Award from the Concern Foundation, and Sidney Kimmel Cancer Center (SKCC) Support Grant NIH P30 CA056036. This work was supported by grants to R.L.L. including a Cycle For Survival Innovation Grant, National Cancer Institute R35 CA197594, National Cancer Institute ROI CA173636, a grant from the Samuel Waxman Cancer Research Foundation, and SCOR grants from the Leukemia and Lymphoma Society.

Author contributions L.A.M., R.L.B., A.D.V., and R.L.L. conceptualized studies. L.A.M., R.L.B., A.T.O., R.D.-D., P.M., C.A., B.D., C.L.D., A.R.A., M.M., S.S., M.B., and R.L.L. designed and optimized experimental methodologies and bioinformatic workflow. L.A.M., R.L.B., T.R.M., I.S.C., and A.T.O. performed experiments. C.F., M.A.P., A.D.G., K.L.B., A.Z., R.R., M.P.C., S.E.M., A.D.V., and R.L.L. provided patient-associated resources and/or patient samples for use in the studies. L.A.M., R.L.B., and R.L.L. wrote the original manuscript. L.A.M., R.L.B., S.E.M., and R.L.L. assisted with review and editing of the manuscript. R.L.L supervised studies and manuscript preparation.

Competing interests L.A.M. and A.D.V. received travel support and honoraria from Mission Bio. A.T.O., R.D.-D., P.M., C.A., M.M., and S.S. are employed by Mission Bio and own equity in Mission Bio. A.R.A. is a cofounder and shareholder of Mission Bio. A.Z. has received honoraria from Illumina. M.P.C. has consulted for Janssen Pharmaceuticals. A.D.G. has served on advisory boards or as a consultant for AbbVie, Aptose, Celgene, Daiichi Sankyo, and Genentech, received research funding from AbbVie, ADC Therapeutics, Aprea, Aptose, AROG, Celularity, Daiichi Sankyo, and Pfizer, and received honoraria from Dava Oncology. R.R. has consulted for Constellation, Incyte, Celgene, Promedior, CTI, Jazz Pharmaceuticals, Blueprint, Stemline, Galecto, Pharmessentia, and Abbvie, and received research support from Incyte, Stemline, and Constellation. A.D.V. is on the Editorial Advisory Board of Hematology News. R.L.L. is on the supervisory board of QIAGEN and is a scientific advisor to Mission Bio, Loxo (until February 2019), Imago, C4 Therapeutics, and Isoplexis. He receives research support from and consulted for Celgene and Roche and has consulted for Lilly, Jubilant, Janssen, Astellas, Morphosys, and Novartis. He has received honoraria from Roche, Lilly, and Amgen for invited lectures and from Celgene and Gilead for grant reviews. R.L.B., T.R.M., I.S.C., C.F., M.A.P., M.B., B.D., C.L.D., K.B., and S.E.M. disclose no competing interests.

Additional information Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2864-x.

Correspondence and requests for materials should be addressed to R.L.L.

Peer review information Nature thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.
Extended Data Fig. 1 | See next page for caption.
Extended Data Fig. 1 Single-cell DNA-sequencing patient cohort. a. Oncoprint of patient samples analysed by scDNA-seq. b. Table describing patient cohort characteristics. Standard deviation calculated for mean age of patients at sample collection date. Absolute number of samples denoted with percent of total samples in parentheses. c. Number of individual mutations identified for each gene covered on our custom amplicon panel by scDNA-seq ($n = 146$ biologically independent samples for c-f). Genes are ranked by the number of identified protein coding mutations from highest to lowest. Genes with zero identified mutations are not listed. d. Number of patients with protein coding mutations in a given gene. Genes are ranked by decreasing number of patients identified with mutations. e. Number of patients with a given number of identified mutant genes via single-cell sequencing. f. Number of patients with a given number of identified protein altering variants via single-cell sequencing. g. Correlation of bulk sequencing SNV data VAF versus single-cell SNV data VAF from MSKCC samples. Statistical significance was calculated by Pearson correlation coefficient. h. Violin plot of computed VAF from scDNA-seq for mutations found in both scDNA-seq and in bulk sequencing (identified; red), or mutations only identified in scDNA-seq (missed; blue) (top panel). Samples identified by scDNA-seq only were found to be low VAF mutations ($P < 2.2 \times 10^{-16}$; two-sample Mann–Whitney test). Bar plot of the number of new mutations in each sample identified by scDNA-seq only (bottom panel).
Extended Data Fig. 2 | See next page for caption.
Extended Data Fig. 2 | Analysis of clonal architecture by disease type and gene mutation. 

