A general approach for detecting expressed mutations in AML cells using single cell RNA-sequencing

Allegra A. Petti  
*Washington University School of Medicine in St. Louis*

Stephen R. Williams  
*10x Genomics, Inc.*

Christopher A. Miller  
*Washington University School of Medicine in St. Louis*

Ian T. Fiddes  
*10x Genomics, Inc.*

Sridhar N. Srivatsan  
*Washington University School of Medicine in St. Louis*

See next page for additional authors

Follow this and additional works at: https://digitalcommons.wustl.edu/open_access_pubs

Please let us know how this document benefits you.

Recommended Citation

https://digitalcommons.wustl.edu/open_access_pubs/8073
Virtually all tumors are genetically heterogeneous, containing mutationally-defined subclonal cell populations that often have distinct phenotypes. Single-cell RNA-sequencing has revealed that a variety of tumors are also transcriptionally heterogeneous, but the relationship between expression heterogeneity and subclonal architecture is unclear. Here, we address this question in the context of Acute Myeloid Leukemia (AML) by integrating whole genome sequencing with single-cell RNA-sequencing (using the 10x Genomics Chromium Single Cell 5’ Gene Expression workflow). Applying this approach to five cryopreserved AML samples, we identify hundreds to thousands of cells containing tumor-specific mutations in each case, and use the results to distinguish AML cells (including normal-karyotype AML cells) from normal cells, identify expression signatures associated with subclonal mutations, and find cell surface markers that could be used to purify subclones for further study. This integrative approach for connecting genotype to phenotype is broadly applicable to any sample that is phenotypically and genetically heterogeneous.
C

one转基因到表型在单个细胞水平是广泛被

视作一个中央挑战在分析和解释scRNA-seq数据，因为这意味着细胞分化表型和子克隆会对下游问题产生影响。

Meanwhile, a growing body of work has demonstrated that tumors are also transcriptionally heterogeneous, but it has been challenging to relate this epigenetic heterogeneity to genetic heterogeneity in individual tumors14,16. We sought to address this challenge by detecting cells that express somatic single nucleotide variants (SNVs) in scRNA-seq data.

Detecting genetic variants in scRNA-seq reads is difficult due to the low transcript abundance, allelic dropout, and incomplete transcript coverage inherent to this platform. Despite these challenges, previous studies of intratumoral heterogeneity have demonstrated that single-cell copy number alterations (CNAs) can be robustly detected in full-length cDNAs, commonly generated using the Fluidigm C1/SMART-seq platform, in dozens3,4,5 to hundreds6,8,11,14,17,18 of cells per tumor, and specialized tools have been developed for this purpose12,19. Others have built upon plate-based scRNA-seq technologies to detect specific mutations with variable sensitivity20,21.

The ability to detect CNAs in single cells has advanced the study of cancers where structural alterations and/or aneuploidy are common3,4,13,14,18. However, CNAs are rare in some tumor types, such as AML12,23. Moreover, CNAs rarely capture the complete subclonal complexity of any tumor, and are often subclonal progression-associated events24. The ability to detect multiple, arbitrary SNVs in scRNA-seq reads is an ideal attribute for any generally-applicable approach to the study of intratumoral heterogeneity. Although previous studies have established that some SNVs can be identified from full-length cDNAs, low numbers of identified mutant cells made downstream analyses difficult14,16.

In working with the 10x Genomics Chromium Single Cell 3’ (v2) and 5’ (v1) Gene Expression workflows, we observed sequence coverage far from the 3’ and 5’ ends of genes (respectively). This was unexpected, given the end-bias of the Chromium library design, and raised the possibility that the resulting scRNA-seq data could be used for variant detection. Because this platform can sample up to 10,000 cells per library, we hypothesized that rare or complex mutations might be detectable.

Results
eWGS and bulk RNA-sequencing. Four cases of de novo AML and one of secondary AML were selected for study (clinical details in https://github.com/genome/scrna_mutations). eWGS was used in conjunction with well-established variant detection pipelines to generate a set of high-confidence mutation calls for each case, and bulk RNA-sequencing was used to determine which mutations were expressed in each tumor sample (Methods)25. eWGS (Fig. 1a) revealed that these cases were genetically representative of AML, containing on average 26 mutations within coding regions, with many in well-established driver genes (e.g. DNMT3A, FLT3, NPM1, TP53, NRAS, IDH1, CEBPA, etc.). To define the clonal architecture of each tumor, the SciClone algorithm26 was used to cluster mutations and infer subclones. At least one subclone was identified in every case (Table 1, Supplementary Data 1)23. Bulk RNA-sequencing showed that on average, fewer than half of the mutations detected by eWGS were expressed (Table 1).

Single-cell transcript coverage and representation. We first compared genome-wide transcript coverage obtained from the 5’ (v1) and 3’ (v2) 10x Genomics Chromium Single Cell Gene Expression workflows. For one case, UPN 508084, we generated two scRNA-seq libraries with each workflow, and sequenced them to high depth, targeting 200,000 reads/cell. Transcriptome-wide coverage for each data set was assessed using 20,090 genes, each having one annotated isoform (Methods). Both workflows yielded consistent low-level coverage at least 10 kbp from the 5’ and 3’ ends of the average transcript (Fig. 1b). However, the 5’-3’ kit yielded slightly higher transcript-wide coverage in distant regions of transcripts. For the average gene assayed using that kit, at least 2.5% of the unique sequenced transcripts mapped to any given base up to 10 kbp away from the 5’ transcription start site of the gene. Coverage metrics for 200 cancer-relevant genes are summarized in Supplementary Data 2 and provided at nucleotide resolution at https://github.com/genome/scrna_mutations. Subsequent sequencing and analyses were performed using only the 5’ workflow application.

We then asked whether bulk and single-cell RNA-seq data capture the same transcript structure for the mutated genes in this study. Using one canonical isoform for each gene, we compared coverage in the single-cell data (unique barcode/UMI pairs at each position) to that in the bulk RNA-seq data (quantified using bamCoverage and 1 bp bins), and visualized it using the UCSC Genome Browser (Methods). The results demonstrate that, for each gene studied, bulk- and single-cell data identified the same set of transcripts (Fig. 1c). Coverage plots for all mutated genes in this study are provided at https://github.com/genome/scrna_mutations.

