PhD degree in Systems Medicine (curriculum in Molecular Oncology)

European School of Molecular Medicine (SEMM),

University of Milan and University of Naples “Federico II”

Settore disciplinare: MED/04

Integrated single-cell mutation, gene expression and isoform analysis to deconvolve acute myeloid leukemia heterogeneity

Chiara Caprioli

European Institute of Oncology (IEO)

Tutor: Prof. Pier Giuseppe Pelicci

European Institute of Oncology (IEO)

PhD Coordinator: Prof. Saverio Minucci

Anno accademico 2021-2022

Table of Contents

List of abbreviations in alphabetical order

Abbreviation	Meaning
AlloHSCT	Allogeneic hematopoietic stem cell transplant
AML	Acute myeloid leukemia
AS	Alternative splicing
BCR	B-cell receptor
BM	Bone marrow
BQ	Base quality
CAR	Chimeric antigen receptor
CB	Cellular barcode
CCUS	Clonal cytopenia of unknown significance
cDNA	Coding DNA
CH	Clonal hematopoiesis
CLP	Common lymphoid progenitors
CNV	Copy number variation
CR	Complete remission
dsDNA	Double strand DNA
ELN	European Leukemia Net
FDR	False discovery rate
FSM	Full-splice-matched
gDNA	Genomic DNA
GEM	Gel beads-in-emulsion
GMP	Granulocyte-monocyte progenitors
Gran	Granulocyte progenitors
GVHD	Graft versus host disease
GVL	Graft versus leukemia
HCA	Human Cell Atlas
HMA	Hypomethylating agents
HSPC	Hematopoietic stem and progenitor cells
LMPP	Lymphoid-primed multipotent progenitors
LOH	Loss of heterozygosis
LR	Long-read
LSC	Leukemia stem cell
MCF	Mutant cell fraction
MDS	Myelodysplastic syndrome
MKP	Megakaryocyte progenitors

MNC	Mononuclear cells
MPN	Myeloproliferative neoplasms
MQ	Mapping quality
MRD	Measurable residual disease
Multilin	Multilineage progenitors
NK	Natural killer
NMD	Nonsense-mediated decay
ONT	Oxford Nanopore Technologies
ORF	Open reading frame
PB	Peripheral blood
PCA	Principal component analysis
PCR	Polymerase chain reaction
PFM	Position frequency matrix
PV	Phenotypic volume
QC	Quality control
SCM-seq	Single cell and molecule sequencing
scRNA-seq	Single cell RNA sequencing
scVAF	Single cell variant allelic frequency
SNP	Single nucleotide polymorphism
snRNPs	Small nuclear ribonucleoproteins
SNV	Single nucleotide variant
SR	Short-read
TCR	T-cell receptor
UMAP	Uniform manifold approximation and projection
UMI	unique molecular identifier
VAF	Variant allelic frequency
WES	Whole exome sequencing
WGS	Whole genome sequencing
WHO	World Health Organization
WNN	Weighted nearest neighbor

Figure and table index

Abstract

Acute myeloid leukemia (AML) is an aggressive cancer arising from the hematopoietic stem cell (HSC). As other tumor types, AMLs are characterized by multiple and interconnected levels of intra-tumor heterogeneity, including genetic (DNA mutations), phenotypic (transcriptional patterns) and ecological (interactions with host immune-cells) diversity. Emerging evidence suggest that intra-tumor heterogeneity impacts directly on leukemogenesis, disease prognosis and sensitivity/resistance to available treatments. How the different layers of intra-tumoral heterogeneity interact with each other and shape the different leukemia phenotypes at single-cell level, however, is still missing. One major limit is the lack of technologies allowing ecosystem-wide characterization of tumor samples, including the simultaneous multiomic analyses of both malignant and immune populations at single-cell level. In this work, we have developed a novel high-throughput multiomics approach to integrate gene mutation, expression and isoform information at single-cell resolution. SCM-seq (Single Cell and Molecule sequencing) combines high-throughput droplet-based scRNA-seq to Nanopore single-molecule sequencing of full-length whole transcriptome and enriched mutated transcripts. This technology allows the integration at single cell-level of expression profiles and lineage-imputation (from scRNA-seq data) with mutation burden and transcript isoform diversity (from Nanopore data). We have applied this methodology to the analysis of three AML samples sharing a mutation in a spliceosome factor, with the aim to investigate how phenotypic heterogeneity is related to genetic complexity in both the malignant and immune compartments of a coherent AML subgroup. Results showed that SCM-seq allows multiomic characterization at single-cell level with sufficiently high throughput to represent sample complexity. We identified mutations at cell-level with high sensitivity and were able to stratify groups of cells based on their genetic complexity and mutations co-occurrences. For selected variants, we were also able to genotype both mutant and wild-type cells, which is the premise to investigate genotype-phenotype interactions. HSC/progenitor-like AML cells accumulated higher numbers of mutations and shared specific transcriptional features, including leukemia stem cell properties, thus enabling the identification of the putative malignant compartment of the AML samples. We found, however, that mutant cells were also represented in all remaining hematopoietic lineages, including differentiated myeloid cells and lymphocytes, recapitulating the genetic hierarchy observed in HSCs. Increasing genetic complexity in HSC/progenitor-like AML cells was associated to increasing transcriptional heterogeneity and correlated with the expression of genes and signatures related to cell cycle control, proliferation, stress response, RNA splicing regulation, MTORC1 signaling and MYC targets. Moreover, HSC/progenitor-like AML cells with high mutation burden displayed limited isoform abundance, as related to the number of expressed genes, indicating a progressively restricted repertoire of isoforms in the presence of increasing genetic complexity. In all lineages, the presence of the *SRSF2* mutation was associated to increased isoforms

diversity, with mutated cells carrying significantly higher proportions of genes expressed with more than one isoform or expressing novel or alternative transcripts, as compared to AML *SRSF2*-wild-type cells. Together, these preliminary data show the capability of our method to integrate different sources of AML heterogeneity and their relevance within the tumor ecosystem.

1. Introduction

1.1 Acute myeloid leukemia epidemiology, classification, state-of-the-art clinical management and unmet needs

Acute myeloid leukemia (AML) is an aggressive blood cancer that results from the clonal expansion and uncontrolled proliferation of immature cells (i.e., blasts) involving one or more myeloid lineages in the bone marrow (BM), peripheral blood (PB) and rarely other tissues, leading to impaired hematopoiesis and life-threatening BM failure(1,2). The disease remains largely incurable, with less than 30% of patients surviving more than 5 years despite many advances in biological knowledge and clinical management(3).

1.1.1 Epidemiology

Although AML represents just 1% of cancers, it is the most frequent type of acute leukemia occurring in adult patients. Median age at diagnosis is around 65 years with a worldwide incidence of 2.5-3 cases by 100,000 population *per* year, higher in Western countries and slightly increasing over the last decades(4). A slight male predominance has been reported, and there are no known specific risk factors apart from those generally associated to cancer (such as aging, exposure to radiation or other chemical genotoxic agents, tobacco smoking) and rare inherited germline predisposition (i.e., Fanconi anemia, Schwachman Diamond syndrome, Down syndrome, ataxia-telangiectasia, and others)(5).

1.1.2 Classification

AML is the most aggressive clinical phenotype within a broader spectrum of hematological cancers named myeloid neoplasms, which arise from the hematopoietic stem or progenitor (HSPC) cells in the bone marrow (BM) and express phenotypic features of the myeloid lineage. Besides AMLs, myeloid neoplasms also include myeloproliferative neoplasms (MPN), which are featured by the hyperproliferation of near-normal maturing blood-cells, and myelodysplastic syndromes (MDS), characterized by ineffective hematopoiesis, abnormalities in cell maturation and cytopenias(1).

Advances in clinical and biological characterization across years have been paralleled by constant efforts to develop a meaningful AML classification, acknowledging the heterogeneity of disease presentation. The WHO classification has gradually shifted from morphology- towards genetically-oriented classifications, which now aid both the recognition of distinct pathological entities and the definition of clinical prognosis(1). Along this line, the International Consensus Classification of AML has recently updated the prior revised 4th WHO edition by introducing new genetic entities, thus further expanding the spectrum of classification identified by cytogenetic and mutational profiles (Table 1)(2). Importantly, genetic features are given priority in defining disease entities because of their impact on disease phenotype and disease outcome. For instance, in the presence of recurrent genetic abnormalities (Table 1), the BM blast threshold needed to define AML (previously set at 20%) has been lowered, because the clinical behavior of myeloid neoplasms with these rearrangements reflects the specific genetic abnormality, even for cases presenting with <20% blasts. The same applies to AML cases that lie on the border between AML and MDS in terms of their biology and prognosis (i.e., AML or MDS/AML with MDS-related gene mutations or cytogenetic abnormalities). For the definition of this particular setting, genetic characteristics have been demonstrated to carry more relevance than clinical history or dysplastic morphology(6–8). Additional predisposing features (prior chemotherapy, radiotherapy or immune interventions, prior MDS or MDS/MPN, germline predisposition) may be appended as qualifiers of the primary diagnosis.

Beyond the WHO classification (that is committed to the definition of disease entities), the European Leukemia Net (ELN) has integrated emerging data to develop a genetic-based risk classification of AMLs (Table 2)(3). This classification aims at capturing prognosis factors of first-line chemo-treated AML patients at diagnosis, and it is largely used to guide treatment choice or identify target populations for enrolment into clinical trials. However, since the ELN AML risk classification has been developed based on chemo-treated patients, it may warrant further validation for patients receiving less intensive or novel target therapies. Figure 1 shows the frequencies of ELN risk categories across AML patients treated with chemotherapy within two phase III trials of the German AML Cooperative Group (AMLCG-1999, clinicaltrials.gov identifier NCT00266136, and AMLCG-2008, NCT01382147). Ccorresponding 5-year overall survival rates are reported in Table 2; of note, patients older than 60 years carry the most dismal prognosis(9).

Figure 1. Frequencies of ELN risk categories across AML patients treated with chemotherapy.
Patients received induction chemotherapy within two prospective randomized clinical trials, and were retrospectively stratified by age and updated ELN risk assessment. From Herold et al.(9)

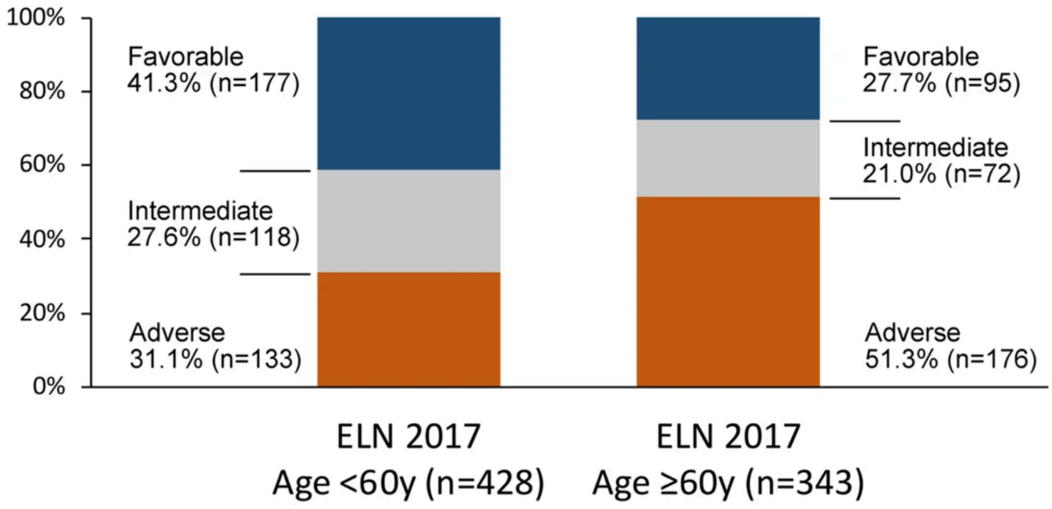


Table 1. International Consensus Classification of AML (2022 update).

Disease entity	Percentage of blasts required for diagnosis
Acute promyelocytic leukemia (APL) with t(15;17)(q24.1;q21.2)/ *PML*::*RARA*	10%
APL with other *RARA* rearrangements*	10%
AML with t(8;21)(q22;q22.1)/*RUNX1*::*RUNX1T1*	10%
AML with inv(16)(p13.1q22) or t(16;16)(p13.1;q22)/*CBFB*::*MYH11*	10%
AML with t(9;11)(p21.3;q23.3)/*MLLT3*::*KMT2A*	10%
AML with other *KMT2A* rearrangements†	10%
AML with t(6;9)(p22.3;q34.1)/*DEK*::*NUP214*	10%
AML with inv(3)(q21.3q26.2) or t(3;3)(q21.3;q26.2)/*GATA2*; *MECOM(EVI1)*	10%
AML with other *MECOM* rearrangements‡	10%
AML with other rare recurring translocations	10%
AML with t(9;22)(q34.1;q11.2)/*BCR*::*ABL1*§	20%
AML with mutated *NPM1*	10%
AML with in-frame bZIP *CEBPA* mutations	10%
AML and MDS/AML with mutated *TP53*†	10-19% (MDS/AML) and 20% (AML)
AML and MDS/AML with myelodysplasia-related gene mutations (*ASXL1*, *BCOR*, *EZH2*, *RUNX1*, *SF3B1*, *SRSF2*, *STAG2*, *U2AF1*, or *ZRSR2*)	10-19% (MDS/AML) and 20% (AML)
AML with myelodysplasia-related cytogenetic abnormalities (complex karyotype, del(5q)/t(5q)/ add(5q), 27/del(7q), 18, del(12p)/t(12p)/add(12p), i(17q), 217/add(17p) or del(17p), del(20q), and/or idic(X)(q13) clonal abnormalities)	10-19% (MDS/AML) and 20% (AML)
AML not otherwise specified (NOS)	10-19% (MDS/AML) and 20% (AML)
Myeloid sarcoma	-

*Includes AMLs with t(1;17)(q42.3;q21.2)/*IRF2BP2::RARA*; t(5;17)(q35.1;q21.2)/*NPM1::RARA*; t(11;17)(q23.2;q21.2)/*ZBTB16::RARA*; cryptic inv(17q) or del(17) (q21.2q21.2)/*STAT5B::RARA*, *STAT3::RARA*; Other genes rarely rearranged with *RARA:TBL1XR1* (3q26.3), *FIP1L1* (4q12), *BCOR* (Xp11.4). †Includes AMLs with t(4;11)(q21.3;q23.3)/AFF1::*KMT2A*$^{\#}$; t(6;11)(q27;q23.3)/*AFDN::KMT2A;* t(10;11)(p12.3;q23.3)/*MLLT10::KMT2A*; t(10;11)(q21.3;q23.3)/*TET1::KMT2A;* t(11;19)(q23.3; p13.1)/*KMT2A::ELL;* t(11;19)(q23.3;p13.3)/*KMT2A::MLLT1* (occurs predominantly in infants and children). ‡Includes AMLs with t(2;3)(p11$23;q26.2)/*MECOM*::?; t(3;8)(q26.2;q24.2)/*MYC*, *MECOM*; t(3;12)(q26.2;p13.2)/*ETV6*::*MECOM*; t(3;21)(q26.2;q22.1)/*MECOM*::*RUNX1*.
§The category of MDS/AML will not be used for AML with *BCR::ABL1* due to its overlap with progression of CML, *BCR::ABL1*-positive.

Table 2. European Leukemia Net genetic risk classification at initial diagnosis (2022 update).