a, scDNA-seq data processing and analysis workflow. FASTQ sequencing files for each sample were uploaded and processed through Mission Bio Tapestri Insights platform for variant calling and cell finding (Commercial Platform). Included samples for further analysis harboured ≥ 1 variant which leads to a protein sequence change (non-synonymous/insertion/deletion) and included 50 cells with definitive genotyping for all protein coding variants within the sample (n = 146). This data was used for analysis in Fig. 1. Clones present in each sample were identified and samples removed if they contained less than 2 clones for clonal analysis studies. Samples were subjected to random resampling of cells using a bootstrapping approach to identify the stability of identified clones (n = 132). Following bootstrapping, clones with lower 95% confidence intervals >10 were removed as were variants identified only within those clones. Samples which harboured only 1 variant or presented with <2 clones after bootstrapping analysis were removed (n = 111). The number of samples at each step of processing is shown below the different steps of the workflow. 

b, Number of mutations in the most dominant clone identified in each sample (n = 111 biologically independent samples) stratified by cohort. Mean value for each cohort shown by height of bar with standard error of measurement (SEM) depicted with error bars. A two-sided t-test with FDR correction was used to determine statistical significance pairwise between all groups. For clarity, only significant P values referenced in the text are shown. *P < 0.1; **P < 0.01; ***P < 0.001. 

c, Association between clone size and the number of mutant alleles in the clone. Every clone (n = 111 biologically independent samples) identified in clinical cohort is depicted by black circle. Centre line: median; box: IQR; whiskers 1.5 × IQR. 

d, Bar plot depicting the prevalence of dominant clones for each DTAI gene across patient cohorts. Colour of bar plot annotates if mutation occurs in the dominant clone (red) or subclone (grey). Absence of bar denotes no clones were identified with the indicated mutation in a given cohort. 

e, Association of VAF with presence of mutation in either the dominant clone (red) or subclone (grey) for select genes (n = 101 biologically independent samples). Standard error of measurement depicted with error bars. A two-sided t-test with FDR correction was used to determine statistical significance pairwise between all groups. *P < 0.1; **P < 0.01; ***P < 0.001. Absence of P value for IDH2 and JAK2 due to lack of samples with subclonal mutations. 

f, Pairwise interaction matrix of mutually exclusive (red square) and inclusive (blue square) on a per-sample basis. Pairwise interactions with no colour did not garner a significant P value.
Extended Data Fig. 3 | See next page for caption.
Extended Data Fig. 3 | Clonal dominance, initiating mutation, and co-mutation patterns in patients with myeloid malignancies. 

a, Upset plot of co-occurring DTAI mutations in CH samples with more than one DTAI variant. Bar graph (top panel) depicts the number of samples with each mutant gene(s) and colour of bar annotating whether mutation(s) occur in the dominant clone (red) or subclones (grey). Black circles and connecting line in bottom panel demark the combination of mutations in each corresponding bar plot.

b, Divergent frequency of co-mutated cells for epigenetic modifier genes (red) and signalling genes (blue). Individual samples (n = 6 samples) shown with black square. Centre line: median; box: IQR; whiskers 1.5 × IQR. A two-sided Student’s t-test was used to determine statistical significance *P < 0.1; **P < 0.01; ***P < 0.001.

c, Fraction of mutant samples harbouring a homozygous mutation for the indicated given gene (at least >10% of cells). Homozygous sample denoted in blue.

d, Correlation of VAF computed by scDNA sequencing to fraction of a mutant sample explained by the genetic trajectory starting with an initiating mutation in a given gene. Genes used as the initiating mutation for a given sample are denoted by colored squares (colours described in figure). Statistical significance calculated by Spearman’s rank correlation coefficient test (p = 0.93; P ≤ 2.2 × 10⁻¹⁶).

e, Number of samples where a monoallelic clone for a given gene is observed. Dark blue denotes total number of mutant samples where single-mutant clone is present for a given gene and grey represents mutant samples where single-mutant clone is unobserved.