Mutation identification in single cells. We next sought to identify cells containing any of the somatic variants discovered using eWGS. For each cell and each variant position in the eWGS data, unique wild-type, and mutant reads were counted using cb_sniffer, a tool that extends the PySam library to do barcode-aware pileup analysis (see Methods: https://github.com/genome/cb_sniffer). In most high-throughput scRNA-seq datasets, the median gene is represented in the median cell by one transcript read. Consistent with this, most SNV locations were covered by a single read in most cells (although SNVs in several highly expressed, high-coverage genes (e.g. U2AF1, NPM1, SRSF2, and NRAS) were more likely to have multiple reads per cell) (Supplementary Fig. 1a). For a heterozygous mutation, therefore, there is a 50% chance that the observed transcript is mutant, and a 50% chance that it is wild-type, leading to the phenomenon known as allelic dropout. This has two main consequences: first, it is impossible to conclude that a cell is wild-type; secondly, the sensitivity of mutation detection is reduced by a factor of two. Therefore, barring sequencing errors, one can in principle classify a cell’s genotype as “mutant” if it contains one or more mutant transcripts, and “unknown” if it does not. We measured the
frequency of false positives originating from sequencing errors by examining the positions of known somatic mutations in samples that did not harbor those mutations. The false-positive rate (the rate at which wild-type UMIs are called mutant in the control samples) was site-specific, and had a maximum rate of only 0.39% (Supplementary Data 3). We also searched for these variants in 8057 bone marrow cells from four healthy donors, and found no false positives.

We therefore labeled a cell “mutant” if it contained at least one variant-containing read, and “unknown” if only wild-type reads or no reads were detected. We found an average of 49 mutant cells per variant (range: 1–3944), and 3732 mutant cells (22% of the total cells) per sample, but this varied widely among samples (range: 396–8200, or 1.8–52%), depending on the mutations present in each (Table 1). Most mutant cells contained one detected mutation, with one read mapping to the variant position (Supplementary Fig. 1). Founding clone mutations, subclonal mutations, and putative driver mutations were detectable in each case, and these included SNVs, insertions and deletions (indels, including FLT3-ITD and NPMc), and one gene fusion (NUP98-NSD1) (Table 1, Supplementary Data 1). Although the vast majority of mutant cells contained only one mutation (88–98%, depending on the sample), a small fraction of cells in each case contained multiple mutations, particularly when a founding clone mutation was readily detectable (Table 2). Specifically, two mutations were found in 1.6–12% of the mutant cells in a sample; three mutations were found in 0.21–0.29% of mutant cells; and four mutations were found in one cell (0.012%) in sample 721214. The observed mutation combinations were consistent with the known subclonal architecture (although the mutation data was generally not dense enough for accurate de novo subclonal inference). For example, case 548327 contained an NPM1W288fs mutation in the founding clone, and several hundred cells contained both this mutation and one subclonal mutation. Case 721214 is composed of three subclones sequentially nested within the founding clone. One cell was found to have one mutation from each (sub)clone.

Mutation detection in single cells was compared to that in bulk RNA-seq and eWGS data using a read-based metric, “single-cell Variant Allele Frequency” (scVAF), which enabled us to compare VAFs across data types, and two cell-based metrics, “Mutant Cell Fraction” (MCF) and “Mutant Cell Detection Rate” (MCDR), which allowed us to measure and compare the sensitivity of mutant cell identification (Methods). In terms of mutant reads,
<table>
<thead>
<tr>
<th>Sample</th>
<th>508084</th>
<th>548327</th>
<th>721214</th>
<th>782328</th>
<th>809653</th>
<th>Mean [SD]</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. cells</td>
<td>14,964</td>
<td>11,620</td>
<td>20,474</td>
<td>21,731</td>
<td>21,038</td>
<td>17,965 [3979]</td>
</tr>
<tr>
<td>Reads/cell</td>
<td>192,427</td>
<td>346,965</td>
<td>176,035</td>
<td>214,284</td>
<td>189,751</td>
<td>223,892 [62,745]</td>
</tr>
<tr>
<td>Reads mapped</td>
<td>84.9</td>
<td>87.2</td>
<td>92.2</td>
<td>79.8</td>
<td>90</td>
<td>87 [4.29]</td>
</tr>
<tr>
<td>Read confidence</td>
<td>64.2</td>
<td>63.7</td>
<td>73</td>
<td>61.9</td>
<td>68.8</td>
<td>66 [4.04]</td>
</tr>
<tr>
<td>genome (%)</td>
<td>84.9</td>
<td>87.2</td>
<td>92.2</td>
<td>79.8</td>
<td>90</td>
<td>87 [4.29]</td>
</tr>
<tr>
<td>Median genes</td>
<td>2405</td>
<td>1383</td>
<td>2260</td>
<td>1376</td>
<td>1829</td>
<td>1851 [428]</td>
</tr>
<tr>
<td>Median genes (%)</td>
<td>80%</td>
<td>7</td>
<td>17</td>
<td>12</td>
<td>8</td>
<td>10.4</td>
</tr>
<tr>
<td>Median genes (%)</td>
<td>80%</td>
<td>7</td>
<td>17</td>
<td>12</td>
<td>8</td>
<td>10.4</td>
</tr>
<tr>
<td>Median genes (%)</td>
<td>80%</td>
<td>7</td>
<td>17</td>
<td>12</td>
<td>8</td>
<td>10.4</td>
</tr>
<tr>
<td>Total genes</td>
<td>22,645</td>
<td>22,503</td>
<td>23,376</td>
<td>25,389</td>
<td>23,102</td>
<td>23,403 [1041]</td>
</tr>
<tr>
<td>WGS variants</td>
<td>19</td>
<td>13</td>
<td>41</td>
<td>31</td>
<td>28</td>
<td>26.4</td>
</tr>
<tr>
<td>Expressed WGS</td>
<td>10</td>
<td>5</td>
<td>18</td>
<td>7</td>
<td>8</td>
<td>9.6</td>
</tr>
<tr>
<td>variants (bulk)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percentage</td>
<td>8%</td>
<td>7%</td>
<td>17%</td>
<td>12%</td>
<td>8%</td>
<td>10.4</td>
</tr>
<tr>
<td>Percentage</td>
<td>8%</td>
<td>7%</td>
<td>17%</td>
<td>12%</td>
<td>8%</td>
<td>10.4</td>
</tr>
<tr>
<td>Percentage</td>
<td>8%</td>
<td>7%</td>
<td>17%</td>
<td>12%</td>
<td>8%</td>
<td>10.4</td>
</tr>
<tr>
<td>Total mutant</td>
<td>669</td>
<td>6042</td>
<td>8200</td>
<td>3354</td>
<td>386</td>
<td>3732</td>
</tr>
<tr>
<td>Total mutant</td>
<td>669</td>
<td>6042</td>
<td>8200</td>
<td>3354</td>
<td>386</td>
<td>3732</td>
</tr>
<tr>
<td>Mutant cells per</td>
<td>13-453</td>
<td>1-3012</td>
<td>1-3944</td>
<td>1-2619</td>
<td>1-207</td>
<td>3-4-2047</td>
</tr>
<tr>
<td>per cell</td>
<td>13-453</td>
<td>1-3012</td>
<td>1-3944</td>
<td>1-2619</td>
<td>1-207</td>
<td>3-4-2047</td>
</tr>
<tr>
<td>Key WGS variants</td>
<td>32</td>
<td>48</td>
<td>30</td>
<td>111</td>
<td>21.5</td>
<td>48.5</td>
</tr>
<tr>
<td>(median)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IKBKBV616M</td>
<td>(150)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLT3-ITD</td>
<td>(707)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NUP98-NDS1(1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Additional</td>
<td>RNFL10</td>
<td>(103)</td>
<td></td>
<td></td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>variants</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(number of cells</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>with coverage)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
the sensitivity of mutation detection was comparable in single cell and bulk RNA-seq data: on average, a slightly higher fraction of known mutations was detected in the scRNA-seq data, but not necessarily in a large number of cells (Table 1). We then examined the relationship between bulk VAF (either eWGS or bulk RNA-seq) and single-cell VAF (scVAF) and detection sensitivity (MCF); for expressed mutations that were identifiable in bulk and single-cell RNA-seq data, single-cell MCFs and scVAFs were more highly correlated with eWGS VAFs ($r = 0.69$ and $0.68$, respectively) than were bulk RNA VAFs ($r = 0.52$; Table 1). scVAFs were positively correlated with bulk RNA-seq VAFs ($r = 0.34$; Fig. 1d, Supplementary Fig. 1c). Although mutation-detection in scRNA-seq is sensitive from a read-based perspective, sensitivity from a cell-based perspective is very low and mutation-specific; we identified only a small fraction of the cells that would be expected to contain mutations based on eWGS VAFs (MCDR, Fig. 1d). This fraction depends on multiple variables, including bulk RNA VAF, distance from the 5’ end of the transcript, and single-cell gene expression (Fig. 1d, Supplementary Fig. 1c, d).