Risk Category[a]	Genetic Abnormality	5-year overall survival[b]
Favorable	• t(8;21)(q22;q22.1)/RUNX1::RUNX1T1[b,c] • inv(16)(p13.1q22) or t(16;16)(p13.1;q22)/CBFB::MYH11[b,c] • Mutated NPM1[b,d] without FLT3-ITD • bZIP in-frame mutated CEBPA[e]	• <60 years: 64% • ≥60 years: 37%
Intermediate	• Mutated NPM1[b,d] with FLT3-ITD • Wild-type NPM1 with FLT3-ITD • t(9;11)(p21.3;q23.3)/MLLT3::KMT2A[b,f] • Cytogenetic and/or molecular abnormalities not classified as favorable or adverse	• <60 years: 42% • ≥60 years: 16%
Adverse	• t(6;9)(p23;q34.1)/DEK::NUP214 • t(v;11q23.3)/KMT2A-rearranged[g] • t(9;22)(q34.1;q11.2)/BCR::ABL1 • t(8;16)(p11;p13)/KAT6A::CREBBP • inv(3)(q21.3q26.2) or t(3;3)(q21.3;q26.2)/GATA2, MECOM(EVI1) • t(3q26.2;v)/MECOM(EVI1)-rearranged • -5 or del(5q); -7; -17/abn(17p) • Complex karyotype,[h] monosomal karyotype[i] • Mutated ASXL1, BCOR, EZH2, RUNX1, SF3B1, SRSF2, STAG2, U2AF1, or ZRSR2[j] • Mutated TP53[k]	• <60 years: 20% • ≥60 years: 6%

[a] Mainly based on results observed in intensively treated patients. Initial risk assignment may change during the treatment course based on the results from analyses of measurable residual disease.

[b] As of Herold T., et al., Leukemia 2020.

[c] Concurrent of KIT and/or FLT3 gene mutation does not alter risk categorization.

[d] AML with NPM1 mutation and adverse-risk cytogenetic abnormalities are categorized as adverse-risk.

[e] Only in-frame mutations affecting the basic leucine zipper (bZIP) region of CEBPA, irrespective whether they occur as monoallelic or biallelic mutations, have been associated with favorable outcome.

[f] The presence of t(9;11)(p21.3;q23.3) takes precedence over rare, concurrent adverse-risk gene mutations.

[g] Excluding KMT2A partial tandem duplication (PTD).

[h] Complex karyotype: ≥3 unrelated chromosome abnormalities in the absence of other class-defining recurring genetic abnormalities; excludes hyperdiploid karyotypes with three or more trisomies (or polysomies) without structural abnormalities.

[i] Monosomal karyotype: presence of two or more distinct monosomies (excluding loss of X or Y), or one single autosomal monosomy in combination with at least one structural chromosome abnormality (excluding core- binding factor AML).

[j] For the time being, these markers should not be used as an adverse prognostic marker if they co-occur with favorable-risk AML subtypes.

[k] TP53 mutation at a variant allele fraction of at least 10%, irrespective of the TP53 allelic status (mono- or biallelic mutation); TP53 mutations are significantly associated with AML with complex and monosomal karyotype.

1.1.3 State-of-the-art clinical management

The goals of AML management are control of life-threatening BM failure symptoms (anemia, bleeding, infections) and, if patients are fit enough to tolerate active treatments, eradication of disease. Unfit patients are instead typically managed with supportive and/or palliative care measures, or enrolment into a clinical trial whenever possible (3).

Active treatment aims at inducing complete remission (CR) (defined as less than 5% BM blasts) after initial therapy, and is followed by consolidation and/or maintenance strategies to deepen the obtained remission and maximize response duration. For patients considered fit for intensive therapy, the combination of cytarabine and anthracyclines (so called '7+3') remain the mainstay of first-line intensive chemotherapy, a regimen that induces >90% of CR. Some alternative options include the addition of fludarabine or mitoxantrone, a novel dual-drug liposomal formulation that encapsulates cytarabine/daunorubicin in a 5:1 fixed molar ratio(10). In recent years, this chemo-based backbone has been associated with some novel agents, including the kinase inhibitors midostaurin(11) or the more potent quizartinib (for patients with *FLT3*-mutant AML)(12), or the anti-CD33 cytotoxic agent gemtuzumab-ozogamicin (of particular benefit on core binding factor AML)(13).

For patients ineligible for intensive chemotherapy the treatment backbone consists of hypomethylating agents (HMA), which achieve clinical responses in only around 35% of cases when administered alone(14,15). Such outcome has substantially improved with the addition of some recently approved biological agents. Most notably, the combination of the BCL2 inhibitor venetoclax has improved clinical response (CR 66.4% vs 28.3% as compared to HMA alone) and median overall survival (14.7 vs 9.6 months), establishing a new standard of care for older or unfit patients even in adverse genetic risk(16). Another successful example of treatment combination is represented by inhibitors IDH1 (ivosidenib) or IDH2 (enasidenib) added to HMA in patients with IDH1/IDH2 mutations (CR 52.8 vs 17.6% and median overall survival 24.0 vs 7.9 months as compared to HMA alone)(17). However, after the achievement of CR, such non-intensive treatment approaches are supposed to be continued until disease progression or decreased tolerability(3).

Consolidation strategies for chemo-treated patients generally include different courses of intermediate-dose cytarabine; in the subset of patients receiving induction with a FLT3 inhibitor, gemtuzumab-ozogamicin or CPX-351, these agents may be incorporated into consolidation. Patients with an estimated relapse risk exceeding treatment-related mortality (35%-40% based on genetic risk) are best managed with consolidation allogeneic hematopoietic stem cell transplant (alloHSCT)(18). In addition to initial risk assessment, monitoring of measurable residual disease (MRD) (by means of multiparameter flow cytometry or qPCR) assists the choice of the best consolidation strategy, as it provides a quantitative methodology to establish a deeper remission status and predict risk of relapse(19,20); indeed, a survival benefit has been shown for non-adverse risk AML patients allocated to alloHSCT due to MRD persistence(19–21).

AlloHSCT is a highly effective treatment to obtain sustained remission and long-term survival due to the immunological eradication of therapy-resistant leukemia cells [i.e., graft-versus-leukemia (GVL) effect(22)]. Potential benefit has to be carefully judged against a generally high burden of morbidity and treatment-related mortality, which precludes access to alloHSCT for the majority of leukemia patients. However, there has been continuous improvement in the development of better selection criteria for candidate patients and donors, less toxic conditioning procedures, graft-versus-host-disease (GVHD) and infection prophylaxis and management, aiming to reach as many patients as possible.

Maintenance treatment was only recently introduced in the scenario of AML treatment. As of the definition of the Food and Drug Administration, it is an extended but time-limited course of treatment with a minimally toxic therapy, administered after the achievement of CR with specific regimens and capable of reducing the risk of leukemic relapse. For instance, patients receiving midostaurin during induction and consolidation chemotherapy may continue these agents, although the value of adding maintenance therapy remains uncertain(11). In addition, two randomized studies have explored the use of HMA maintenance therapy in older patients in first remission after two cycles of intensive induction and not considered candidates for alloHSCT, but the approach is still controversial(23,24).

1.1.4 Unmet clinical needs and areas of research

Long-term survival outcomes of AML patients remain unsatisfactory, especially for those with adverse genetic features, older age or unfit for intensive treatment. In all cases, including more favorable risk groups, treatment resistance is identified as the main cause of death, with less than 30% of patients surviving more than 5 years(25).

Failure to achieve CR after two cycles of induction (including at least one cycle of intermediate-dose cytarabine, i.e. primary refractory AML), is the earliest scenario of treatment resistance, which occurs in 10-50% of patients depending on risk category. Patients in this group have a dismal prognosis, are unlikely to benefit from further cycles of conventional chemotherapy and are typically referred to a clinical trial with alternative agents and/or alloHSCT, which is the only curative therapy in this setting even though anti-tumor activity is less powerful in patients with active disease. As a consequence, one area of active research aims at i) defining predictors of primary chemoresistance, ii) developing frontline strategies devised to minimize the risk of failure and obtain deeper remission, and iii) finding effective salvage regimens. While many novel drugs have have been tested in recent years, the administration of a single-agent usually does not achieve long-term disease control, and combination therapies - i.e., novel agents with a chemo- or HMA-based backbone - represent the most promising approach. In selected cases single agents can work as a bridge to transplant, such as with the potent FLT3 kinase inhibitor gilteritinib or IDH1/IDH2 inhibitors(26–28).

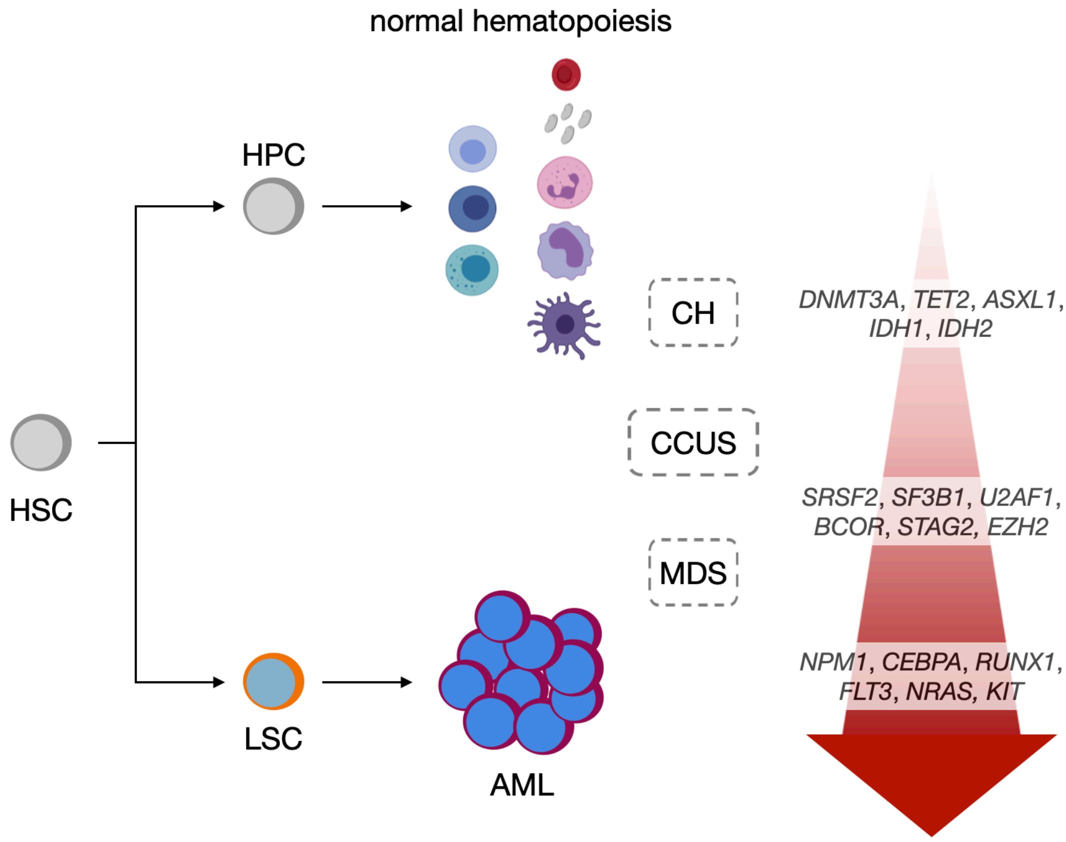
The main cause for treatment resistance and poor long-term survival, however, is disease relapse after attaining CR, a situation occurring in 40–50% of young patients and the great majority of elederly. Relapse can be seen as a delayed expression of therapy resistance, as it is driven by drug-resistant leukemia stem cells (LSCs) with dormancy and self-renewal properties(29). Understanding the biological context, mechanisms and dynamics of AML relapse remains the primary objective of base, translational and clinical research. In this setting, prognosis is generally poor but can widely vary depending on the timing of relapse (as late relapses can frequently respond again to chemotherapy), genetic profile and the possibility of performing alloHSCT. Relapse is also a common event after alloHSCT, most commonly within 2 years after transplant and with particularly poor outcomes when occurring within the first 12 months(3). There is large interest in developing strategies to prevent post-transplant relapse, basing on three main players: vulnerable pharmacological targets (such as HMA, FLT3 or IDH inhibitors), choice and management of immunosuppression (to potentiate GVL while minimizing GVHD) and immunotherapies [donor lymphocyte infusion, bi-specific T-cell engaging antibodies, checkpoint inhibitors, and chimeric antigen receptor (CAR) T cells or NK cells]. This latter approach stems from the observation of potential GVL after alloHSCT and aims to overcome therapy resistance by harnessing the immune system against tumor cells. However, although the strategy is conceptually promising, results of immunotherapy clinical trials have been far less successful in AML as compared to other cancer types(30).

1.2 Landscape and clinical correlates of AML intra-tumor heterogeneity

1.2.1 Genomic patterns and clonal evolution

Genetics plays a leading role in driving development of AML, as leukemia progresses from the accumulation of somatic mutations in HSC/HPCs, which populate the bulk leukemia with multiple cellular clones endowed with the capability to self-renew and propagate(31). The process may start much earlier than the appearance of a full-blown leukemia, as somatic mutations can be identified even decades before AML diagnosis in the peripheral blood of healthy individuals - a condition known as clonal hematopoiesis (CH) and reflecting the expansion of mutated HSCs. CH increases in prevalence with age (>10% of individuals over age 65), involves both HSPCs and differentiated hematopoietic lineages and can eventually progress to AML(32) following accumulation of further genomic alterations along a spectrum of diverse clinical and phenotypic features, such as clonal cytopenia of unknown significance (CCUS) or MDS (Figure 2) (clonal evolution).

Figure 2. Patterns of clonal evolution during AML development.
In healthy BM, HSC cells and HPC originate all the differentiated lineages of hematopoiesis. The appearance of gene mutations causes the switch from normal hematopoiesis to CH, which, however, still preserves differentiation and most of related functions. Eventually, with the expansion of clonality and the acquisition of phase-specific mutations, hematopoietic differentiation is progressively impaired. At AML diagnosis, hematopoiesis is almost completely substituted with immature blasts.


normal hematopoiesis
HPC
HSC
LSC
AML
CH
CCUS
MDS
DNMT3A, *TET2*, *ASXL1*, *IDH1*, *IDH2*
SRSF2, *SF3B1*, *U2AF1*, *BCOR*, *STAG2*, *EZH2*
NPM1, *CEBPA*, *RUNX1*, *FLT3*, *NRAS*, *KIT*

Overt AML implies the coexistence of genetically-distinct clones, whose combinations of mutations is influenced by patterns of biological cooperativity and mutual exclusivity among mutated genes. More than 95% of AML cases harbor at least 1 somatic alteration, with approximately a dozen of alterations identified *per* AML sample including an average of 3 driver mutations (7,33–35). While mutation burden and heterogeneity are relatively low in AML as compared to other cancer types, differences in genomic profiles account for a major source of inter-patient heterogeneity and have been exploited to derive a clinical classification of AMLs. This classification has also biological significance (7,33–35), as most commonly mutated genes in AMLs can be grouped according to different functional categories. Of note, mutations within the same functional categories are largely mutually exclusive, suggesting synergistic effects among recurrent mutations of distinct functional groups. Main genomic functional categories are summarized in Table 3.

Table 3. Functional categories of main driver genes in AML.
Main AML driver genes are presented by their functional role in leukemogenesis; incidence is given according to the TCGA study(33), and might change in different cohorts. From Bullinger L. et al.(36).