f, Number of DNM34 mutant samples where single-mutant clones are observed (red) or unobserved (grey) with samples categorized by DNM34 R882 hotspot mutations, nonsense mutations, or missense mutations. A two-sided Fisher’s exact test was used to determine statistical significance (P ≤ 0.04) between DNM34 R882 and other missense mutations.

g, Differences in dominant and subclone size in DNM34 mutant samples (n = 61 biologically independent clones). Fraction of sample in the dominant clone or subclone(s) for DNM34 nonsense (red), R882 missense (green), and non-R882 missense (blue) mutations shown. Centre line: median; box: IQR; whiskers 1.5 × IQR. Each mutant clone denoted by black square. A two-sided t-test correction was used to determine statistical significance pairwise between all groups. For clarity, only significant P values referenced in the text are shown.

h, As in Fig. 3e, fraction of sample in single- and double-mutant clones in DNM34/IDH2 mutant samples. Each sample is indicated by a connecting line, absence of a line for single mutants indicates absence of clone.
Extended Data Fig. 4 | See next page for caption.
Extended Data Fig. 4 | Clonal evolution in patients with myeloid malignancies. 

**a**, Paired samples from patients (n = 6) that underwent MPN to AML transformation were analysed. Samples with significant changes in clonal architecture or clonal sweeps were evaluated using a two-sided two proportions z-test; ***P < 0.001. Sample A (red) denotes the MPN sample and sample B (blue) denotes the AML sample. Clonotype plot depicts the frequency of a clone with given genotype in Sample A and B ranked by decreasing frequency based on Sample A (top panel). Heat map (bottom panel) shows the genotype of each identified protein coding mutation in the given clone with zygosity (wild type = light pink, heterozygous = orange, homozygous = red). Paired samples MSK75/76 are highlighted in Fig. 3f.

**b**, Clonal sweeps, or significant clonal architecture alterations, following gilteritinib therapy of FLT3-mutant patients (n = 3). Line graphs for each pair of samples depict individual clones and the change in clone frequency between pre- (left) and post- (right) therapy samples. Clones harbouring FLT3 mutations (red), RAS mutations (blue), or wild-type (WT) clones (light blue) are significantly altered after gilteritinib therapy in each patient. FLT3-RAS mutations (orange) and clones harbouring additional mutations (Other; grey) are also included. Statistical significance was assessed using a two-sided two proportions z-test; ***P < 0.001 (a, b).

**c**, As in a, clonotype plot of paired sample (n = 1 sample/time point) from a patient with AML (MSK95/96) who underwent gilteritinib therapy: sample A (red, pre-therapy) and sample B (blue, post-therapy).
Extended Data Fig. 5 | Contribution of clonal haematopoiesis mutations to mature cell lineages. Bar graphs of the mutant cell percentage found in myeloid (CD11b high; green), B (CD19 high; orange), and T (CD3 high; purple) cells in samples from patients with CH. *DNMT3A* and/or *TET2* mutations found in each sample are listed above each graph. Double-mutant samples are shown on the left and single-mutant samples are depicted on the right.
Extended Data Fig. 6 | Simultaneous molecular and immunophenotypic profiling of samples from patients with AML. a, UMAP plot of MSK54 with cells clustered by immunophenotype. Genotype (wild type = grey; \( \text{DNMT3A} \) = red; \( \text{IDH2} \) = green; \( \text{DNMT3A/IDH2} \) double mutant = blue) overlaid onto each cell. b, UMAP from a with protein expression (high expression = red; low expression = blue) for each of the six antibody targets (CD3, CD11b, CD34, CD38, CD45RA, CD90) overlaid onto each cell. Relative protein expression is normalized across individual samples by CLR. c, Immunophenotype changes based on co-occurring mutations in clones. Heat map of normalized protein expression of CD34 (top panel) and CD11b (bottom panel) in \( \text{DNMT3A} \) and \( \text{IDH1/2} \) single-mutant clones versus \( \text{DNMT3A} \) and \( \text{IDH1/2} \) mutant clones with co-occurring \( \text{NRAS} \) or \( \text{FLT3} \) mutations. High protein expression depicted in red and low protein expression depicted in blue.
Extended Data Fig. 7 | Clonal architecture analysis using single-cell DNA+Protein sequencing of select AML samples. Samples shown have significant differences in community representation between the dominant clone and subclones further discussed in Extended Data Fig. 8. MSK71 (depicted with ***) is highlighted in Fig. 4c–f. Clonotype plot depicts the number of cells identified with a given genotype and ranked by decreasing frequency (top panel). Mean cell counts for each clone are depicted with 95% confidence intervals derived from random resampling analysis. Heat map (middle panel) shows the genotype of each identified protein coding mutation in the given clone with zygosity (wildtype = light pink, heterozygous = orange, homozygous = red). Heat map of the relative protein expression for each cell-surface protein (n = 7) in each identified clone (purple = high expression; green = low expression).
Extended Data Fig. 8 | Neighbourhood analysis of all single-cell DNA+Protein AML samples. a, Divergences in cell-surface protein expression of CD34, CD38, CD11b, and CD45RA determined by presence of signalling effector mutation. Density plots of cells from MSK71 (further detailed in Figure 4c–f and Extended Data Figure 7) of DNMT3A mutant cells (yellow = single-mutant) with co-occurring FLT3 (black), KRAS (orange), or NRAS (light blue) mutations. Concentration of cells with a given immunophenotype depicted by the density of lines. b, UMAP plot of samples (n = 17) analysed by DNA+Protein single-cell sequencing with cells clustered by cell-surface protein expression of 6 antibody targets (CD3, CD11b, CD34, CD38, CD45RA, CD90). Cells from the same sample are denoted with same colour. c, Neighbourhood analysis of all samples from UMAP b with communities of cells identified by neighbourhood analysis in overlaid colours.
Extended Data Fig. 9 | See next page for caption.
Extended Data Fig. 9 | Clone- and gene-specific alterations to cell-surface protein expression and community representation in AML samples.