The ability to detect mutations in scRNA-seq data therefore depends on a number of variables, including VAF, expression level of the mutated gene, position of the mutation in the transcript, sequencing depth, fraction of tumor cells in the sample, and number of cells sequenced. The probability of finding at least one cell containing a particular heterozygous mutation $m$ is approximately:

$$P(m) = na(f[1 - (1 - ct)^e] + (1 - f)e) \approx na(1 - f)e$$

Where $f$ is the twice the variant allele frequency of the mutation in the eWGS data, $r$ is the relative expression level of the gene (e.g. in counts per million), $r$ is the average number of UMIs per mutant cell, $c$ is the fraction of UMIs that have coverage at the mutant position, $e$ is the site-specific false-positive rate (frequency with which a wild-type cell is called mutant), $a$ is the fraction of cells in the sample that are tumor cells, and $n$ is the total number of cells sequenced.

Using SNVs to distinguish between tumor and normal cells. Single-cell CNA detection is often used to identify tumor cells in samples that contain a mixture of tumor and normal cells, but sensitivity is limited by the fact that CNAs are frequently subclonal, even in the (non-AML) tumors that contain them. Therefore, we investigated the utility of single-cell SNV detection for this purpose. A straightforward approach would involve selecting only those cells that contain a mutation; we detected an average of 3732 mutant cells per sample (Table 1). Despite the wide range (396–8200), this is substantially more than the total number of cells/sample analyzed in previous single-cell mutation-detection studies. However, we retained the additional cells in each sample (which contained valuable expression information), and instead used single-cell SNVs as markers for tumor vs. wild-type cell clusters.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Total mutant cells</th>
<th>1 mutation (%)</th>
<th>2 mutations (%)</th>
<th>3 mutations (%)</th>
<th>4 mutations (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>508084</td>
<td>669</td>
<td>658 (98)</td>
<td>11 (1.6)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>548327</td>
<td>6042</td>
<td>5290 (88)</td>
<td>734 (12)</td>
<td>18 (0.29)</td>
<td>1 (0.02)</td>
</tr>
<tr>
<td>721214</td>
<td>8200</td>
<td>7702 (94)</td>
<td>477 (5.8)</td>
<td>20 (0.24)</td>
<td>1 (0.02)</td>
</tr>
<tr>
<td>782328</td>
<td>3354</td>
<td>3176 (95)</td>
<td>171 (5.1)</td>
<td>7 (0.21)</td>
<td>0</td>
</tr>
<tr>
<td>809653</td>
<td>396</td>
<td>386 (97)</td>
<td>10 (2.5)</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

We first used principal component analysis to summarize the expression heterogeneity in each case (Methods) to better understand the composition of each sample. As expected, this revealed complex relationships among clusters (such as partially overlapping expression signatures), and multiple sources of heterogeneity in all samples, including variable expression of known hematopoietic cell-type markers (e.g. CD3δ (T-cells), CD79α, or CD19 (B-cells), and HBA1 (erythrocytes)), cell cycle genes (e.g. TUBA1B, TOP2A), markers of myeloid lineage (e.g. AZU1, ELANE, MPO, PRTN3), mitochondrial genes, and ribosomal genes (Fig. 2a, b; Supplementary Fig. 2–5, Supplementary Data 4). This indicated that the distribution of cell types is a major source of expression heterogeneity, and varies among samples, as expected.

To investigate sample composition in a more unsupervised manner, we identified the nearest hematopoietic lineage of each cell by matching each cell’s expression profile to the most similar lineage-specific expression profile in the DMAP database. (Methods, Figs. 3c and 4). The inferred sample composition varied widely among subjects, particularly with respect to the fraction of lineage-defined cells (e.g. cells resembling myelomonocytic cells, T-cells, B-cells, and erythrocytes). All five samples contained clusters of immature cells, including cells resembling hematopoietic stem cells (HSCs), common myeloid progenitors (CMPs), and megakaryocyte-erythroid progenitors (MEPs), which could represent either immature non-malignant cells or AML cells.

To clarify the identity of these clusters, we combined single-cell mutation detection with expression-based clustering and lineage inference. Using the bone marrow sample from 809653 (which contained many non-AML cells, based on morphology and flow cytometry) we overlaid mutation data on the t-SNE projections by highlighting mutant cells (Fig. 3e–g). A highly expressed germline SNP in the BAG1 gene served as a positive control, marking SNP-containing cells in all expression clusters (Fig. 3h).