Functional category	Selected gene members	Role in leukemogenesis	Incidence in the TCGA cohort (%)[a]
Signaling	Kinases, phosphatases, RAS family members	Proliferative advantage through RAS/ RAF, JAK/STAT, and PI3K/AKT signaling pathways	59
DNA methylation	*DNMT3A*, *TET2*, *IDH1*, *IDH2*	Deregulated DNA methylation patterns leading to transcriptional deregulation of AML-relevant genes	44
Myeloid transcription factors	*RUNX1*, *CEBPA* (mutations); t(8:21), inv(16)/t(16;16) (fusions)	Aberrant transcription factor function resulting in transcriptional deregulation and impaired hematopoietic differentiation	18; 22
Chromatin modification	*ASXL1*, *EZH2* (mutations); *KMT2A* (fusions)	Transcriptional deregulation	30
Tumor suppression	*TP53*, *WT1*, *PHF6*	Transcriptional deregulation and impaired degradation through MDM2 and PTEN	27
Spliceosome complex	*SRSF2*, *SF3B1*, *U2AF1*, *ZRSR2*	Impaired spliceosome function and deregulated RNA processing resulting in aberrant splicing patterns	16
Cohesin complex	*STAG2*, *RAD21*	Impaired chromosome segregation and impact on transcriptional regulation	14
Nucleophosmin	*NPM1*	Aberrant cytoplasmic localization of NPM1 and NPM1-interacting proteins	13
[a]Median age 55 years; incidence may differ for patients cohorts with higher median age.			

Analyses of the variant allele frequency (VAF) in whole-genome sequencing studies have shown the complexity of the clonal architecture of AMLs and served as a proxy to infer the temporal order of the acquisition of mutations and to identify the founding leukemia clone (the clone showing the highest VAF values) from more recently-acquired subclones (clones with relatively lower VAFs). Thus, the clonal architecture is considered a dynamic process that can vary across the different phases of disease progression or following treatment (Figure 2). Mutations in genes involved in epigenetic, splicing and apoptosis regulation tend to occur as early founder events in preleukemic progenitor cells and are prevalent in CH (e.g., *DNMT3A*, *TET2*, *ASXL1, IDH1, IDH2*) or MDS (*SRSF2*, *SF3B1*, *U2AF1, EZH2, RUNX1, STAG2, TP53, PPM1D*), and also found in overt AML. At the contrary, mutations in *NPM1* and transcription factors are later leukemogenic events, while mutations in signaling pathways often represent subclonal mutations (e.g., *FLT3*, *NRAS*, *KRAS*, *KIT*, *PTPN11*)(36).

As discussed in chapter 1.1, individual genomic profiles provide prognostic information (especially in patients with normal or intermediate-risk cytogenetics) and may be used to guide treatment choice, such as the use of alloHSCT in high-risk patients(3). Also, given the high frequency, prognostic impact and functional implications of certain gene mutations, in recent years many efforts have been put to the development of targeted therapeutic strategies directed at specific mutations; a non-exhaustive view is provided in Table 4.

Table 4. Selected target therapies directed at gene mutations and related pathways in AML.

Targets	Targeted agents
Tyrosine and signaling kin	Midostaurin(11), sorafenib, crenolanib(37), quizartinib(12), gilteritinib(26) (FLT3) BGB324 (AXL) BKM120, BYL719, and TGR-1202 (PI3K) Trametinib (RAS(38)) GSK2141795 (AKT)
DNA methylation	Ivosidenib (IDH1)(27) Enasidenib (IDH2)(28) Hypomethylating agents (azacitidine(15), decitabine(14), CC-486(24))
Spliceosome complex	H3B-8800 (SF3B1)(39) GSK3326595 and JNJ-64619178 (PRMT5(40))

The improvement in survival conferred by such agents in clinical randomized studies is usually modest and difficult to generalize to real-life patients, but the observed disease-modifying activity in some cases (e.g., molecular clearance achieved with IDH inhibitors) is a key aspect for the implementation of novel chemo- or HMA-based treatment. Most promising results might be obtained combining standard regimens with one or more target agents, aiming to maximize MRD clearence and benefits of alloHSCT. The establishment of most rational combinations and effective treatment strategies is indeed still matter of intense investigation.

However, as mentioned above, the clonal composition of each AML case is complex and unique, and it can be argued that the overall effect of drugs designed to target a specific mutation might be minimal when the targeted mutation is represented at subclonal level. In addition, clonal evolution occurs also after AML treatment, including both standard chemotherapy and novel agents targeting AML-associated mutations, as demonstrated by genomic sequencing of paired diagnosis and relapse AML samples(41–43).

Historically, chemotherapy is thought to induce acquisition of mutations conferring drug resistance, thus causing therapeutic failure. In some experimentally documented cases, however, relapse was supported by minor genetic subclones already present at diagnosis, suggesting that cells bearing drug-resistant mutations may exist before treatment and be selected by therapeutic pressure, rather than being directly induced(29). Although the mechanisms of chemoresistance and the specific impact of genomic changes are not yet fully elucidated, the emergence of secondary resistance mutations following targeted therapy is a well- documented issue, especially in the single-agent setting(44,45).

In summary, genomic studies from recent years have boosted our understanding of AML diversity, allowing to solve inter-patient heterogeneity into discrete categories that have entered clinical practice due to their impact on prognosis and, in selected cases, therapeutic decisions. The recognition of prevailing mutational patterns and functional associations has been fundamental to start the transition from standard chemotherapy to more personalized treatments targeting precise molecular mechanisms. For instance, the Beat AML study provided associations of drug response with mutational status, including instances of drug sensitivity that are specific to combinatorial mutational events, and involved gene networks(35). However, knowledge about intra-tumor genetic clonal evolution has not yet translated into effective strategies to predict and tackle therapeutic resistance, which stands as the main cause of death in AML patients.

1.2.2 Intra-tumor phenotypic heterogeneity beyond genetics

Various factors beyond somatic mutations may contribute to intra-tumor heterogeneity. Indeed, studies on patient-derived xenografts have demonstrated that AML cells sharing the same driver mutation can exhibit functional differences(46), while some observations in CH studies proposed that clones driven by the same or similar mutations can behave differently between individuals, suggesting that non-genetic factors might influence clonal evolution as well(47,48). As a clinical correlate, genomic profiling has limited predictive value for therapies targeting specific biological processes, as in the case of the anti BCL2 agent venetoclax(16,49).

A first level of phenotypic heterogeneity is related to AML cell lineage hierarchies, which mimic normal blood development with varying degrees of distortion depending on the cell of origin. AML has been shown to be maintained by rare LSCs that stay at the apex of cellular hierarchies and display stem cell properties, including self-renewal, quiescence and therapy resistance(50). The cellular composition of the leukemic hierarchy likely reflects the functional consequences of specific mutations on LSCs(51). Importantly, LSCs may propagate disease relapse through non-genetic mechanisms, and LSC-related gene expression signatures have emerged as predictors of chemoresistance(52). Functional and transcriptional stemness properties are shared between distinct phenotypic patterns of relapse, highlighting the need to therapeutically address stemness to prevent relapse(29). In animal models, LSCs showed intrinsic, non-genetic properties of malignant clonal dominance(53). One recent seminal study beautifully deconvolved the relationship between LSC properties and genetic alterations, by analyzing patient-specific variation in hierarchy composition across large AML cohorts and integrating information from genomic profiles, functional stem cell properties and clinical outcomes. Notably, variations in hierarchy composition were associated with response to chemotherapy or drug sensitivity profiles of targeted therapies, confirming that genetics alone fails to predict response to treatment(54).

In parallel, there is growing interest in understanding how AML cell hierarchies and fates are shaped by immune-related properties. The tumor immune microenvironment consists of multiple players, including adaptive and innate immune cells and stromal components, which may either antagonize or promote tumor progression. AML cells themselves may exhibit immunomodulatory properties, especially differentiated, non-LSC AML cells that can interact with microenvironmental components of the tumor niche(55,56). In particular, the expression of immunomodulatory factors can instruct surrounding T cells to switch from cytotoxic to suppressive functions, thus promoting the evasion of AML cells from immune surveillance. The immune milieu itself, in turn, can act as an extrinsic regulator of tumor fitness and quiescence properties of LSCs(55,57). There are multiple lines of evidence highlighting the role of different mechanisms of immune-evasion during AML development and clonal expansion: for instance, dysregulation of innate immune and inflammatory cells and signalling contributes to the competitive advantage of CH-mutant HSC during aging, particularly in the context of *TET2*, *DNMT3A* and *JAK2* mutations(58). In addition, mutations associated with CH are nearly always present in circulating innate immune cells and, less frequently, in the T and B lymphoid compartment(59), which might affect immune surveillance against emerging tumor cells and response to immune therapies. Proof-of-concept studies in murine models have proposed that immune-related pathways could be targeted for reducing the selective advantage of CH or MDS transforming clones(60,61). The risk of leukemic transformation significantly varies across CH-individuals and pre-leukemic patients and is associated to diverse synergistic combinations of mutations(31,48,62), which suggests the existence of connections between genetic evolution, immune response (i.e., tolerance vs escape) and clinical correlations. Preliminary evidence supports this perspective with different mechanisms, including: i) the expansion of specific immune populations (e.g. in MDS, where chromosome 8 trisomy and consequent WT1 overexpression fuel CD8+ expansion(63)); ii) the up/downregulation of immune effectors activity (e.g., fusion proteins *PML-RARa* and *AML1-ETO* impair NK cytolytic activity by downregulating their receptor's ligand CD48 on AML cells(64)); iii) the enhancement of specific signalling and immune activation pathways (such as for mutations in *JAK2*(65,66) or spliceosome genes(67,56), which are early genetic events in AML, or for signalling effector mutations, which occur in late AML subclones). Although not specifically focusing on the immune microenvironment, Miles et al. observed differential skew to the myeloid, B or T cell lineages, depending on which CH gene was mutated. Genotype-driven changes in cell-surface protein expression were also reported in the leukemic phase, with signaling effector mutations leading to increased CD11b expression(68).

As a clinical correlate for the role of the immune microenvironment in AML, alloHSCT currently is the only strategy to overcome chemoresistance and obtain sustained remission by immunological eradication of therapy-resistant LSC, even in high-risk genetic subsets. However, post-transplant relapse is a common occurrence, due to several leukemia-driven immune-escape mechanisms(69,70,56). These observations point toward the strong need of developing more potent, specific and possibly less toxic immunotherapeutic strategies for AML patients. To date, these include harnessing T and NK-cell-mediated tumor clearance by checkpoint inhibitors, monoclonal antibodies, bispecific antibodies or CAR T cells; however, their effect has been less successful in AML than in other cancers(30,71,56,55). Reasons for this failure include a limited power of the currently used immunological markers to predict clinical response, the absence of a suitable leukemia-specific target antigen and elusive resistance mechanisms. Thus, ongoing research efforts are committed to the discovery of druggable targets or mechanisms and more effective therapeutic combinations, which would benefit from a better understanding of the various cellular and functional components of the immune microenvironment. In this context, common AML-associated translocations (*AML1-ETO, DEC-CAN, PML-RAR*a*, BCR-ABL*) or mutations (*FLT3*-ITD, *NPM1*, *IDH1*^{R132H}, mutations in spliceosome genes and some *TP53* hotspots, *JAK2*, *CALR*) produce tumor-specific immunogenic proteins that may become ideal antigen targets for the development of immunotherapies(56,72). Finally, gene expression and spatial profiling of the BM immune context in AML patients have allowed to define distinct functional classes within the immune microenvironment; in particular, IFNγ-related mRNA profiles are associated to outcomes of chemotherapy and flotetuzumab immunotherapy, beyond the capabilities of single molecular markers(73). On the same line, a recent preprint paper reported that transcriptomic features of senescence in cytotoxic T cells correlate with adverse-risk molecular lesions, stemness, poor survival and response to immune checkpoint blockade(74).

To sum up, several indications from different angles suggest that genetics alone dose not explain all AML cellular fates and functions, thus highlighting the need of interrogating intra-tumor heterogeneity from a wider perspective and with more comprehensive approaches.

1.2.3 Alternative mRNA splicing: linking phenotypic heterogeneity, AML genetics and immune response

An additional layer of AML phenotypic heterogeneity is related to the repertoire of transcript isoforms that are generated upon expression of certain genes under certain conditions, along with the underlying alternative splicing (AS) events (Figure 3). AS is a fundamental biological mechanism that occurs in the vast majority of human genes to generate multiple potential protein-encoding mature mRNAs from a single gene, thus resulting in proteome diversity and expanding the possibility of phenotypic adaptation. RNA splicing is also an essential regulator of gene expression, as it can generate mRNA species that are targeted for degradation and regulate the expression of non–protein-coding RNAs. In AML, altered splicing can occur as a mutation-dependent or mutation-independent mechanism.

Mutations in genes encoding for members of the spliceosome complex (*SF3B1*, *SRSF2*, *U2AF1*, and *ZRSR2*) are frequent in myeloid neoplasms and can be identified in 10-15% of AML patients(7,8,34,35), especially those with older age and an antecedent history of MDS(6,8). Of note, the presence of mutations belonging to this functional group confers high-risk prognosis at the diagnosis of AML, with poor rates of response to standard treatment and decreased survival(6,8,75). Many of the genetic alterations influencing splicing affect proteins involved in the initial steps of spliceosome assembly. Physiologically, splicing occurs in the cell nucleus and stems from the processing of 5' and 3' splice sites at either ends of each intron and involves branch point sequences upstream of the 3' splice site and polypyrimidine tracts that are critical for splicing accuracy. In the initial steps of mRNA splicing, the 5' site is identified and bound to the branch point sequence by the U1 and U2 small nuclear ribonucleoproteins (snRNPs), the latter containing SF3B1, which on its turn binds to the branch point sequence. U2AF1 is involved in recognizing the 3' splice site, while SRSF2, a member of the serine/arginine protein family, functions as a splicing enhancer, with a role in exon recognition. ZRSR2 appears to have a role analogous to U2AF1 but as part of the minor splicing complex(76,77) (Figure 4).

Figure 3. Basic modes of alternative splicing events.

In exon skipping (the most common event in mammalians), an exon may be spliced out of the transcript or retained. In alternative donor site, an alternative 5’ or 3’ splice junction (donor site) is used, changing the boundary of exon. In mutually exclusive exons, one of two exons is retained in mRNAs after splicing, but not both. In intron retention (the least common event in mammalians), a sequence may be spliced out as an intron or simply retained. If the retained intron is in the coding region, the intron must encode amino acids in frame with the neighboring exons, or a stop codon or a shift in the reading frame will cause the protein to be non-functional.