a, Column normalized heat map of cell-surface protein expression for each community identified in phenoGraph analysis on UMAP from Extended Data Figure 8b. Expression is depicted by colour with blue being low expression and red annotating high expression. b, Community representation changes across all samples (n = 14) in the wild type, the dominant clone, and all subclones. The fraction of each sample within each community is shown with communities depicted by corresponding colour. Samples without communities shown for wild-type cells were found to not have any wild-type cells present in analysis. Changes in immunophenotype due to community representation changes for samples MSK94 (P ≤ 9.95 × 10⁻³) and MSK130 (P ≤ 2.45 × 10⁻³) are highlighted in c. A two proportions z-test for each sample was used to determine statistical significance between dominant clone communities and communities present in subclone ***P < 0.001. c, Cell-surface protein expression of CD11b, CD34, and CD38 between the dominant clone (red) and subclones (black) in an FLT3-ITD mutant sample (MSK130; right panel; n = 2274 total cells) and JAK2 mutant sample (MSK94; left panel; n = 6012 total cells). Each error bar represents a distinct community that is significantly expanded or contracted, (error bar indicates ± standard error of measure, from the mean expression of indicated protein in a given community). A Student’s t-test was used to determine statistical significance *P < 0.1; **P < 0.01; ***P < 0.001.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
- Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection
FASTQ files for single cell DNA libraries were analyzed through the Tapestri Pipeline (version 1.9) using Bluebee's high performance genomics platform. Processed data is then consolidated into a multiple sample VCF file and output as a loom file for subsequent processing. Initial steps for filtering low quality genotypes or cells was performed in Tapestri Insights (version v2.0) and R.

Data analysis
FASTQ files for single cell DNA libraries were analyzed through the Tapestri Pipeline (version 1.9) using Bluebee's high performance genomics platform. All scripts and processed data files for data analysis are available at https://github.com/bowmanr/scDNA_myeloid.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All scripts and processed data files are available at https://github.com/bowmanr/scDNA_myeloid. Raw and processed data are available on dbGAP under accession number phs002049.v1.p1.
Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☑ Life sciences  ☐ Behavioural & social sciences  ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

No sample size calculation was performed. Samples were first selected based on mutation coverage by the Mission Bio Custom amplicon panel, a 109 amplicon panel covering 31 of the most frequently mutated genes in myeloid malignancies. We then further selected samples with variant allele frequencies of all covered mutations (5% VAF for each gene covered on panel), and number of cells collected (5 × 10⁶ cells) per frozen aliquot. In total, we sequenced 146 samples from 123 patients with myeloid malignancies, including clonal hematopoiesis (CH; n=14), myeloproliferative neoplasms (MPN; n=14), and AML (n=91).