By scRNA-seq, we detected cells expressing mutations in 8 genes, including TP53, NTRAS, and CEBPA (Table 1, Supplementary Data 1). Several clusters were significantly enriched ($p \leq 0.05$, one-sided Fisher exact test) for mutant cells; other cells in these clusters presumably contained undetected mutations in these genes (Fig. 3b–g). Two of these clusters were composed of cells that had stem/progenitor expression signatures (HSCs and MEPs). The other two were composed of cells expressing erythrocyte or monocyte markers; in terms of gene expression, these clusters are distinct from normal cell clusters, but they could not have been labeled as AML-derived using expression data alone.

This was the only case with multiple CNAs, allowing us to benchmark SNV-based cell classification against the better-established CNA-based methods. CONICSmatt was used to identify cells containing the CNAs discovered by eWGS, and high concordance was observed with SNV/expression-based classification of AML cells: 95.5% of cells classified as AML by copy number were also classified as AML by SNV and expression signature (Fig. 2c). Conversely, 94.9% of cells classified as AML by SNVs and expression were confirmed by CNA analysis.
**Fig. 2** Clustering, overview of expression heterogeneity, and copy number analysis in 809653. (a) t-SNE projection of scRNA-seq data, with cells colored according to graph-based cluster assignment; putative AML clusters (based on later analyses) circled. (b) Hierarchical clustering of the most heavily weighted genes in each principal component, averaged within graph-based clusters. Each column represents a cluster from panel a. (c) CNV analysis: blue, cells with detected CNVs; gray, no detected CNVs.
NRAS integrative genomic approach validates the concept that AML supported by the extent of differentiation of each tumor. Our conclusions were inference with single-cell mutation identi-
Evaluating tumor differentiation state
monocytes in 809653 (Fig. 3c), and monocytes and NK-T cells in natures consistent with differentiated cells: erythrocytes and considerable fraction of the mutant cells had expression sig-
782328 (Fig. 4d). Likewise, case 548327 contained mutant cells that some AML cells display lineage in that co-clustered with wild-type B- and T-cells, again suggesting
Expression signatures of mutation-containing cells. In the approach described above, we treated mutation-containing cells as markers for entire clusters of putative AML cells. However, the ability to map mutations in many cells (3732 mutant cells/sample on average) facilitates more conservative, direct analyses of intratumoral expression heterogeneity, using only the cells that express a confirmed somatic mutation. For each sample, therefore, we analyzed the mutationally defined AML cells separately (Methods, Supplementary Figs. 6 and 7). As expected, all samples showed intercellular heterogeneity in the expression of cell cycle genes (as expected) and genes that function in the immune system, especially the MHC Class II genes and/or CD74. All but one case (782328) showed intercellular variability in expression of TP53-interacting genes. Three cases (508084, 548327, and 721214) showed intercellular heterogeneity in genes that interact with the vascular cell adhesion gene VCAM1, and three (721214, 782328, and 809653) showed heterogeneous expression of myeloid differentiation genes. There were also case-specific signatures, such as “response to reactive oxygen species” in 721214. As discussed further below, a GATA2 expression signature is evident in cells expressing this mutation. Thus, the reduced, mutant- only data set is sufficient to capture much of the expression heterogeneity observed in the total sample.
Mutation-associated expression signatures. We next investigated the extent to which mutational heterogeneity was associated
Fig. 4 Single-cell mutation detection and interpretation in additional cases ordered by the differentiation signature of AML cells. a 721214, top to bottom: clonality inferred from eWGS; cells colored according to closest inferred lineage (RBC = red blood cell, HSC = hematopoietic stem cell, CMP = common myeloid progenitor); cells colored according to cell cycle phase; cells colored according to single-cell genotype at the indicated site: blue, at least one mutant read detected; yellow, wild-type reads only; gray, no coverage. b 548327, putative AML cells circled. c 508084. d 782328
with transcriptional heterogeneity in each case. A subclonal mutation that drives an expression signature should be restricted in expression space. In contrast, a founding or subclonal mutation not associated with an expression signature should be present throughout expression space. Furthermore, this should not depend on restricted expression of the mutant gene. To this end, we highlighted mutant cells on the t-SNE projection of each sample, and identified mutations that are nonuniformly distributed, even after controlling for that gene's expression (Figs. 3e–g and 4, Supplementary Fig. 8). We performed two versions of this analysis: a “whole-sample” analysis using all cells, and a “mutant-cell” analysis using only mutation-containing cells. The whole-sample analysis capitalized on the high throughput of this platform by incorporating expression information from all ~20,000 cells per sample, thereby improving our ability to discern distinct expression signatures. On the other hand, the limitations of genotype assignment imposed by low coverage and allelic dropout required that we compare a relatively small number of mutant cells to a much larger number of cells of unknown genotype (representing a mixture of mutant and wild-type cells). Therefore, instead of performing a straightforward comparison between mutant and wild-type cells, we looked for evidence that the mutant cells were nonuniformly distributed among the “unknown” cells: when defining expression signatures, as described further below, we searched for genes whose expression was correlated with the density of mutant cells.

The results of the whole-sample analysis indicated that the relationship between expression heterogeneity and mutational heterogeneity is case- and mutation-dependent. Two cases, 721214 and 508084, contained subclonal mutations with nonuniform distributions (Fig. 4a, c). Based on eWGS, 721214 contained subclonal mutations with nonuniform distributions, even after controlling for that gene's expression (Figs. 3e–g and 4, Supplementary Fig. 8). We performed two versions of this analysis: a “whole-sample” analysis using all cells, and a “mutant-cell” analysis using only mutation-containing cells. The whole-sample analysis capitalized on the high throughput of this platform by incorporating expression information from all ~20,000 cells per sample, thereby improving our ability to discern distinct expression signatures. On the other hand, the limitations of genotype assignment imposed by low coverage and allelic dropout required that we compare a relatively small number of mutant cells to a much larger number of cells of unknown genotype (representing a mixture of mutant and wild-type cells). Therefore, instead of performing a straightforward comparison between mutant and wild-type cells, we looked for evidence that the mutant cells were nonuniformly distributed among the “unknown” cells: when defining expression signatures, as described further below, we searched for genes whose expression was correlated with the density of mutant cells.