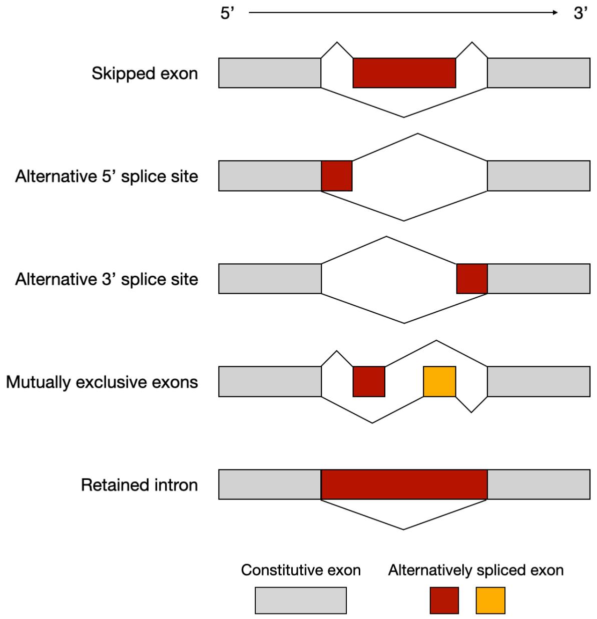
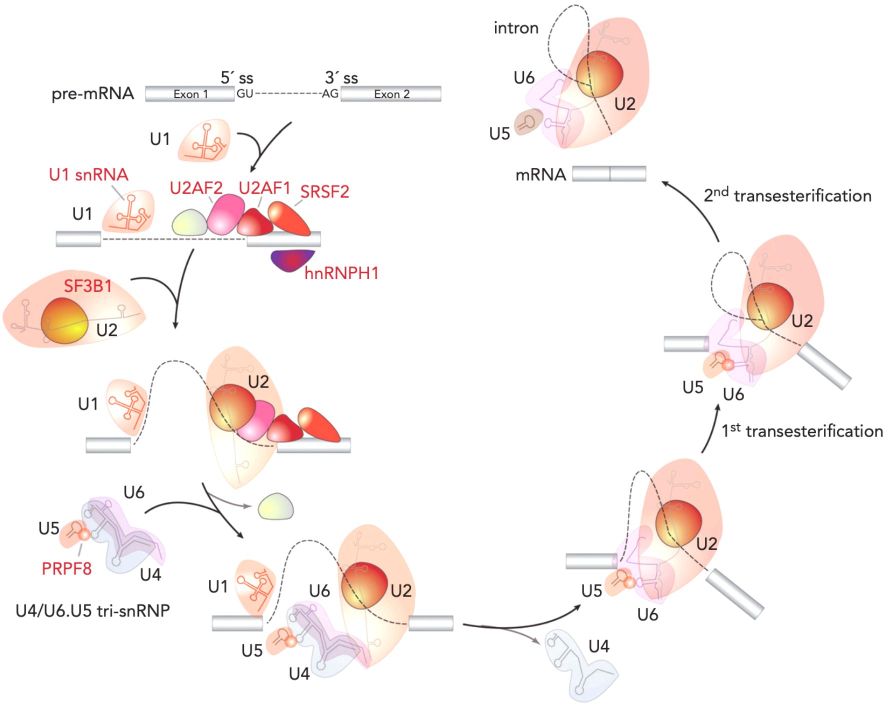

5’
3’
Skipped exon
Alternative 5’ splice site
Alternative 3’ splice site
Mutually exclusive exons
Retained intron
Constitutive exon
Alternatively spliced exon

Figure 4. The spliceosome complex and main steps of pre-mRNA splicing.

The enzymatic steps of splicing are carried out by groups of proteins complexed with small nuclear RNAs termed snRNPs. Factors labeled in red in this diagram undergo recurrent mutations in patients with hematologic malignancies.



Mutations in splicing factors result in characteristic alterations in pre-mRNA splicing based on their specific functions: mutations in *SF3B1* causes alternative branchpoint usage, mutations in *U2AF1* are associated with altered splice site recognition, while mutations in *SRSF2* result in altered recognition of exon splicing enhancers, inducing altered splicing of a number of proteins including critical regulators of hematopoiesis, like EZH2(78). Of note, spliceosome mutations are mutually exclusive and always co-expressed with the wild-type allele, indicating that one functional allele is required for cellular integrity - which suggests potential for exploiting synthetic lethality strategies(7,79–81). Instead, there is frequent co-occurrence of mutations affecting spliceosome factors and epigenetic regulators. A recurrent association between mutations in *IDH2* and *SRSF2* has been shown to promote leukaemogenesis through coordinated effects on the epigenome and RNA splicing, providing functional evidence that mutations in splicing factors drive myeloid malignancy development(82).

The effect of splicing factor mutations extends beyond mRNA splicing itself, because splicing occurs in conjunction with mRNA transcription and any alterations in splicing can also affect the efficiency and integrity of transcription. Hence, AS slows the replication fork and results in DNA:RNA hybrid structures termed R-loops, which are unstable structures that displace the non-hybridized DNA strand resulting in single- stranded DNA, which, in turn, activates DNA damage responses mediated by ataxia telangiectasia and rad3 related (ATR) signaling(83,84). Interrupting this DNA damage response may hamper genomic integrity leading to selective apoptosis of splicing factor mutant cells. Moreover, there is increasing interest to understand whether and how mutations in RNA splicing factors might affect nonsense-mediated mRNA decay (NMD), a translation-coupled surveillance pathway that reduces errors in gene expression by eliminating mRNA transcripts that contain premature stop codons; for instance, it has been reported that SRSF2^{P95} mutants, but not wild-type SRSF2, enhance NMD(85).

AS is frequently altered in cancer and may impact cancer functions beyond the presence of mutations affecting the spliceosome machinery(86–88). One recent seminal study analyzed 32 non-hematopoietic cancer types from the TCGA using WES and bulk RNA sequencing in parallel with tissue-matched normal controls, and showed increased AS events in tumors as compared to normal tissues, with higher numbers of novel (i.e., not previously annotated in reference databases) exon-exon junctions(88). Alterations in splicing were associated to known mutations in splicing factors, but also to new and unexpected variants, as similarly reported in another study(81). Results from Kahles et al. suggested that tumor-specific AS events are far more abundant than somatic single-nucleotide variants (SNVs) and, together with another TCGA study(87), predicted that polypeptides generated from cancer-specific novel junctions have the potential to bind MHC-I and serve as a neoantigens. Although the immunogenicity of such AS-derived neoantigens requires further elucidation and experimental validation, these data represent an important link between genetic and phenotypic determinants of cancer diversity, together with related immune response. In myeloid neoplasms, characteristic

RNA splicing isoform expression patterns have been found to distinguish normal HSCs, aging HSCs and malignant progenitors of MDS and AML(89). Cumulative DNA damage response to AS events, as well as aberrant isoforms themselves, might result in tumor-specific antigens that enhance immunogenicity on their own or in combination with immunotherapies. In AML, one study has assessed patterns and impact of AS on RNA sequencing data from low- and high-risk patients, as determined by genetic risk stratification and clinical prognosis, excluding cases with mutations in spliceosome factors(90). The Authors found that widespread and recurrent AS differences exist between AML patients with good or poor prognosis, the latter being associated to aberrant splicing of protein translation genes that triggers the induction of an integrated stress response and concomitant inflammatory response. However, the functional associations to genetic complexity and immune response, including immune evasion, are currently unknown.

Altered dependency on the spliceosome as well as AS are seen as attractive therapeutic targets in myeloid neoplasms, due to the prevalence of splicing factor mutations and the associated high-risk prognosis. Besides, these mutations usually present early in the course of disease and tend to persist after treatment even when clinical response occurs, suggesting that targeting may be more likely to impact the malignant founder clone, eventually eliciting durable responses(91,92). In pre-clinical models, treatment with spliceosome modulators has shown to impair AML LSC maintenance by abolishing pro-survival splice isoforms(89). Approaches exploiting synthetic lethality strategies seem particularly reasonable in the context of spliceosome-mutated AMLs. Studies in animal models have led to hypothesize a mechanism of action for spliceosome modulators, based on which a cell can tolerate a limited amount of splicing dysfunction before undergoing selective cell death(93,94). However, despite the solid rationale, a phase I first-in-human trial with the oral SF3B1 modulator H3B-8800 yielded poor response in AML patients(95). Combinations with immunotherapies may be warranted in the future.

1.3 Single-cell multi-omics for the characterization of intra-tumor heterogeneity

1.3.1 Limits of bulk sequencing and advantages of single-cell sequencing approaches

Traditional bulk-sequencing approaches rely on the analysis of whole samples through next-generation sequencing platforms, which generate multiple sequencing reads covering individual RNA or DNA molecules. As such, the output of bulk sequencing represents an “average” of the transcriptomic or genomic features of all sample cells, which poses a challenge in the precise deconvolution of intra-tumor heterogeneity. Dedicated bioinformatic tools have been developed to estimate the relative proportions of cell types in complex tissues from their gene expression profiles (e.g. CIBERSORT(96)), but low intensity signals from rare cell populations might result undetectable using bulk sequencing approaches, which precludes identification of rare (yet potentially relevant functionally) cell populations. Conversely, single-cell approaches allow the characterization of individual cells, thus providing a more faithful representation of the heterogeneity of tumor ecosystems and allowing the identification of even rare cell types(97–99). The use of single-cell technologies for research purposes is rapidly spreading, favored by combined academic and industrial efforts to improve standardization, develop several different applicati ons and technological platforms and decrease costs. Some key aspects of single-cell approaches offer relevant advancement in the characterization of intra-tumor heterogeneity. Both tumor and immune cells can be acquired in parallel without prior marker-based sorting, thus avoiding the bias of predefined lineage-markers, because the analysis of individual-cell transcriptional states can independently reconstruct cellular phenotypic traits(98). The high resolution of single-cell methods enables the investigation of even small groups of cells, which can be analyzed for both their phenotypic traits (e.g., surface markers, cell types) and functional states (e.g., over- expressed pathways, genomic features, activation of signalling pathways). Finally, the throughput of some sequencing platforms (up to thousands of cells)(100) provides unprecedented statistical power and allow grouping cells with shared features of interest.

1.3.2 Main single-cell approaches and challenges

Recent advances in cell isolation methods and automated micro-fluidics techniques have improved tremendously the accuracy, sensitivity, reproducibility, and throughput of single-cell RNA sequencing (scRNA-seq), by which it is now possible to measure and model gene expression profiles from thousands of cells simultaneously(101–103,100,104). A few scRNA-seq platforms are available on the market, differing by protocol complexity, costs, number of output cells, sequencing depth and full or partial coverage of transcripts. Such elements, as well as downstream analysis-pipelines, should be considered in view of the specific research-question. Library construction methods that allow full transcript coverage(105,106) are optimal for scoring expressed mutations, splicing isoforms(107) and T/B-cell receptor sequence(107–110), while molecular-counting methods based on the sequence of the 5’ or 3’ end transcripts are better suited for cost-effective profiling of high numbers of cells and transcripts(102). The introduction of unique molecular identifiers (UMI) during library preparation allows counting and grouping specific mRNA molecules prior to PCR amplification, thus increasing accuracy and reducing technical artifacts(111). High-dimensional scRNA-seq data need to be processed with specific computational algorithms, which incorporate various steps of quality control, normalization and dimensionality reduction to enable bidimensional representation(112). Opportunities from downstream analyses include unbiased clustering to identify groups of transcriptionally related cells, differential gene expression(113), and reconstructing dynamic biological processes, such as cellular differentiation and immune response, by inferring developmental ‘trajectories’ to reveal transitional states and cell fate decisions of distinct cell subpopulations(114).

Single-cell DNA sequencing (scDNA-seq) overcomes the limits of bulk sequencing allowing the direct identification of intratumoral genetic subclones - as defined by mutations co-occurring within the same cell - including rare clones, which may significantly impact tumor evolution and the acquisition of therapeutic resistance. The technique’s core involves whole-genome amplification (WGA) of single cells, which allows detection of single nucleotide variations, chromosomal copy number alterations or more complex genomic rearrangements. Droplet-based platforms currently enable high-throughput and cost-effective characterization of hundreds of amplicons in thousands of cells(115). However, a drawback of scDNA-seq methods is the high rate of false negative and false positive hits, due to artifacts introduced during genomic amplification, non-uniform genome-coverage and allelic dropout events.

Along with scientific opportunities, the adoption of single-cell technologies implies dealing with specific experimental and computational/statistical challenges, which are often shared across the different single-cell applications(113). From the experimental point of view, the generation of single-cell data from a biological sample typically requires some common key steps, including dissociation of cells from the tissue of interest, cell purification and isolation, library construction and sequencing. Each step impacts significantly the output results for downstream analyses. For instance, in scRNA-seq protocols, sample preparation and handling have to be carefully planned to avoid unnecessary stressful conditions, which are known to induce extensive cellular responses, thus introducing artifactual modifications of transcriptional states(116). The emergence of microfluidics techniques for cell isolation and combinatorial indexing strategies scaled up the number of cells being sequenced in one experiment and recently enabled multiplexing of different samples. Experimental steps, however, may result in considerable batch effect during later analysis and become the source of technical noise; this might be the case with protocols that use WGA, or with carrying over of empty droplets during library preparation, cell doublets or dying cells. In parallel, recurring computational challenges exist, due to inherent features of the sequencing data(29). The amount of material sequenced from each single cell is considerably less than that available from bulk experiments, which leads to high levels of missing data. Missings may be due to technical dropouts (depending on platform and sequencing depth) or reflect true biological signal (as for variations in expression levels of a given gene). This condition requires strategies to impute missing values, which have been more successful for genotype data than for transcriptomic data(117). Conversely, any increase in the number of analyzed cells and features translates in the need of scalable data analysis models and methods. As a further complication, high-dimensional single-cell data have to be processed for easier tractability, while preserving the salient biological signals of the overall dataset. Another common challenging task is the integration of multiple datasets for comparative analyses across multiple samples (even from same samples from different experiments or experimental conditions). Computational approaches have been devised to score pairwise correspondences between single cells across datasets, enabling batch-effect correction and identification of populations with common sources of variation. This procedure, however, brings the inherent risk of overcorrection and should be applied cautiously(118–121). With the advancement of innovative methodologies, the number and scale of publicly available datasets are continuously increasing; this offers the opportunity to integrate and interrogate multiple datasets for the validation of previous discoveries or, conversely, the generation of new hypotheses to be experimentally validated, and will possibly allow the construction of a specific cell-type atlas for both cancer and immune cells(122,123). Proper curation, quality control and reliable computational strategies for integration are essential to the full exploitation of available data. However, comprehensive integration is challenging because datasets are typically generated through a variety of different approaches and heterogeneous study designs. To this aim, achieving standardization of experimental protocols will play an important role.

1.3.3 Single-cell multiomics approaches and opportunities to address key questions in AML

Several emerging single-cell technologies are committed to recording complementary types of cellular and molecular information from the same cell, including its transcriptome, genome, epigenome and proteome. The application of multiomics approaches enables the integration of different molecular layers within single cells at the same time and, possibly, with respect to their surrounding environment, thus providing a more holistic view of cellular processes and an unprecedented description of the cancer ecosystem.

Because of the prominent role of genetics in cancer biology and clinical management, most efforts have converged on the development of technologies that jointly capture a single cell's genomic profile along with its phenotypes defined by either surface markers or functional features. A number of strategies have been reported, each with its own strengths and limits. Table 5 provides a summary of the main approaches devised to couple genetic and phenotypic information. Direct approaches analyzing genomic DNA along with mRNA are technically limited by the low DNA sequencing coverage that can be achieved at single-cell level, and are consequently hampered in their sensitivity(124,125). This limit can be circumvented using indirect approaches that aim at identifying expressed genomic variants in scRNA-seq data. In this context, short-read, Illumina-based protocols make mutation analysis challenging due to the 3′ or 5′ transcript end-bias and lack of coverage across mutation hotspots(100). Experimental and computational methods are under continuous development to achieve the broadest applicability. Another approach was featured in the seminal paper by Miles et al. and consists in combining scDNA-seq with cell-surface protein expression, which the Authors exploited to characterize CH, MPN and AML patients(68). Furthermore, the T- or B-cell receptor repertoire of individual lymphocytes can be scored in parallel with their gene expression profiles, using properly devised experimental(126) and computational(127) methods on scRNA-seq data, thus providing connections between lymphocyte clonality and functional responses.