Data exclusions

Initial steps for filtering low quality genotypes or cells was performed in Tapestri Insights and R, where the minimum variant quality score was set to 30 with a minimum of 10 reads per variant per cell. We further removed variants present in <50% of cells and removed cells in which <50% of potential variants reported informative genotypes. Samples were excluded if they harbored less than 100 cells with definitive genotype.

Replication

We have performed duplicate single cell DNA sequencing on at least 5 patient samples within our study. Additionally, we have 12 patients where we sequenced serial samples. In each case of duplicated sequencing of the same sample, we observed similar mutations with similar VAF and clonal architecture. In serial samples, we observed similar mutations with minor alterations to clonal architecture and VAF, which would be expected given time between samples.

Randomization

Randomization was not part of our study. All comparisons we performed in our study were either based on diagnosis of patient at time of sample collection or based on the expression of mutant genes/clones in a sample, neither of which could be randomized.

Blinding

Blinding was not relevant to our study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a Involved in the study
☑ Antibodies
☐ Eukaryotic cell lines
☐ Palaeontology
☐ Animals and other organisms
☑ Human research participants
☐ Clinical data

Methods

n/a Involved in the study
☐ ChIP-seq
☑ Flow cytometry
☐ MRI-based neuroimaging

Antibodies

Antibodies used

All antibodies were purchased from Biolegend. These studies used the following antibodies: FITC-CD3 (clone UCHT1) FITC-CD19 (clone HIIB19) and FITC-CD56 (clone HCD56). Human TruStain FcX was also purchased from Biolegend. The DNA+Protein oligo-conjugated antibodies were produced and purchased from Biolegend. The antibodies in the conjugate pool were the following: CD3 (clone SK7), CD11b (clone RCRF44), CD19 (clone HIIB19), CD34 (clone S81), CD38 (clone HIT2), CD45RA (clone HI100), and CD90 (clone SE10).

Validation

All antibodies were purchased from Biolegend. Validation information is available for each antibody at www.biolegend.com

Human research participants

Policy information about studies involving human research participants.

Population characteristics

Patients characteristics are summarized in Extended Table 2 and Figure 1A-B. Samples from patients with myeloid neoplasms or
Population characteristics

Acute myeloid leukemia between 2014 and 2019 were studied. Male and female patients are equally represented (male=63; female=60) with average age at sample collection being 66.4 years (SD=13.0 yrs). Cohort contains a majority of patients with the following diagnosis/disease status at time of samples collection: clonal hematopoiesis (n=14), myeloproliferative neoplasms (n=14), and acute myeloid leukemia (n=91). The majority of AML patients are relapsed/refractory (n=56) with other patients being newly diagnosed (n=25), newly transformed (n=11), therapy related-AML (n=4), or secondary AML (n=18). Samples were selected based on mutation coverage by the Mission Bio Custom amplicon panel, variant allele frequencies of all covered mutations (>5% VAF for each gene covered on panel), and number of cells collected (5–106 cells) per frozen aliquot. Specifically, samples were prioritized if they harbored 1) more than one mutation in epigenetic modifier genes DNMT3A, TET2, ASXL1, or IDH1/2, 2) a NPM1 mutation, 3) mutations in NRAS, KRAS, and/or 4) mutations in FLT3 (either internal tandem duplication (ITD) or tyrosine kinase domain (TKD) mutations).

Recruitment

Patients with myeloid neoplasms or acute myeloid leukemia between 2014 and 2019 were studied. Informed consent was obtained from patients according to protocols approved by the institutional IRBs and in accordance with the Declaration of Helsinki. Diagnosis and disease status was confirmed and assigned according to World Health Organization (WHO) classification criteria. Bone marrow from healthy individuals was obtained with informed consent according to procedures approved by the institutional review boards Memorial Sloan Kettering Cancer Center and Hospital for Special Surgery. Samples processed in this study were either bone marrow mononuclear cells or peripheral blood mononuclear cells.

Ethics oversight

This study was approved by MSKCC Institutional Review Board (protocol #15-017) and Thomas Jefferson University (TJU) Institutional Review Board (protocol# 17D.083).

Note that full information on the approval of the study protocol must also be provided in the manuscript.