The results of the whole-sample analysis indicated that the relationship between expression heterogeneity and mutational heterogeneity is case- and mutation-dependent. Two cases, 721214 and 508084, contained subclonal mutations with nonuniform distributions (Fig. 4a, c). Based on eWGS, 721214 contained a subclone defined by GATA2R361C. In the scRNA-seq data, cells expressing GATA2R361C were largely restricted to the same space on one side of the t-SNE projection, suggesting that AML cells containing this mutation have a unique expression signature (Fig. 4a). Two cases (809653 (Fig. 3f–g) and 782328 (Fig. 4d)) exhibited complex mutation-associated expression profiles, and a third, 548327 (Fig. 4b), showed expression heterogeneity in the absence of discernable genetic heterogeneity. The GATA2R361C gradient in 721214 was of particular interest, because GATA2 encodes a transcription factor that is a key regulator of hematopoiesis, and is recurrently mutated in AML.23,30. We therefore sought to characterize the associated expression signature. As noted above, scRNA-seq data allows us to distinguish between mutant cells and cells of unknown genotype; we cannot conclusively label a cell as “wild-type.” To address this limitation while incorporating expression information from cells of unknown genotype, we employed a regression-based method that identifies genes whose expression is correlated with the density of mutant cells (Methods), for a given mutation or set of subclonal mutations. This approach makes use of expression clusters to smooth the expression and density data, but does not depend on the exact clustering, and does not require us to identify cluster-specific gene expression. It also permitted us to control for potential covariates, such as the expression of GATA2, which was slightly correlated with the mutation density gradient. We applied this method to the full data set in two ways: first, by searching for genes whose expression was correlated with any mutation in the GATA2R361C subclone (Fig. 5a), and second, for genes associated with GATA2R361C per se. Each analysis yielded several hundred genes whose expression was positively correlated with GATA2R361C density (FDR-adjusted p-value for the regression coefficient <0.05; Fig. 5b, c, Supplementary Data 5a, b). Clusters with a higher density of expressed GATA2R361C subclonal mutations exhibited higher expression of genes involved in immune response, apoptosis, and leukocyte adhesion.29 (Fig. 5c, Supplementary Table 1). Notably, the gene whose expression was most strikingly correlated with this subclone is VIM, which encodes a type III intermediate filament and is an established target of the GATA2/SPI1 (PU.1) transcriptional circuit (Fig. 5d).31,32

As described above, we also analyzed this sample using only the cells of known (i.e. mutant) genotype, and again found that GATA2R361C subclonal mutations were nonuniformly distributed in expression space. We compared mutant-rich clusters (mutation fraction >10%) to the remaining clusters using a Wilcoxon rank sum test for differential expression.33 (Fig. 5e–h). Consistent with the above results, subclonal mutations were associated with higher expression of VIM, CRIP1, AHI1AK, CD74, and other genes associated with immune response, apoptosis, and cell adhesion (Supplementary Data 5c).

Many of the GATA2R361C subclone-associated genes are highly correlated with each other (and with VIM), in the TCGA AML gene expression data. To quantify the overlap between TCGA VIM-associated genes and GATA2R361C subclone-associated genes, we identified 2191 genes that are highly correlated with VIM in TCGA (q < 0.001, Pearson correlation, Benjamini-Hochberg correction), and used a hypergeometric test to compare them to the GATA2R361C subclone-associated genes (Fig. 5h). The intersection of these gene sets was statistically significant (p = 3.5 × 10^-95, hypergeometric test), suggesting the existence of a “regulon” whose expression is influenced by one or more mutations in the GATA2R361C subclone. To further characterize this regulon, we examined the functional enrichment of the 198 genes in the intersection (Supplementary Table 2, Supplementary Data 5d), and found that they are enriched for Gene Ontology (GO) terms related to immune response (in particular, the Fc-gamma receptor pathway), cytoskeletal organization, and focal adhesion, and for genes that interact with WAS. WAS, which encodes Wiscott-Aldrich Syndrome Protein, transduces signals from the cell surface to the actin cytoskeleton in response to infection, and is required for a variety of immunological cell functions. WAS mutations are associated with a broad spectrum of clinical manifestations, including immunodeficiencies and hematologic malignancies. Like VIM, WAS may also be regulated by PU.1.36 Together with our data, this suggests that a subset of PU.1 target genes coordinates immune function with cytoskeletal reorganization in hematopoietic cells, and that at least one of the mutations in the GATA2R361C subclone influences the expression of these genes. Because GATA2 is a transcription factor that negatively regulates PU.1, it is likely that the GATA2R361C mutation itself is at least partly responsible for the observed transcriptional effects in this sample. Furthermore, GATA2 has well-documented roles in both immune function and hematological malignancies: autosomal dominant mutations in GATA2 can also lead to immunological disorders and hematologic malignancies.

The success of this method depends on a number of factors, including steepness of the expression gradient and number of mutant cells (the more subtle the expression signature, the more mutant cells required). Moreover, irrelevant or hidden variables can affect the distribution of mutant cells in expression space, such as expression level of the mutated gene, cell cycle phase, ribosomal transcript content, mitochondrial transcript content, or other variables for which we could not account. We therefore used an independent experimental approach to test for the GATA2R361C-associated expression gradient, in which we compared the frequency of GATA2R361C in genomic DNA from cells drawn from each extreme of the expression gradient. First, scRNA-seq was used to identify cell-surface markers whose expression was correlated with the GATA2R361C mutation.
gradient. This analysis yielded CD99, a well-described cell-surface marker associated with AML and MDS cells (Fig. 6a). Then, peripheral blood and bone marrow samples from this patient were stained with CD99 FITC-conjugated antibody, and flow cytometry was used to isolate cells with high or low CD99 expression (top or bottom 15%, Fig. 6b, c; Methods). Genomic DNA was prepared from each population, and targeted sequencing was used to measure the frequency of GATA2 R361 mutations (as well as a control DNMT3A R882 mutation, which is found in all AML cells in this sample) in each cell population. This demonstrated that GATA2 R361C is significantly more abundant (p = 0.0081 (marrow), p = 0.0432 (peripheral blood), Fisher Exact test) in the genomes of the CD99hi cells (Fig. 6d). The control mutation, DNMT3A R882, was not significantly enriched in the CD99hi cells, because it is present in all of the AML cells in this sample (it was the initiating event). GATA2 R361C abundance varies more dramatically in the scRNA-seq data, possibly due to allele-specific expression of the mutant allele, a documented phenomenon in GATA2-mutated AML. These results support the conclusion that GATA2 R361C is associated with a distinct gene expression profile, and shows that SNV detection in scRNA-seq data can be used to identify mutation-associated expression signatures. Moreover, the ability to identify cell surface markers for the purification and analysis of subclones is an important application of scRNA-seq data that should have broad applications.