Technologies are also available that allow concomitant analyses of protein and transcripts at single-cell levels. These are particularly useful to investigate post-translational regulatory events and to relate functionally-defined phenotypes to protein markers, which might assist tumor classification, biomarker assessment for prognostic purposes, and development of therapeutic targets. Surface proteins can be detected by implementing gene-expression libraries with oligonucleotide- labeled antibodies, as for CITE-seq(128) and REAP-seq(129). Notably, the CITE-seq workflow is compatible with the most frequently used commercial platforms for scRNA-seq, and there's no upper limit to the number of antibodies that can be used. PLAYR, instead, relies on mass spectrometry and allows the detection of up to 40 proteins(130). This technique might be critical when high-quality antibodies are unavailable; also, it can be deployed for index sorting and imaging approaches to enable spatial resolution. Using other techniques, intracellular proteins can be accessed as well with scaling throughput(131). Multiomics applications can potentially address many research questions related to AML heterogeneity and its associated impact. A first important application of combined genomic-phenotypic approaches is the distinction of neoplastic from non-neoplastic cells within tumors, which remains inaccurate when solely based on the expression of specific genes or surface markers, due to the occurrence of technical artifacts in scRNA-seq or aberrant gene expression in either cell-populations. Mapping single-nucleotide variants and/or copy-number variations across phenotypically defined cells can enhance the confidence of such imputation(99). In principle, the acquisition of thousands of unselected cells (e.g., total CD34+ or BM/PB MNCs) would allow the characterization of both neoplastic and non-neoplastic/immune compartments in parallel, to study the functional properties specific to each compartment and clone. Only a few studies have exploited such approaches in myeloid neoplasms, focusing on mapping single mutations. Giustacchini et al. obtained scRNA-seq profiling of BCR-ABL positive vs negative HSC from patients with chronic myeloid leukemia, and found restricted expression in BCR-ABL negative HSC of inflammatory genes with suppressor functions on HSC (i.e., IL6 and its downstream mediators, TGF-β and TNF-α pathways)(132). Another study used transcriptional and mutational single-cell data to feed a machine-learning model for the identification of malignant vs non-malignant AML cells, and found heterogenous malignant cell-types whose abundance correlated with genotypes and survival(133). Beyond the distinction of neoplastic from non-neoplastic cells, multiomics single-cell methods represent promising strategies to dissect mechanisms of therapy resistance and relapse, which involve genetic dynamics of cell clones in parallel with changes in functional and immunomodulatory properties of both tumor and immune cells. Finally, multiomics characterization might assist the identification of biomarkers for predicting disease evolution and response to treatment.

Table 5. Overview of selected single-cell multiomics approaches devised to couple genetic and phenotypic information.

Method	Overview	Throughput (N cells with multiomics characterization)	Features	Limits
Genome + Transcriptome				
G&T-seq(124)	Experimental method Physical separation of RNA and DNA with subsequent parallel amplification and sequencing	- / +	- CNV (direct scoring) - SNV (direct scoring) - Full-length transcriptome (including fusions)	- Low throughput - Low coverage
HoneyBADGER(134)	Computational method Integration of normalized scRNA-seq profiles as compared to: - putative diploid reference of comparable cell type - allelic frequency of heterozygous germline SNP	/	- CNV (inferred from scRNA-seq) - LOH (inferred from scRNA-seq) - Transcriptome	- No information on DNA alterations smaller than 10 megabases - Best performance with scRNA-seq protocols that achieve full-transcript coverage
SCmut(135)	Computational method Variant calling implemented to both scRNA-seq and WES data	/	- Expressed SNV (inferred from scRNA-seq) - Transcriptome	- Relies on quality of the alignment and transcript annotation - Detection sensitivity of a mutation depends on the corresponding gene expression - High rate of false positives and negatives
Van Galen et al.(133)	Experimental method Target amplification of transcript and locus of interest, integration with long-read sequencing	+	- Expressed SNV (inferred from scRNA-seq), insertions, deletions and fusions - Transcriptome	- Depends on expression for mutation detection
Petti et al.(136)	Experimental method Variants scored in WGS and then detected in scRNA-seq data	++	- Expressed SNV (inferred from scRNA-seq), indels - Transcriptome - High-throughput that preserves biological complexity - General applicability	- 5'-end bias - Heavily depends on expression for mutation detection - No clonal reconstruction (wild-type status not defined)
GoT(137)	Experimental method Target amplification and circularization of transcript and locus of interest	++	- Expressed SNV (inferred from scRNA-seq) - Transcriptome - Overcomes end bias by transcripts circularization	- Depends on expression for mutation detection (mitigated by target amplification)
TARGET-seq(138)	Experimental method Release of gDNA and mRNA followed by target amplification	+	- SNV, indels - Transcriptome - Parallel information from coding and non coding DNA - Clonal reconstruction - Low allelic dropout	- End-bias with 'high-troughput' protocol
Genome + Proteins				
Tapestri (Mission Bio, Inc)(68,115)	Experimental method Microfluidic workflow for target amplification of DNA amplicons and proteins	++	- SNV - CNV - Cell-surface proteins - Standardized commercial platform - Customizable gene and antibody panel - Clonal reconstruction at single-cell level - Integrated pipeline for multi-omics analysis	- No information on gene expression and regulatory networks

(Continued)

Method	Overview	Throughput (N cells with multiomics characterization)	Features	Limits
Transcriptome + Proteins				
CITE-seq(128)	Experimental method Antibody-bound oligos act as synthetic transcripts that are captured during most large-scale oligodT-based scRNA-seq library preparation protocols	++	- Transcriptome - Surface proteins - Adaptable to RNA interference assays, CRISPR, and other gene editing techniques. - No upper limit in number of antibodies	- No spatial information - No intracellular proteins
PLAYR(130)	Experimental method Labelling of RNA and proteins with isotope-conjugated probes and antibodies for mass spectrometry detection	+	- Transcriptome - Surface and intracellular proteins	- No spatial information - Limited number of proteins
Transcriptome + T cell receptor				
Tessa(127)	Computational method Bayesian model trained on bulk and scRNA-seq of TCR and T cells	/	- TCR sequences - Transcriptome	- No information on splicing isoforms
RAGE-seq(126)	Experimental method Combined targeted capture and long-read sequencing of full-length transcripts	++	- TCR / BCR sequences - Transcriptome - Splicing isoforms - Accurate antigen receptor sequences at nucleotide resolution - Information on splicing isoforms - Adaptable to any scRNA-seq platform using 3′ or 5′ cell-barcode tagging	- Low recovery of cell barcodes due to low accuracy of long-read sequencing - Possible PCR artifacts
CNV, copy number variation; LOH, loss of heterozygosity; SNP, single nucleotide polymorphism; SNV, single nucleotide variant; WES, whole exome sequencing; WGS, whole genome sequencing; gDNA, genomic DNA; cDNA, coding DNA; PCR, polymerase chain reaction; TCR, T cell receptor; BCR, B cell receptor.				

2. Aim of the project

AML intra-tumor heterogeneity originates from manifold, yet complementary genetic and phenotypic sources that impact leukemogenesis and disease prognosis. Although the genomic and transcriptional landscapes of AMLs are fairly-well characterized as independent traits, functional consequences of specific recurrent genotypes remain poorly elucidated, and, as a consequence, we have poor information on how the sub-clonal structure of AMLs is established, and how cellular and genetic clones evolve during leukemogenesis or upon treatment. A further level of complexity and intra-tumoral heterogeneity is the tumor-associated immune milieu, which plays a critical role in maintaining AML fitness and contribute to treatment resistance. The connections between genetic/phenotypic traits and the immune microenvironment, however, are not fully understood, thus challenging identification of vulnerabilities for the development of novel immunotherapies.

Based on these premises, we see an urgent need to achieve an ecosystem-wide characterization of AMLs (as well other tumors) to capture the genetic and phenotypic landscapes of both malignant and immune populations. In this work, we have developed a novel multiomics approach to integrate gene-mutation and -expression information at single-cell resolution, with sufficiently high throughput to represent cancer biological complexity. In addition to gene expression, we investigated an additional layer of phenotypic heterogeneity, e.g. the repertoire of transcript isoforms and associated alternative splicing patterns, which generate protein diversity and are frequently altered in cancer, possibly affecting immune evasion.

From a technological point of view, we combined the power of two sequencing approaches that allowed simultaneous analyses of mutation and transcript-isoform diversity (long-read sequences) and of transcriptional profiles (short-read sequences) at single-cell level. To showcase our approach, we have applied this methodology to the analysis of AML samples sharing a mutation in a spliceosome factor, with the aim to investigate how phenotypic heterogeneity is related to genetic complexity in both the malignant and immune compartments of a coherent AML subgroup.

3. Materials and methods

3.1 Sample selection and collection

We selected 3 newly diagnosed adult AML patients all carrying mutations of the *SRSF2* spliceosome gene. Mutational status was determined during the diagnostic workflow by targeted DNA sequencing of BM aspirate using the Oncomine™ Myeloid Research Assay (ThermoFisher) and Sophia Myeloid Solution kit (Sophia Genetics SA). Two patients had *de novo* AML (hereafter named AML4 and AML5), while the third suffered of AML secondary to MDS (sAML1). All patients provided written informed consent to study procedures under the IEO *Institutional Review Board* research protocols. BM was collected by posterior iliac spine aspiration before chemotherapy administration; no prior treatment with immunosuppressive agents was reported. Mononuclear cells (MNCs) from BM aspirate were isolated using Ficoll density-gradient centrifugation and cryopreserved in alpha-MEM with dimethylsulfoxide 10% in liquid nitrogen, before further use. The main demographic, clinical and pathological characteristics of patients are summarized in Table 6.

Table 6. Demographic, clinical and pathological characteristics of AML patients.

Sample ID	Age at diagnosis	Sex	Previous MDS	BM blasts (%)	BM dysplastic morphology	Immunophenotype	Karyotype
AML4	75	M	no	80	no	CD34+/CD117+/CD33+/ CD13+	46,XY[20]
AML5	50	M	no	70	no	CD34+/CD117+/TdT+/HLA-DR+/CD38+/CD56+/ CD13+/CD33+/CD4+/ CD15-/+	46,XY[20]
sAML1	67	M	yes	80	yes	CD34+/CD117+/CD13+/ CD33+/CD45/HLA-DR+/ CD38+/CD123+/CD25+/ CD99+/CD45+/CD133+	46,XY[20]

3.2 SCM-seq workflow

To achieve integration of gene expression, mutation and splicing isoforms profilings at single-cell levels, each sample was processed independently according to the corresponding protocols and sequencing technologies. Results were analysed separately and then combined by dedicated computational pipelines. Here, we summarize the overall SCM-seq workflow from sample processing to downstream analyses (Figure 5), while a more detailed description of each step is provided in the following paragraphs.

Single cells were captured through a microfluidic chip using the 10x Chromium system coupled with the 10X Genomics 3′ v3 Single Cell Gene Expression Solution into Gel Beads-in-emulsion (GEMs). Inside the GEMs, poly-adenylated mRNA transcripts undergo tagging with a cellular barcode (CB) and a unique molecular identifier (UMI), followed by reverse transcription (RT) for the production of full-length cDNA. One key advantage of the 10x Chromium system over other platforms that produce full-length cDNA from single cells (e.g., the Fluidigm C1 system) is the higher number of cells that can be recovered from each sequencing run, with consequent increased representation of biological complexity, a feature that is critical for our scopes.

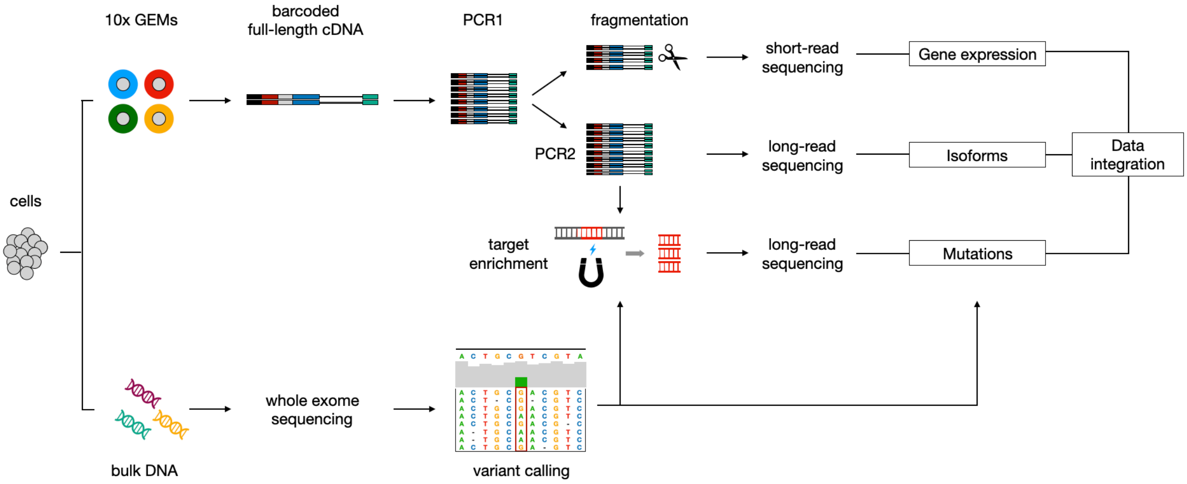
After GEMs are broken, the CB- and UMI-tagged full-length cDNA is amplified by PCR (PCR1) and split in two for its usage in short- (SR) and long-(LR) read sequencing. SR (Illumina) sequencing requires cDNA enzymatic-fragmentation before library construction and yields state-of-the-art single-cell gene expression analysis. LR sequencing, instead, can profile the entire length of cDNA molecules, thus avoiding transcript end-bias and enabling reliable analyses of transcript isoforms and expressed mutations.

As SR and LR sequencing platforms, we used Nanopore and Illumina, respectively. Nanopore is a nucleic acid sequencing approach developed by Oxford Nanopore Technologies (ONT). The Nanopore technology is based on ratcheting a nucleic acid strand through a proteic nanopore in the presence of an ionic current across the pore itself. As the DNA or RNA molecule is threaded through the pore, it alters the ionic flow in a manner that is specific to the the chemical composition of the nucleotides residing within the pore at any given moment. Thanks to a current sensor coupled to each pore, these current alterations are continuously recorded and produce a trace of current (picoAmperes) over time. This signal can then be translated into a nucleic acid sequence by a process called basecalling, using dedicated algorithms.

Prior to LR sequencing of barcoded full-length cDNA, further PCR amplification (PCR2) is applied to obtain enough material for: i*)* LR sequencing (without further processing), to obtain transcriptome-wide profiling of isoforms and associated splicing events, and ii*)* target enrichment of regions carrying sample-specific SNVs (see below), to optimize coverage at sites of interest. During computational downstream analysis, shared CB are used to link SR and LR data, thus enabling the integration of multiomics information.

Sample-specific SNVs are obtained by whole exome sequencing (WES) of bulk-cell DNA from the same samples, and used to: *a)* design a mutation-specific enrichment panel to target regions of interest in barcoded cDNA, and *b)* guide mutation mapping and validation in the LR dataset.