Fig. 5 GATA2 R361C Subclonal expression signature. a t-SNE projection showing mutation-expressing cells in blue (GATA2 R361C) and pink (TIMM17B L122fs). b Cells colored according to graph-based cluster assignment. c Heatmap of top 50 mutation-dependent genes, with bar graph showing mutant cell fraction in each AML cluster (labeled to the right of the heatmap). d Cells colored according to VIM expression (left), and scatterplot showing average VIM expression in each cluster as a function of the subclonal mutation fraction of each cluster (right). e t-SNE plot constructed from mutant cells, which are colored according to the mutation they contain: GATA2 R361C, yellow; DNMT3A R882H, pink; FLT3-ITD, green; FLT3 F612L, purple; NPM1 W288Fs; other somatic mutation(s), gray. f Mutant cells colored according to graph-based cluster. g Heatmap of top 25 subclonal mutation-dependent genes, with bar graph showing mutant cell fraction in each cluster (labeled to the right of the heatmap). Genes that are highly correlated with VIM in TCGA are indicated with blue dots. h Venn diagram indicating gene sets used to identify the VIM regulome.
In case 508084, a mutation in RNF10, a putative transcription factor of unknown function, was also restricted to a subset of expression clusters (Fig. 7a, b). Compared to clusters with few or no mutant cells, mutant-rich clusters displayed a clear expression signature marked by high expression of genes involved in immune-related cell adhesion. This included genes involved in MHC Class II receptor activity and T cell aggregation, as well as genes that interact with MCM2 (a regulator of TP53) and NPM (which is frequently mutated in AML) (Fig. 7c, Supplementary Table 3, Supplementary Data 5e).

The remaining three cases showed more complex relationships between expression and mutations. Based on the eWGS results, 809653 contained TP53 E286G and CEBPA R142fs in the founding clone, and a subclone defined by NRAS G12D and NF1 I679fs. In the single-cell data, however, the distribution of CEBPA R142fs was markedly nonuniform, suggesting that CEBPA R142fs may be in a subclone (Fig. 7d, e); the clonal architecture of this case may be more complicated than can be discerned from a single eWGS sample. Notably, cells expressing CEBPA R142fs were restricted to one AML cluster that differed from the other AML clusters with respect to differentiation state and cell cycle status: compared to CEBPA wild-type AML clusters, the CEBPA-mutant cluster was enriched for cells in S-phase, and cells with progenitor-like expression signatures. Differential expression analysis of the CEBPA-mutant cluster showed that it overexpressed genes associated with a variety of biological processes, most notably ribosome biogenesis, which probably reflects the increased protein synthesis requirements of rapidly proliferating cells (Fig. 7f, Supplementary Table 4, Supplementary Data 5f). The cluster also expressed a variety of key transcription factors involved in myeloid differentiation, particularly targets of Myc, consistent with the observed perturbation of cellular differentiation in this cluster.

In 782328, NRAS G12D and NRAS G12S also display subtle expression signatures (Fig. 4d). They are predominantly localized to cells in the S, G2, and M phases of the cell cycle, suggesting a
The ability to link genetic and transcriptomic information in single cells is important for the study of heterogeneous cell populations. By combining eWGS and scRNA-seq data from a high-throughput platform, we can distinguish between tumor and non-tumor cells, identify tumor cells displaying lineage infidelity, evaluate the differentiation state of individual tumor samples, derive mutation-associated expression signatures, study transcriptional heterogeneity within confirmed tumor cells, and identify cell-surface markers that can be used to isolate specific cells for downstream study. Further, the approach described here should be applicable–without additional modifications or customization–to virtually any tumor type.

Previous studies have demonstrated that CNAs and specific genetic variants, such as the BCR-ABL fusion, can be identified with high sensitivity in full-length transcripts from dozens to hundreds of single cells using plate-based techniques such as the Fluidigm C1/Smartseq platform, and that SNVs can also be identified, albeit with lower sensitivity, using that data. Because CNAs rarely reflect the complete clonal architecture of a tumor (and are rare in most AML samples), we were interested in finding a way to identify SNVs in single cells. We noticed that the 10x Genomics Chromium Single Cell 3’ and 5’ Gene Expression workflows yield unexpectedly high transcript coverage far from the 3’ and 5’ ends of transcripts. Although this distal coverage is sparse, it is sufficient for low-sensitivity variant detection in single cells: SNVs were detectable in 22.7% of the cells in our samples, on average. Coupled with the high throughput of the platform, this sensitivity enables the detection of SNVs in hundreds to thousands of cells per sample. Although these cells can be studied in isolation, we analyzed them in the context of the entire sample, thereby leveraging the expression information provided by the additional, non-genotyped cells.

A common application of variant detection in scRNA-seq data is to distinguish tumor from normal cells in heterogeneous samples. However, because malignant cells can have expression profiles that mimic more highly differentiated normal cells, gene expression data alone is not sufficient to identify bona fide AML cells. Moreover, AML cells sometimes display lineage infidelity, where some AML cells display the characteristics of differentiated cell types from other lineages, such as T-cells. These AML cells, which would have been missed if classification had been performed using expression signatures alone, can be identified when mutation information is also considered.

Transcriptional heterogeneity in AML samples clearly arises from multiple sources, including the differentiation states of
normal and tumor cells, cell cycle states, mutations that are present in subsets of cells (i.e. subclones), or non-genetic heterogeneity that arises as a consequence of stochastic gene expression or other perturbations. Incorporating variant detection into scRNA-seq analysis helps to distinguish among these sources by facilitating the distinction between tumor and normal cells, and by revealing correlations between mutational and transcriptional heterogeneity. Equally, importantly, it suggests that genetic heterogeneity plays a limited role in establishing transcriptional heterogeneity: we routinely observed expression heterogeneity in the absence of detectable genetic heterogeneity. In some cases, this may be due in part to the limited sensitivity of mutation detection. In others, they may exemplify the well-established phenomenon whereby stochastic gene expression gives rise to phenotypic heterogeneity in clonal populations of cells. Non-genetic transcriptional heterogeneity can influence phenotype (such as drug sensitivity, growth rate, and cell fate) and persist across generations and might therefore serve as a substrate for natural selection. This underscores the idea that a combination of genetic and non-genetic sources of heterogeneity may help to govern tumor biology and evolution. As additional scRNA-seq studies of primary tumor samples are undertaken by many groups, the relative contributions of these sources of heterogeneity for each tumor type should become more clear.

The detection of cells with expressed mutations in scRNA-seq data is subject to several limitations. Dropout (including transcript dropout and allelic dropout) occurs with most scRNA-seq platforms. As a result, it is impossible to determine whether a cell is truly wild-type for a given mutation. Additionally, dropout reduces the sensitivity of mutation detection by a factor of two. Partial transcript coverage is specific to end-biased platforms such as the Chromium platform, and also limits the sensitivity of variant detection. Moreover, coverage drops non-linearly across the length of the transcript, so some variants are much more easily detectable than others. The utility of this approach therefore depends on the specific mutational composition of the sample in question, and will likely perform better for other tumor types, almost all of which have higher mutation burdens than AML.