Figure 5. Overview of SCM-seq workflow.
Full-length cDNA tagged with a single-cell CB is obtained and amplified within the Chromium 10x protocol (PCR1). The cDNA pool is split to undergo both short-read sequencing (after fragmentation) and long-read sequencing (after further amplification, PCR2). These steps allow to obtain gene expression and isoform analysis, respectively. In parallel, bulk WES analysis is performed to score high-confidence somatic variants, which are enriched before long-read sequencing for single-cell mutation analysis. Afterwards, the three levels of information are computationally integrated using shared CB from the cDNA pool.



3.3 Single-cell RNA 10x library preparation and sequencing

Cryopreserved BM-MNCs for each sample were thawed at 37°C, subjected to FACSMelody-sorting after DAPI-staining to purify viable cells and subsequently washed twice with PBS + 0.04% BSA. Final concentration of viable cells was determined using an average of two manual counts. For scRNA-seq library preparation, we used the 10x Chromium system and the 10X Genomics 3′ v3 Single Cell Gene Expression Solution following all manufacturer's recommendations. Briefly, an appropriate volume of cell suspension was loaded onto each channels of the 10x Chromium Single Cell Chip to recover approximately 5,000 cells and single cells were then captured in GEMs. Captured polyA-mRNAs were reverse transcribed into cDNAs and amplified to generate amplified full-length cDNA tagged with CB and UMI (PCR1). The quality of cDNA was checked using the 2100 Bioanalyzer (Agilent). About 25% of the prepared cDNA volume was used for subsequent library preparation (as *per* protocol's instructions), while the remaining was stored at -20ºC until further processing. Final libraries were constructed by fragmentation, adapter ligation and a sample index PCR. Library size was analysed by the 2100 Bioanalyzer (Agilent). Sequencing was performed on an Illumina NovaSeq 6000 Sequencing System machine (paired-end 151 bp reads, ~50,000–70,000 reads per cell).

3.4 Single-cell transcriptional analysis

3.4.1 Data pre-processing

For each sample separately, the binary base call files obtained from the Illumina sequencing machine were demultiplexed using the 10x Genomics Cell Ranger (v6.0) mkfastq pipeline to obtain two FASTQ files for each sample for each lane. These files were used as input for the Cell Ranger count pipeline, based on STAR aligner, to align sequencing reads to the GRCh38 reference transcriptome and to quantify transcript levels of each gene in each cell. The pipeline returns feature-barcode matrices with the number of UMIs associated with a feature (gene, row) and a barcode (cell, column).

3.4.2 Generation of high-quality gene-cell matrix

We performed cell and gene quality-control (QC) on the obtained raw count matrices using a custom pipeline in an R environment, again on a *per*-sample basis. Namely, putative cell doublets were filtered out using the scDblFinder method(134). An adaptive threshold was applied to discard cells based on their QC metrics, based on numbers of UMIs and genes *per* cell within 3-5 median absolute deviations around the median. We also discarded cells with percentages of mitochondrial genes >20% and genes expressed by <3 cells. From these steps, we retrieved: a) a list of high-quality CB to be used for matching SR and LR data, and b) a high-quality gene expression matrix for each sample, to be used in the following procedures.

3.4.3 Gene-cell expression matrices normalization, visualization and integration

All scRNA-seq data analyses were carried out using Seurat v4.0, unless otherwise specified(120,135). For each sample separately, we initialized a Seurat object from the raw filtered gene-cell expression matrix using the command *CreateSeuratObject.* The count matrices were normalized on a *per*-sample basis following the *SCTransform* method including the percentage of mitochondrial genes in the regression model as potential source of technical variation and accounting for 3,000 hypervariable features. Such normalization method was selected to maximize variance stabilization of scRNA-seq data, as SCTransform outperforms log-normalization in dealing with technical variability including different sequencing depth across samples. Thereafter, we performed linear dimensionality reduction with principal component analysis (PCA) followed by non-linear dimensionality reduction with the Uniform Manifold Approximation and Projection (UMAP) algorithm on the first 50 principal components, which allowed to visualize the high-dimensional scRNA dataset in a bidimensional space. To correct for batch effect and identify shared cell populations, we performed data integration of the SCT-normalized objects according to the Seurat v4 method, using the top 3,000 hypervariable features across datasets as ranked by the function *SelectIntegrationFeatures*. After data integration, we performed again PCA and UMAP on the first 50 principal components to jointly visualize and analyze the three samples.

3.4.4 Cell lineage imputation

To classify cells based on their lineage identity, we mapped each of our query samples (unselected BM-MNCs from AML patients) to a reference dataset of 30,672 BM-MNCs from a healthy human donor (GSE128639(119); hereafter named 'BM reference'), obtained by CITE-seq(128), a protocol that uses oligonucleotide-conjugated lineage-specific antibodies to identify terminally differentiated cells together with scRNAseq data to classify individual cells across the full spectrum of hematopoietic differentiation. A more refined lineage annotation for the same dataset has been provided by Hao et al., who developed a framework to weight the combination of RNA and protein data and assess subtler levels of heterogeneity through a weighted nearest neighbor (WNN) graph(120). To map our AML samples to this multimodal BM reference, we obtained the corresponding RNA-cell expression matrix with cell lineage annotation and associated WNN graph (provided as 'bmcite' object in the SeuratData package) and followed the approach illustrated in Hao et al.

First, as the BM reference count matrix was log-normalized, we log-normalized our query data according to Seurat default settings (i.e., divided individual RNA counts by the total count of RNA in the cell, multiplied by a scale factor of 10,000 and log-transformed), using the command *NormalizeData.* On the reference BM, we computed a supervised principal component analysis (sPCA) via *RunSPCA*, aiming to identify the transformation of the RNA data that best encapsulates the structure of the WNN graph. This step allows a weighted combination of the protein and RNA measurements to 'supervise' the PCA and highlights the most relevant sources of variation in the dataset. We then computed the first 50 neighbors in the sPCA space of the BM reference, to guide finding anchors between each query dataset and the multimodal reference. Finally, based on integration anchors and using the command *MapQuery*, each query AML was mapped onto the WNN coordinates of the BM reference, while cell lineage labels were transferred accordingly. Mapping visualization was obtained through the *RunUMAP* command (with the argument 'return.model' set to TRUE to enable projection of the query datasets onto the BM reference). Cell lineage labels were transferred from the log-normalized AML datasets back to the SCT-normalized object and stored in the metadata. To check the accuracy of lineage imputation, we produced a prediction score for label transfer as elaborated by Stuart et al. and implemented in the *TransferData* function in Seurat. Briefly, lineage label predictions were computed by multiplying the anchor classification matrix (which contains the classification information for each anchor cell in the reference dataset) with the transpose of the weights matrix (which defines the strength of association between each query cell and each anchor). This returns a prediction score for each lineage for every cell in the query dataset that ranges from 0 to 1, and sums to 1.

For the orthogonal validation of cell lineage imputation, we used a set of lineage-specific signatures released from the Human Cell Atlas and derived from an independent dataset of over 100,000 BM cells spanning 8 healthy donors (4 females and 4 males; age 26-52 years), which allows to account for sex, age and donor-related variation. Of note, these signatures include both discrete and transitioning states associated to early progenitors and committed precursors(123).

3.4.5 Cell cycle phase imputation

Cell cycle phase assignments were generated for each cell from scores derived using the *CellCycleScoring* function within Seurat, as originally described in Tirosh et al.(136) and based on the following list of genes (Table 7).

Table 7. List of genes used for cell cycle phase imputation.

Cell cycle phase	Genes				
S	PCNA	UHRF1	POLD3	MSH2	EXO1
	TYMS	MLF1IP	SLBP	ATAD2	TIPIN
	FEN1	HELLS	POLA1	RAD51	DSCC1
	MCM2	RFC2	CHAF1B	RRM2	BLM
	MCM4	E2F8	BRIP1	CDC45	CASP8AP2
	RRM1	CCNE2	CLSPN	CDC6	USP1
	UNG	DTL	UBR7	RAD51AP1	RPA2
	GINS2	PRIM1	WDR76	GMNN	NASP
	MCM6	MCM5	CDCA7		
G2M	TOP2A	HMGB2	CDCA2	UBE2C	CENPE
	NDC80	CDC25C	CDCA8	CDK1	CTCF
	CKS2	KIF2C	ECT2	HMMR	NEK2
	NUF2	RANGAP1	KIF23	AURKA	G2E3
	CKS1B	NCAPD2	CDCA3	PSRC1	GAS2L3
	MKI67	DLGAP5	HN1	ANLN	CBX5
	TMPO	CKAP2L	CDC20	LBR	CENPA
	CENPF	CKAP2	TTK	CKAP5	GTSE1
	TACC3	AURKB	KIF11	TUBB4B	KIF20B
	FAM64A	BUB1	ANP32E	NUSAP1	HJURP
	SMC4	TPX2	BIRC5	CCNB2	

3.4.6 Additional transcriptional analyses

Marker genes of cell groups of interest (i.e., overexpressed genes) were derived by Wilcoxon test using the *FindAllMarkers* function in Seurat, discarding genes detected in less than 10% cells of either of the two considered populations and applying Bonferroni correction to the computed p-values to obtain False Discovery Rate (FDR) values.

Single-cell average expression levels of signatures of interest were calculated using the *AddModuleScore* function in Seurat.

Over-represented pathways were investigated by computing the overlap between genes of interest and terms of the Biological Process Gene Ontology, using the clusterProfiler tool(137). FDR values were obtained by correcting the computed p-values with the Benjamini-Hochberg method and selected pathways enriched with 0.05 p-value and 0.05

FDR. To simplify the redundancy of the analysis while still preserving biological diversity, we applied semantic similarity reduction with the GOSemSim package(138), using a cut-off of 0.7%.

To assess and quantify the diversity of phenotypes between cell groups of interest, we calculated a metric of phenotypic volume (PV) as proposed by Azizi et al(139). PV of a given population can be defined as the pseudo-determinant of the gene expression covariance matrix for that population, which considers covariance between all gene pairs in addition to their variance. Higher PV in one cell group as compared to another shows the independency of active phenotypes, suggesting the activation of additional mechanisms and pathways. We computed PVs for each cell group of interest and, to correct for the effect of differences in the number of cells across groups, we sampled 80% of cells from each group with replacement and computed the empirical covariance between genes, based on imputed expression values for that subset of cells. This was followed by singular-value decomposition of each empirical covariance matrix and computation of the product of nonzero eigenvalues. PVs were also normalized by the total number of genes. To increase robustness, this process was repeated 10 times to achieve a range of computed PVs for each group. Finally, we represented this range with violin and boxplots and calculated the statistical significance of changes between groups by Mann-Whitney U-test.

3.5 Bulk analysis of somatic mutations

3.5.1 Fibroblast isolation

As the availability of germline tissue is essential to reliable calling of somatic tumor variants, we isolated fibroblasts (i.e., non-hematopoietic cells) from BM-MNCs of each patient. After thawing with standard procedures, 2.5x10^6 BM-MNCs were plated in 2 wells of a 6-well plate with alpha-MEM, 2 mM L-glutamine, 100 U/ml penicillin/streptomycin and 20% Fetal Bovine Serum; fresh medium was supplied on an every-other-day basis and cell cultures were monitored for the appearance of adherent, spindle-shaped cells. Such cells were let expand to 70% confluence, re-plated and collected at second passage for DNA extraction(140).

3.5.2 DNA extraction and quality control

Bulk genomic DNA from BM-MNCs and fibroblasts was extracted using a DNeasy Blood & Tissue Kit (QIAGEN) as *per* manufacturer's instructions. Purity was checked by NanoDrop while quantity was assessed by a Qubit fluorometer using the High Sensitivity dsDNA Assay Kit (Thermo Fisher Scientific).

3.5.3 Whole exome library preparation and sequencing

Libraries for WES were prepared following manufacturer's protocol using SureSelectXT Low Input Reagent Kits (Agilent Technologies) for dual-indexing, target enrichment and capture. Paired-end sequencing (151 bp) was performed on an Illumina NovaSeq 6000 Sequencing System machine.

3.5.4 Somatic variant calling by WES

After alignment against human reference genome hg38, we proceeded to the identification and marking of duplicate reads using the Picard suite (http://broadinstitute.github.io/picard/) and base quality score recalibration with the GATK tools v4.1.2.0-1, following best practices declared by the Broad Institute(141).

Variant calling was performed with MuTect2 from the GATK toolkit against sample-matched germline-sequence. MuTect2 is a widely used algorithm in WES analyses to call somatic mutations, particularly single nucleotide variants (SNV) and small insertions and deletions (INDELs). The tool uses a Bayesian somatic genotyping model to call short somatic mutations via local assembly of haplotypes. To improve the final quality of called variants, in addition to standard quality filters, we discarded all variants in sites with sequencing depth inferior to 10 total reads and variants called with less than 3 reads supporting the alternative allele in the cancer sample. Functional annotation of variants passing the described filters was performed by Funcotator (FUNCtional annOTATOR), a dedicated tool from the GATK suite, using information from the Cancer Gene Census, ClinVar, the Catalogue of Somatic Mutations in Cancer, dbSNP, GENCODE, the Genome Aggregation Database, the HUGO Gene Nomenclature Committee, the Open Regulatory Annotation Database and UniProt.

3.6 Single-cell target enrichment of somatic mutations

3.6.1 Re-amplification of barcoded full-length cDNA (PCR2)

To obtain enough product for target enrichment and ONT sequencing, we performed further amplification of PCR1 product (full-length cDNA tagged with CB and UMI from the Chromium workflow). To maintain the representation of CB and UMI and minimize amplification biases, we used the same primer mix employed in the 10x Genomics protocol and a high-fidelity KAPA DNA Polymerase (Roche) following settings shown in Tables 8. For each reaction, we used 3-15 ng cDNA as starting input, adjusting with nuclease-free water to a final volume of 25 µL.

Amplified cDNA was purified with 0.6x AMPure XP beads and eluted in 15 µL nuclease-free water. The final product was then quantified with a Qubit fluorometer using the High Sensitivity dsDNA Assay Kit (Thermo Fisher Scientific), while fragment size distribution was analyzed by a Bioanalyzer system High Sensitivity DNA Assay on a 2100 Bioanalyzer instrument (Agilent). At least 3 reactions *per* sample were prepared (each yielding 50-150 ng/uL cDNA, depending on the sample) and pooled as needed for subsequent use.

Table 8. Thermal cycler setting for PCR2.

Segment Number	Reaction	Number of Cycles	Temperature	Time
1	Initial denaturation	1	95°C	3 min
2	Denaturation	8	98°C	30 s
	Annealing		64°C	30 s
	Extension		72°C	5 min
3	Final extension	1	72°C	10 min
4	Cooling	1	4°C	hold

3.6.2 Design of mutation-specific enrichment panel

For the design of a mutation-specific target enrichment panel, following WES analysis we selected 11, 15 and 16 non-synonymous, coding gene mutations (out of a total of 23, 22 and 24 genes for the three samples) with at least 0.1 variant allelic frequency (VAF) (Table 15). RNA capture libraries were built using the SureSelect DNA Advanced Design Wizard software (Agilent) according to the human reference genome hg38. Specifically, the panel consists of 4,005 probes (average size 423 bp) covering 41 amplicons, which correspond to the genomic positions of selected variants and span 17.038 kbp (Table 9).

Table 9. Genomic positions of mutation-specific target amplicons.