A number of other approaches to identify expressed mutations in single-cell RNA-sequencing data have been described. Each method has different strengths and weaknesses that should influence the choice of platform for a specific experimental question. Key variables include library insert size, end-bias, and complexity, sequencing depth and read length, dropout rate, and throughput. Furthermore, technologies that enable simultaneous DNA and RNA sequencing of single cells, such as G&T-seq, may become very powerful with increased throughput. The rapid pace of technological advancement in this area will likely increase the power of scRNA-seq to identify and distinguish among different sources of transcriptional heterogeneity in primary tumor samples.

Methods

Ethical approval and consent. Samples were obtained as part of a study that was approved by the Human Research Protection Office at Washington University School of Medicine (HRPO # 201011766). All the patients provided written informed consent that permitted whole-genome sequencing, in accordance with a protocol that was approved by the institutional review board at the Washington University School of Medicine.

eWGS, germline SNP detection, and somatic variant detection. For each case, we performed enhanced whole genome sequencing (eWGS) on bone marrow and matched normal tissue to identify germline and somatic variants. eWGS combines whole-genome sequencing with targeted exon capture to yield high coverage (~150x) of the exome, and lower genome-wide coverage in the tumor (~45x) and normal (~25x) samples. Using a previously described protocol, eWGS sequencing, including WGS of libraries (350 bp inserts) and targeted libraries (250 bp inserts), were constructed with a KAPA HTP kit on a SciCone instrument. Targeted libraries were captured with the IDT exome reagent spiked with AML recurrently mutated genes (~40 Mb). These were sequenced on an Illumina HiSeq4000, producing ~150X coverage of each enhanced region. Sequence data were aligned to reference sequence builds GRCh37-lite builds73 using BWA-MEM2 version 0.7.10 (params: -t 8), then merged and deduplicated using Picard version 1.113 (https://broadinstitute.github.io/picard/). Germline mutations were called using GATK Haplotypecaller v3.575 (parameters -stand_emit_conf 10 -stand_-

Bulk RNA-sequencing. RNA libraries were prepared using the TruSeq stranded kit, sequenced on the Illumina HiSeq platform, and aligned as described previously. Expression quantification was performed using Kallisto 0.43.163 and transcripts from ensemble version 74.

Flow sorting for live cells for scRNA-seq. Cryoival of AML cells were thawed as follows: while 9 ml of Fetal Bovine Serum (FBS) was allowed to come to ~24°C, AML cryoivals were removed from liquid nitrogen, and warmed at a 37°C water bath until the cells began to thaw. After 1 min, 1 ml of room temperature FBS was added to the warming cryoival and allowed to mix with thawing cells. The freshly added FBS was removed from the cell pellet and transferred back to the FBS stock. This process was repeated 3–4 times until all cells from the cryoival could be poured directly into the FBS stock. The empty cryoival was warmed once more with the FBS mixture. Cells were then pelleted on a 1000x g centrifugation at 300G for 5 min and resuspended in Phosphate-buffered saline (PBS) at a concentration of 1 x 10⁶ cell/ml in 1x PBS. Cells were then pipetted through a 70-µm filter into a 5-ml tube for sorting. Cells were then stained with 1 µl 7-AAD and 1 µl 7-AAD into a 5-ml tube for sorting. Cells were then stained with 1 µl 7-AAD and 1 µl 7-AAD into a 5-ml tube for sorting.
a target of 150,000 reads/cell (2 × 150 paired end reads), yielding a median per-cell depth of 192,427 reads per cell.

**Evaluating transcript coverage as a function of distance.** Transcript alignment, counting, and inter-library normalization were performed using the Cell Ranger pipeline (default settings, V2.1.1, GRCh38 Genome63). For the genes TP53, NPM1, GATA2, and DNMT3A, the depth at each transcript was evaluated using both scRNA-seq data as well as bulk RNA-seq data. For each gene, a canonical isoform was chosen by consulting the APPRIS database65 (ENST00000454888.6, ENST00000296930.9, ENST00000341105.6, and ENST00000347997, respectively). For the scRNA-seq data, the number of unique barcode/UMI pairs was counted at each position. For the bulk RNA-seq data, the tool bamCoverage66 was used to generate a wiggle file over the transcript at 1 bp bin size. The resulting tracks were visualized using the UCSC Genome Browser67. To reduce visual noise from intergenic reads, positions not overlapping the canonical isoform were excluded. Coverage plots for all mutated genes in this study are provided at https://github.com/genome/scrna_mutations.

To evaluate transcription-wide coverage, we used the annotation set GENCODE V27 to extract 20,090 genes with one only annotated isoform between 250 bp and 11,000 bp, with an average size of 1569 bp and median size of 829 bp. Restricting to single isoforms reduced noise related to alternative transcription start (TSS) and stop (TTS) sites. For each transcript in each sample in this study, single-cell transcriptome-wide coverage was quantified by counting the number of unique barcode/UMI pairs seen across the whole transcript. Then, for each position along the transcript, the number of unique pairs was divided by this total. This value was calculated as distance from the TSS for 5′ kth kit, and distance from the TTS for 3′ kth kit. To plot the results, the average value across all transcripts for all samples was calculated at each position. For shorter transcripts, positions with no data were not included in the average. The plot was also truncated to 10,000 bp to avoid edge effects related to the transcript selection process. Coverage plots were generated using the DeepTools66 and BiomasR66 R packages, versions 1.22.3 and 2.3.4 respectively. For each locus, both coding and non-coding exonic nucleotides were considered at a 1 bp bin size. Gene region tracks were retrieved directly from Ensembl v93. scRNA total read coverage was generated using bamCoverage, part of the deepTools package66, and scRNA cell barcode coverage can be found at https://github.com/genome/scrna_mutations.

**Copy number analysis.** Gene expression matrices were analyzed with the CON-ICSmat package for R66. The default filtering and normalization procedures were followed, as outlined in https://goopy.gl/ftyyLeH. The mixture model results were obtained, then restricted to regions of known copy number events from the eWGS kit data, and distance from the eWGS VAF. For each cell, expression of each gene was normalized to the sequencing depth of the cell, scaled to a constant depth (10,000), and log-transformed. Variable genes were filtered for dimension reduction if the majority of the variance were statistically significant according to jackstraw resampling; exhibited consistent expression variation in heatmaps; and were not composed entirely of ribosomal, mitochondrial, or immune genes. Dimensionality reduction and visualization were performed with the t-SNE algorithm (Seurat implementation) using the R package selected above. Unsupervised graph-based clustering of cells was performed using the indicated PCs, with resolution = 0.7. Cell cycle phase was determined using methodology provided in Seurat, based on relative expression of phase-specific genes66. The distribution of mutations on the t-SNE projection was robust to filtering for mitochondrial and ribosomal transcripts, the number of PCs used, the clustering resolution, and normalization for cell cycle phase. The mutation distribution was also robust to the particular implementation of the t-SNE algorithm, with the Seurat and Cell Ranger implementations giving consistent results. To assess the relationship between mutation distribution and expression of the mutated gene, we colored each cluster in each t-SNE plot according to the expression-normalized mutant cell fraction (mutant cell fraction divided by the average expression of the mutant gene in that cluster).