Gene	Chromosome	Start	End
CSF3R	chr1	36466443	36466851
MACF1	chr1	39337051	39337459
WDR63	chr1	85081045	85081453
TTF2	chr1	117097213	117097621
CACNA1E	chr1	181757810	181758218
OR2T8	chr1	247921073	247921481
DNMT3A	chr2	25244282	25244780
IDH1	chr2	208248185	208248593
ALPP	chr2	232378616	232379214
SETD2	chr3	47120063	47120471
FRYL	chr4	48510675	48511083
SPATA18	chr4	52071897	52072305
CCDC158	chr4	76369281	76369689
AADAT	chr4	170068989	170069397
ADAMTS16	chr5	5181849	5182257
PCDHB7	chr5	141174687	141175095
SLC35B2	chr6	44255396	44255804
ZAN	chr7	100763685	100764093
DGKI	chr7	137469397	137469805
EZH2	chr7	148826295	148826703
CNTNAP3	chr9	39078642	39079050
ABL1	chr9	130884766	130885174
MALRD1	chr10	19595058	19595466
PDZD7	chr10	101023739	101024147
USH1C	chr11	17530976	17531384
HDAC7	chr12	47796911	47797409
FLT3	chr13	28018301	28018709
ADAMTS7	chr15	78774555	78774963
SRSF2	chr17	76736673	76737081
HRH4	chr18	24476600	24477008
ASXL3	chr18	33744910	33745318
PCSK4	chr19	1483749	1484259
CEBPA	chr19	33301250	33301658
PCSK2	chr20	17481565	17481973
ASXL1	chr20	32434434	32434842
RUNX1	chr21	34880372	34880900
CXorf21	chrX	30560052	30560460
MED12	chrX	71137042	71137450
STAG2	chrX	124037362	124037770

…continued

3.6.3 Enrichment library preparation

Enrichment libraries were prepared following the steps described in the SureSelectXT HS Target Enrichment System protocol (Agilent Technologies), and summarized below.

1. Hybridization of cDNA samples to the capture library

As starting input for hybridization to the capture library, we used the maximum possible amount (800-1000 ng) of amplified full-length cDNA tagged with CB from PCR2. In the initial segments of the reaction (Table 10, segment 1-2), proprietary blocking oligonucleotides were added (SureSelect XT HS and XT Low Input Blocker Mix) to increase capture specificity and reduce nonspecific hybridization of repetitive elements. A capture library hybridization mix was prepared adding 25% RNase Block solution and SureSelect Fast Hybridization Buffer to the capture library, with proportions that are based on the library size (<3 Mb in our case). During capture (segments 3-5), the cDNA is denatured and hybridized to RNA probes.

Table 10. Thermal cycler setting for cDNA hybridization.

Segment Number	Number of Cycles	Temperature	Time
1	1	95°C	5 minutes
2	1	65°C	10 minutes
3	1	65°C	1 minute
4	60	65°C	1 minute
		37°C	3 seconds
5	1	65°C	Hold

2. Capture of the hybridized cDNA by streptavidin-coated beads

As RNA probes contain a biotin molecule, the hybridized cDNA can be captured by magnetic beads coated in Streptavidin, which acts as a receptor to the biotin molecules (Dynabeads MyOne Streptavidin T1, ThermoFisher Scientific). After washing, the beads were added to the hybridized cDNA and mixed vigorously (at 1600 rpm) for 30 minutes. The beads attached to target cDNA were then collected through a magnetic separator and further washed with proprietary wash buffer at 70°C. This step is critical to ensure specificity of the capture and to discard non-target fragments which are still in solution. After washing, the beads are resuspended in nuclease-free water.

3. Post-capture amplification and purification

Post-capture PCR amplification of the captured cDNA is necessary to achieve a library molarity sufficient for sequencing. The amplification mix was added to the captured DNA, when still retained on the streptavidin beads. Reactions were carried out as for PCR2, but for the number of denaturation-annealing-extension cycles, which were instead set based on the size of the capture library following manufacturer’s recommendations, i.e., 14 cycles for our mutation-specific panel (library size < 0.2 Mb). Finally, the amplified captured libraries were purified using x1 AMPure XP beads, eluted in nuclease-free water and kept at -20°C until further use. We repeated steps 1 to 3 to ensure capture specificity given the small size of probes.

4. Quantity and quality assessment

The final enrichment product was quantified with a Qubit fluorometer using the High Sensitivity dsDNA Assay Kit (Thermo Fisher Scientific), while fragment size distribution was analyzed by a Bioanalyzer system High Sensitivity DNA Assay on a 2100 Bioanalyzer instrument (Agilent). At least 3 reactions *per* sample were prepared and pooled as needed for subsequent use.

3.7 Single-cell cDNA ONT library preparation, sequencing, basecalling and data pre-processing

For both whole-transcriptome and mutation-targeted full-length cDNA, ONT sequencing libraries were prepared using the SQK-DCS109 protocol (PCR-free) for direct cDNA sequencing (Oxford Nanopore Technologies), following manufacturer's modification, and sequence on a GridION device (the whole-transcriptome library) or a PromethION device (the mutation-enriched library), the latter aimed at obtaining higher throughput for mutation analysis.

We started each preparation with 100-200 ng cDNA in 20 µl nuclease-free water and performed end repair and dA-tailing by incubating samples at 20°C for 5 minutes and 65°C for 5 minutes in a thermal cycler. Subsequently, we purified the product with x1 AMPure XP beads and 70% EtOH, eluted in 30 µl nuclease-free water and performed ligation of the sequencing adapters with a proprietary adapter mix and a Blunt/TA Ligation Master Mix (New England Biolabs) and incubating the reaction for 10 minutes at room temperature. The adapted and tethered cDNA library underwent further purification by x0.4 AMPure XP and a proprietary wash buffer. The final elution was performed in 13 µl and 25 µl of proprietary elution buffer for samples to sequence on GridION and PromethION flowcells, respectively. As the flowcell's pore occupancy could be compromised when loading low amounts of cDNA, we quantified libraries with a Qubit fluorometer using the High Sensitivity dsDNA Assay Kit (Thermo Fisher Scientific), calculated molarity by the online tool Promega Biomath Calculator (https://www.promega.com/resources/tools/biomath/?calc=molarity) choosing the option dsDNA: µg to pmol and adjusted concentration to obtain 5–50 fmol (GridION) or ~60 fmol (PromethION) of library. In some cases, we stored the libraries at -20ºC before carrying on the sequencing (no later than 15 days). To set and control the sequencing device and acquire data we relied on the MinKNOW software (Oxford Nanopore Technologies). Prior to priming the flowcell and starting sequencing, for each flowcell we checked the number of active pores as suggested in the manufacturer's instructions (as a minimum, this should test around 800 active pores for GridION flowcells and 5000 for PromethION flowcells). For all samples, sequencing was carried on for 72 hours. Basecalling was performed with the proprietary software Guppy, a neural network basecaller that translates changes in the raw electric signal into nucleobases; this was set on high-accuracy basecalling mode and run with versions from 3.2.10 to 5.1.13 across subsequent experiments (GridION) or 6.1.5 (PromethION).

Basecalled reads were analyzed with pycoQC(142) v2.5.0.21 to generate and inspect QC metrics. We evaluated the distribution of read counts, read length, N50 (a parameter that represents the length of the shortest read in the group of longest sequences that together represent ≥50% of the nucleotides in the set of sequences) and read PHRED quality score (which indicates the probability of the base being called correctly, ranging from 0 to

60). For downstream analyses, we filtered reads with PHRED score ≥7 (whole-transcriptome dataset) and ≥10 (mutation-enriched dataset).

3.8 Identification of 10x cellular barcodes in ONT data and generation of ONT gene-cell count matrix

To perform an integrated analysis of transcriptional, mutational and splicing features on a single-cell basis, we first need to establish a common set of CB that matches any given short-read 10x experiment with the corresponding long-read ONT experiment; in this way, information obtained from 10x (i.e., transcriptional analysis) can be combined to that investigated by ONT (i.e., mutation and isoform analysis). For this and further following steps, we exploited many of the functions available from FLAMES (Full-Length Analysis of Mutations and Splicing), a recently published computational framework developed for the analysis of full-length transcriptome at both bulk and single cell level, which also provides a workflow specifically devised for isoform analysis on LR data(143).

The 10x read template (Figure 6) consists of poly-adenylated fragmented transcripts attached to an Illumina adapter, CB and UMI of fixed length, in a fixed sequence. The ONT read template (Figure 7) is essentially the same as the 10x read template but for the addition of the ONT sequencing adapter and length of the transcript.

Figure 6. Structure of 10x short-read templates.

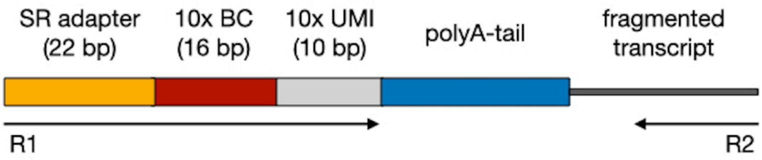
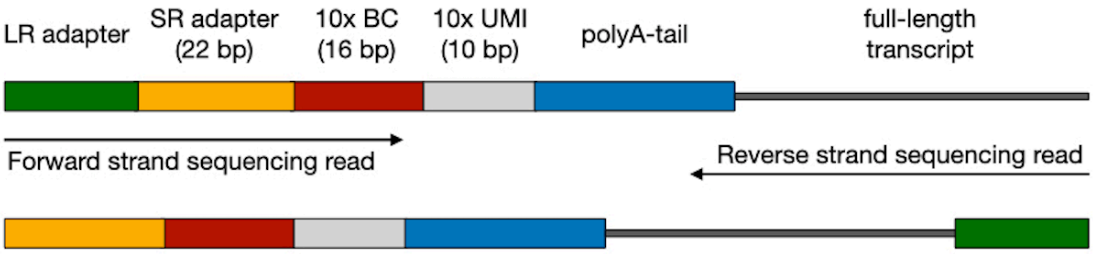


Figure 7. Structure of ONT long-read templates from SCM-seq.



To extract ONT reads that match reference CB from 10x data, we run the FLAMES *match_cell_barcode* function using as input the raw single-end ONT FASTQ files and the list of CB obtained from scRNA-seq 10x data after QC and filtering. More in detail, the program exploits the fixed structure of template reads to identify CB locations in the ONT FASTQ file, by searching for the flanking sequence before the CB; in this process, we accepted up to 1 mismatch in the base sequence. Reads that failed to match any CB were discarded, while the CB, UMI and polyA stretch were trimmed and sequences after the polyA tail were kept. Finally, the CB and UMI sequences were integrated into the FASTQ read header to associate a combination of CB and UMI to each read name using the scPipe package, a Bioconductor package for scRNA data handling(144).

For each AML sample, we performed this step in parallel on each of the FASTQ files obtained from each of the ONT flowcells (4, 6, and 4 *per* each sample for GridION experiments, and 1 flowcell *per* sample for PromethION experiments); afterwards, the processed FASTQ files belonging to the same sample were merged into a single one, which was used as input for genome alignment and further analyses.

All the reads resulting from CB assignment were aligned to the reference genome (GRCh38) using minimap2; from this step, we obtained an Unsorted Binary Alignment Map (BAM) file of aligned reads, which was imported to R to associate reads to exonic, intronic or intergenic features using the Gencode v24 annotation in GTF format. For this step we employed scPipe to finally create the canonical barcodes-by-genes single-cell matrix. Here in particular, we applied the *sc_exon_mapping* function specifying CB and UMI length. As output, we obtained a BAM file with additional tags storing information about mapping status, gene identity, CB and UMI. The final gene count matrix (n (cells) × m (genes)) was generated using the *sc_count_aligned_bam* function from the scPipe package, considering only reads mapping to exonic regions.

3.9 Mutation and genotype analysis on ONT data

3.9.1 Assessment of mutation-specific target enrichment performance and coverage

We first assessed the performance of mutation-specific target enrichment in terms of capture specificity, coverage and uniformity across samples. We exploited the Bioconductor package TEQC since it provides functionalities specifically devised to this aim(145). For each sample, we input a bed file containing genomic positions (chromosome, start, end) of the targeted regions and a BAM file containing genomic positions of sequenced reads aligned to the reference genome.

Capture specificity was defined as the fraction of aligned reads that overlap with any target region, calculated for each sample by the function *fraction.reads.target* (with Offset option set to 0) and graphically represented by barplots.

Coverage was calculated as read coverage for each base that is sequenced and/or located in a target region by the function *coverage.target* (perTarget and perBase options set to TRUE) and graphically represented by histograms (function *coverage.hist*). To analyse reproducibility across samples, we also calculated the *normalized coverage*, defined as the *per*-base coverages divided by the average coverage over all target bases, by using the function *coverage.uniformity*. Coverage normalization is independent of absolute numbers of sequenced reads, thus allowing better comparisons between different samples or different experiments.

3.9.2 Analysis of expressed mutation at single-cell level

To identify reads overlapping with the position of any gene variant of interest, we first extracted start and end position of each mutation by exploiting the information provided by WES analysis. We stored the genomic coordinates in a GRanges object (GenomicRanges v1.42.0) and retrieved ONT reads from the previously produced BAM file, using the *readGAlignments* function from GenomicAlignments v1.26.0, together with a ScanBamParam object for the filtering. With the *sequenceLayer* function from GenomicAlignments we laid read sequences stored in the GAlignments object alongside the reference sequence from BSgenome.Hsapiens.UCSC.hg38 (v1.4.3), using the supplied CIGARs to ensure that each nucleotide in our query was associated with the correct position on the reference sequence. After that, for each read in the filtered BAM, we evaluated the associated base quality (BQ, i.e. the PHRED-scaled quality score, in the range x-y) and mapping quality (MQ, i.e. the probability of the read being incorrectly aligned, in the range 0-40). For each gene and variant of interest, we filtered reads covering the region with BQ ≥10 and MQ ≥ 20. Based on WES information for each gene mutation of interest (position, reference nucleotide and nucleotide change), we extracted reads containing: i*)* the reference nucleotide at the position of the variant (e.g. the wild-type sequence); ii*)* the expected nucleotide change at the same position (e.g. the mutated sequence); or iii*)* neither of the previous at the position of the variant, likely due to sequencing errors (i.e. mismatch reads). Finally, we demultiplexed all classified reads by CB information stored in the read identity, to obtain single-cell information of expressed mutations. CB were selected for genotype imputation if covered by at least 1 mutated or wild-type read in the ONT dataset regardless of the expression level in the 10x dataset, given the possibility of technical dropouts in the latter. In this step, we discarded variants found in <5 cells.

3.9.3 Imputation of genotype at single-cell level

A key challenge in analyzing ONT data derives from the relatively-low accuracy of the official ONT basecaller Guppy (~90%)(146), thus challenging the accuracy of our mutated or wild-type read assignment. Moreover, in most high-throughput scRNA-seq datasets, median gene-expression is represented in the median cell by only one transcript read, a situation that worsens the impact of sequencing errors (as shallow coverage precludes establishing a trustworthy consensus) and leads to the so-called ‘allelic dropout’ phenomenon(147). In particular, in the presence of a single wild-type read *per* cell, it is impossible to impute confidently the genotype, since in the case of heterozygous mutations there is in principle an equal probability to score a wild-type or a mutated transcript. This does not apply in the presence of a single mutated read *per* cell, which instead allows, in principle, to impute the mutated genotype, even though no information is provided on the zygosity of the mutation. Regardless, we have to consider the presence of mismatch reads in the same cell as a further source of technical noise. In conclusion, single-cell genotype cannot be reliably assigned as solely based on the classification of transcript reads as mutated or wild-type. Therefore, we designed two distinct algorithms to determine the confidence of imputing each cell’s mutated or wild-type genotype, respectively, both based on the same available set of mutated, wild-type and mismatch ONT reads.