Mutation-expressing cells were analyzed in isolation using analogous methods, with the exception that fewer PCs were required to capture the variability in the data (508084: 4; 548327: 3; 721214: 6; 782328: 7; 809653: 6). **Expression heatmaps.** An expression heatmap was generated for each sample by sorting the top 10 genes in each of the top 20 PCs, that were clustered based on the graph to the graph-based, clusters, and to examine relationships among clusters, we averaged the expression of each gene within each cluster, and hierarchically clustered the results. For the analogous analysis performed on mutant cells in isolation, we used the top 20 genes from each of the top n PCs, where n was chosen separately for each sample to minimize noise (508084: 4; 548327: 3; 721214: 6; 782328: 7; 809653: 6).

**Lineage inference and AML cell identification.** Cell-type inference was performed in an unsupervised, marker-free manner by training a nearest-neighbor algorithm on expression data from the DMAP database65, using Spearman correlation as the distance metric. Using this approach, cells that co-cluster by graph-based clustering tend to have the same inferred lineage and express the corresponding cell-type markers (when known). In the case of AML cells, the assigned lineage represents the normal lineage to which the AML cell is most transcriptionally similar. To identify AML cells in highly heterogeneous samples (549327 and 809653), a one-sided Fisher exact test was used to identify cell clusters that were enriched for somatic mutations (p ≤ 0.05). In cases where most cells are AML cells, normal cell clusters were identified using a one-sided Fisher exact test for under-enrichment (p ≤ 0.05). **GATA2*366*-associated expression signatures.** Each cell containing a GATA2*366* mutation was assigned to an expression cluster. In order to incorporate expression information from all cells in the data set, including those of undetermined genotype, and to make use of quantitative information about local mutation density, we used a regression model to identify genes whose expression depends on mutant cell concentration. For each gene i, multiple regression was...
used to quantify the relationship between mean expression ($E_i$) and GATA2R363C mutant cell fraction ($m$) across the 12 AML clusters, while controlling for mean cluster-wise GATA2 expression ($g$).

$$E_i = x_i + y_i m + e_i$$  

(2)

We selected genes whose $q$-value ($F$-test with Benjamini-Hochberg correction for multiple hypotheses) for $y_i$ was at most 0.05. To consider the entire subclone containing the GATA2 mutation, we performed this procedure using cells containing any detected mutation in that subclone (GATA2R363C or TIMM17B1225), without the correction for GATA2 expression. To analyze the mutant-only data, we used a Wilcoxon test to perform a binary comparison of mutation-rich clusters to mutation-poor clusters, with analogous $p$-value correction and cutoffs.

Identification and analysis of the VIM regulon. Genes that exhibited subclone-specific expression (above) were compared to genes whose expression was highly VIM-correlated in TCGA ($q < 0.001$, Pearson correlation, Benjamini-Hochberg correction for multiple hypotheses). Genes in the intersection were considered part of the "VIM regulon," and ToppFun27 was used to characterize their functional enrichment.

Orthogonal confirmation of GATA2R363C signature. Primary, human AML peripheral blood and bone marrow aspirate samples (721214) were thawed from cryopreserved stocks and labeled with CD99 FITC-conjugated antibody (clone 3B2/TA8, Thermofisher) in staining buffer (2% fetal bovine serum, 0.25 mM EDTA in PBS) for 30 min at 4 °C followed by viability dye for 5 min at room temperature (SytoxBlue, Thermofisher). Live cells were analyzed using a Sony SY3200 Synergy temperature (SytoxBlue, ThermoFisher). Live cells were analyzed using a Sony SY3200 Synergy temperature (SytoxBlue, ThermoFisher). Live cells were analyzed using a Sony SY3200 Synergy temperature (SytoxBlue, ThermoFisher).

Data availability

Enhanced whole-genome sequence (eWGS), bulk RNA-sequence, and single-cell RNA-sequence (scRNA-seq) data generated during the current study are available in dbGaP (https://www.ncbi.nlm.nih.gov/gap/) with the primary accession code phs000159. The SRA IDs for this study are: SRX7904017, SRX7904018, SRX7904019, SRX7904020, SRR7901035, SRR7901035, SRR7901034, SRR7904016, SRR7903979, SRR7825447, SRR7825459, SRR7825456, SRR7825444, SRR7825449, SRR7825459, SRR7825444, SRR7825459, SRR7825449, SRR7825482, and SRR7939318. Processed single-cell RNA-seq and mutation data pertaining to AML samples and normal bone marrow are also available [https://doi.org/10.5281/zenodo.3345981]. All the other data supporting the findings of this study are available within the article and its supplementary information files and from the corresponding author upon reasonable request. A reporting summary for this article is available as a Supplementary Information file.

Code availability


Received: 10 February 2019 Accepted: 23 July 2019 Published online: 14 August 2019

References


70. Heger, A. (psyam-developers/psyam).


Acknowledgements
Supported by the NCI K12 program (CA167540, Rovinjman, P.) to A.P., an R00 CA211782 to C.M., and an R35 CA197561 and P01 CA101937 to T.L. We thank Jonathan L. Weinstein, PhD for advice on statistical methodology, and Julia Lau for construction of the Chromium single-cell gene expression libraries. We also thank Matthew Christopher, MD, PhD and Amanda Smith, PhD for developing the methods used for preparing cryovials of AML cells for scRNA-seq.

Author contributions
Conceptualization by T.L. and D.M.C.; Data curation and Supervision by D.M.C. and T.L.; Funding acquisition and Supervision by D.M.C. and T.L.; Methodology by A.P., S.W., C.M., I.F., S.S., D.M.C. and T.L.; Data analysis and software by A.P., S.W., C.M., I.F. and S.S.; Experimentation by D.Y.C., C.F. and R.F.; Methodology by A.P., S.W., C.M., I.F., S.S., D.M.C. and T.L.; Writing – original draft by A.P.; Writing – review and editing by A.P., S.W., C.M., I.F., S.S., D.M.C. and T.L.; Funding acquisition and Supervision by D.M.C. and T.L.

Additional information
Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-019-11591-1.

Competing interests: S.W. and I.F. are employed by and hold shares in 10x Genomics. The remaining authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Peer review information: Nature Communications thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access
This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019