Premise for the two algorithms

The purpose of the following algorithm is to classify cells as mutated or wild-type based on the available set of ONT reads. Given a set of cells $C = \{c_1, \ldots, c_n\}$ and a certain number of ONT sequencing reads r_{c_i} available for each cell c_i, we are interested in assigning a label L to each element of C, where $L \in \{\text{mutated, wild} - \text{type}\}$ for a given gene mutation known from WES analysis. Note that, r_c might include mutated, wild-type or mismatch reads (the latter due to sequencing error rate). For any SNV we are interested in, let us assume that: i) the mutated base is called B_M and the wild-type base is called B_W (i.e., the ground truth known from WES data); ii) any read at the known position of the variant as called during ONT sequencing is called B (which is possibly affected by errors, as in the case of mismatch reads) and the real base is called R (whose identity is unknown, due to the error probability). Assuming that the accuracy is independent of the specific base, for any base $X \in \{A, C, G, T\}$ we define the basecalling accuracy as $\alpha = \mathbb{P}(B = X \,|\, R = X)$. Furthermore, we have to take two different approaches to classify cells as mutated or wild-type, because a cell can be defined as wild-type only in the case no mutated reads are scored and also due to the aforementioned allelic dropout phenomenon. As such, we shall present the two strategies separately.

1. Confidence of imputing a cell's mutated genotype

We define a cell c as candidate to be $mutated$ if al least one of its reads r_c presents a mutation. For each cell c such that at least one of its r_c basecalled reads presents a mutation, we have to assign a confidence score that this is indeed the case.

Using the above notations and definitions, we compute the probability of wrongly labelling the cell based on the basecalling accuracy α in that specific cell. Namely, if a cell c contains mutated reads $r_m \leq r_c$ that were predicted to contain a mutation, then the likelihood that all of such r_m mutated reads were wrongly basecalled is given by

(1) $$l_c = mutated(1-\alpha)^{r_m}.$$

As a consequence, the higher is the quantity in (1), the lower is our confidence that the cell c is indeed mutated. Therefore, a meaningful score that quantitatively estimates our confidence in assigning the $mutated$ label to cell c can be defined as

(2) $$C_M(c) = 1 - l_c.$$

2. Confidence of imputing a cell's wild-type genotype

We define a candidate cell c as $wild-type$ if all of its r_c reads do not show a mutation. As before, for a cell c that does not present mutated reads, we have to give a confidence score of labelling it as $wild-type$ by quantitatively assessing the likelihood of basecalling errors.

We reasoned that the likelihood of basecalling errors in wild-type cells should be a function of the accuracy of basecalling α (as before) and of the least number of errors needed to preserve the allelic fraction of the SNV of interest. Namely, for a cell c with r_c reads in the absence of mismatches, r_m of which are mutated (we assume $r_m \leq 1$), let $AF(c) = r_m/r_c$ be the allelic fraction of the SNV we are interested in, i.e. the ratio between mutated and wild-type reads.

Given the set C of cells, let $C' \subset C$ be the subset of cells with at least one mutated read. Moreover, let $\mathscr{D}_{AF}$ be the empirical distribution of the allelic fraction AF on C'. If cell c has r_c reads, the number e of possible basecalling errors can be any number in $\{1,\ldots,r_c\}$. The probability of having e errors can be modelled with a binomial distribution of parameters α and r_c, i.e. if E_c is the number of errors for cell c, one has that

(3) $$\mathbb{P}(E_c = e) = \binom{r_c}{e}\alpha^e(1-\alpha)^{r_c-e}.$$

Now, we turn to consider the allelic fraction distribution $\mathscr{D}_{AF}$: we partition the AF values into modality clusters through the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm, which divides the values into partitions $P_1, \ldots, P_m$, each of which will have its own mean and standard deviation, which we denote with $\mu_1, \ldots, \mu_m$ and $\sigma_1, \ldots, \sigma_m$ respectively. We also weight each partition P_i with $w(P_i) = |P_i| / |C'|$ (the relative frequency of cluster P_i). Now, given a cell c that has only $wild-type$ reads, for each partition P_i we compute the least number of errors e_i that would have had to occur for the corresponding allelic fraction $AF(\tilde{c}_{e_i})$ to be within one standard deviation of the mean μ_i, i.e. we compute the smallest number e_i such that the corresponding cell $\tilde{c}_e$, which is the cell c where we switched e_i reads from wild-type to mutated, satisfies

(4) $$|AF(\tilde{c}_{e_i}) - \mu_i| \leq \sigma_i .$$

Therefore, for a cell c we define the likelihood of wrongly predicting it as wild-type as

$$l_c = \sum_{i=1}^{m} w(P_i)\mathbb{P}(E_c = e_i) = \sum_{i=1}^{m} w(P_i)\binom{r_c}{e_i}\alpha^{e_i}(1-\alpha)^{r_c - e_i} .$$

Finally, similarly to the previous section, we define the confidence of labelling as wild-type a cell c with only wild-type reads as

(5) $$C_W(c) = 1 - l_c .$$

We applied the above algorithm to variants with at least one mutated and wild-type read in ≥10 cells, and filtered cells with accuracy ≥0.90 and confidence ≥0.75.

3.10 Transcript isoform analysis on ONT data

After CB assignment and alignment to the reference genome with minimap2, ONT reads resulting from whole-transcriptome experiments were used as input for the FLAMES function *sc_long_pipeline* to polish the alignment and group reads with similar splice junctions to get a raw isoform annotation. This was then compared against the reference gene annotation (Gencode v24) to correct potential splice site and transcript start/end errors. In particular, transcripts with splice junctions or transcript start/end similar to the reference transcripts were merged with the reference, including isoforms that are likely to be truncated transcripts. Next, this updated transcript assembly was used to realign all reads and perform specific isoform quantification. To increase the robustness of the analysis, we filtered transcripts supported by at least 10 reads and covering at least 75%

of the reference transcript they were aligned to. Finally, we generated an isoform-cell count matrix to obtain single-cell isoform characterization. Isoform structural analysis was performed with the tool SQANTI3 (Structural and Quality Annotation of Novel Transcript Isoforms)(148). SQANTI3 classifies long-read transcripts according to their splice junctions and donor and acceptor sites. For the purpose of our analyses, we focused on full-splice-matched isoforms (i.e., isoforms matching a reference transcript at all splice junctions) and novel isoforms (i.e., isoforms containing new combinations of already annotated splice junctions, novel splice junctions formed from already annotated donors and acceptors or isoforms using novel donors and/or acceptors); all other categories were grouped under the label 'other'. SQANTI3 also allowed us to perform isoform functional annotation by supplying the --isoAnnotLite flag.

3.11 Statistical analyses

Unless otherwise specified, all analyses were performed in R v4.2.0. The usage of parametric or non-parametric statistical tests was based upon results of the Shapiro-Wilk's test for normality distribution. Comparisons of proportions between categorical variables were evaluated by Chi-square (or Fisher exact test if the latter was not applicable). Comparison between continuous variables was performed by two-sided Mann-Whitney U test. Correlation analyses were performed by Pearson's or Spearman's method.

4. Results

4.1. SCM-seq reliably allows the integration of matched 10x and ONT datasets

The aim of the SCM-seq protocol is to integrate gene expression, mutation and splicing-isoforms profilings at single-cell levels, combining short-(10x) and long-(ONT) read technologies. The experimental design and detailed protocols description are reported in the Method Section. To investigate whether our SCM-seq approach does indeed allow data integration of 10X and ONT data at single-cell level, we: i) checked the quality of the 10x and ONT experimental outputs, including artifacts eventually introduced in the SCM-seq workflow; ii) identified a set of cellular barcodes (CB) shared between 10x and ONT datasets to match the corresponding outputs; and iii) ensured that 10x and ONT whole-transcriptome datasets were equally representative of the transcriptional heterogeneity of the analysed AML samples.

4.1.1 Quality control of the ONT sequencing output

To check quality of the two ONT experimental outputs (whole-transcriptome or mutation-enriched), we measured a number of sequencing parameters on reads that passed the QC threshold (PHRED ≥7 and ≥10, respectively), including total numbers of reads *per* sample, total number of bases, N50 (a weighted midpoint of the read length distribution; see Materials and Methods, section 3.7), average read length and average read quality (Tables 11 and 12, respectively). For the whole-transcriptome data, despite differences in read depth, the three AML samples showed homogeneous PHRED quality score and only slight differences in N50, in line with previously reported ONT sequencing of the human full-length transcriptome(149). Likewise, the mutation-enriched datasets showed different read depths across samples reads, while N50 and median read length were lower in all samples, as expected for a selective target enrichment procedure. For this dataset, we applied a higher QC threshold (PHRED score ≥ 10), which reflects into increased median PHRED scores.

Table 11. Summary of ONT sequencing output metrics for whole-transcriptome dataset.

Sample	N reads (x 10^6)	N bases (x 10^9)	N50	Median read length (bp)	Median PHRED score
AML4	31.56	28.04	908.75	775.75	11.65
AML5	36.65	37.32	1108.33	892.66	11.64
sAML1	50.69	54.10	1140	927.25	11.73

Table 12. Summary of ONT sequencing output metrics for mutation-enriched dataset.

Sample	N reads (x 10^6)	N bases (x 10^9)	N50	Median read length (bp)	Median PHRED score
AML4	50.76	38.98	801	720	14.985
AML5	16.35	13.54	860	769	15.062
sAML1	38.16	34.50	945	817	14.544

4.1.2 PCR amplification of barcoded cDNA does not introduce major biases on ONT read length and quality

The SCM-seq workflow includes preparation of full-length cDNA tagged with CB and UMI using the Chromium 10x technology. Pools of cDNA were subjected to two consecutive rounds of PCR amplification (PCR1 and PCR2). The latter increased the amount of available material by a factor of 140, on average, and was aimed at generating enough template for target enrichment and ONT sequencing. To investigate whether PCR2 affects whole-transcriptome representation, we compared the output of ONT sequencing on barcoded full-length cDNA after PCR1 and PCR2 separately (prior to target enrichment) using the sAML1 sample (chosen for the abundance of available material). No major differences were observed in the distribution of ONT read length and quality score (Table 13 and Figure 8-9).

Table 13. Summary of ONT read length and quality after PCR1 and PCR2.

Condition	Read status	N reads (x 10^6)	N bases (x 10^9)	N50	Median read length (bp)	Median PHRED score
PCR1	All	5.15	5.55	1170	937	10.72
	Passing QC threshold	4.38	4.85	1190	962	11.12
PCR2	All	13.09	13.47	1100	897	10.55
	Passing QC threshold	11.01	11.67	1120	921	11.02

Figure 8. Effect of PCR1 and PCR2 on ONT read length.
Distribution and quantiles of ONT read length for sAML1 after PCR1 and PCR2.

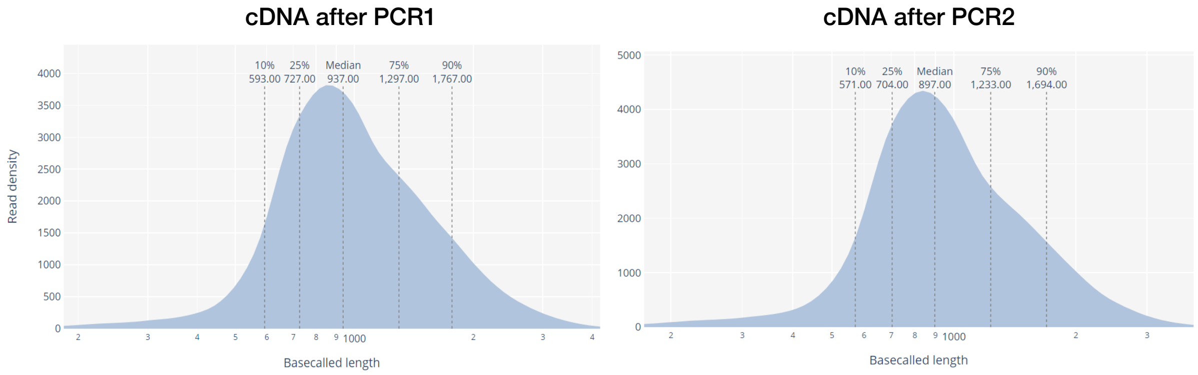
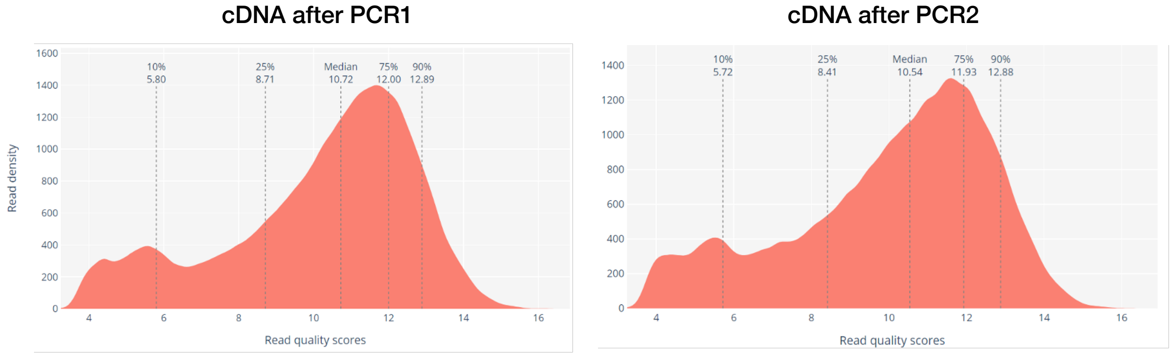


Figure 9. Effect of PCR1 and PCR2 on ONT read quality.
Distribution and quantiles of ONT read PHRED quality scores for sAML1 after PCR1 and PCR2.



Analyses of the effect of PCR2 on the performance of CB identification in ONT reads and transcript/gene identification showed a slight improvement of numbers of CBs matched to 10x data (Figure 10) and of retrieved transcripts/genes (Figure 11).

Figure 10. Effect of PCR1 and PCR2 on CBs identification.
The Venn diagram shows the intersection of sAML1 CBs identified in ONT reads after PCR1 and PCR2. All CBs identified in the ONT datasets also exist in the 10x CB list.

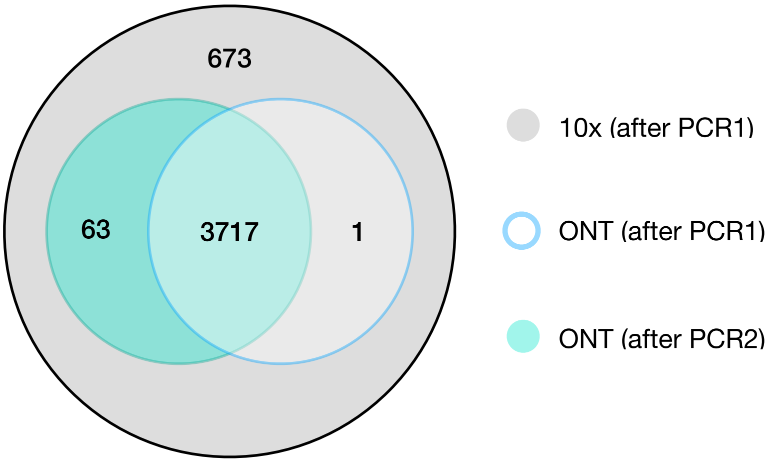
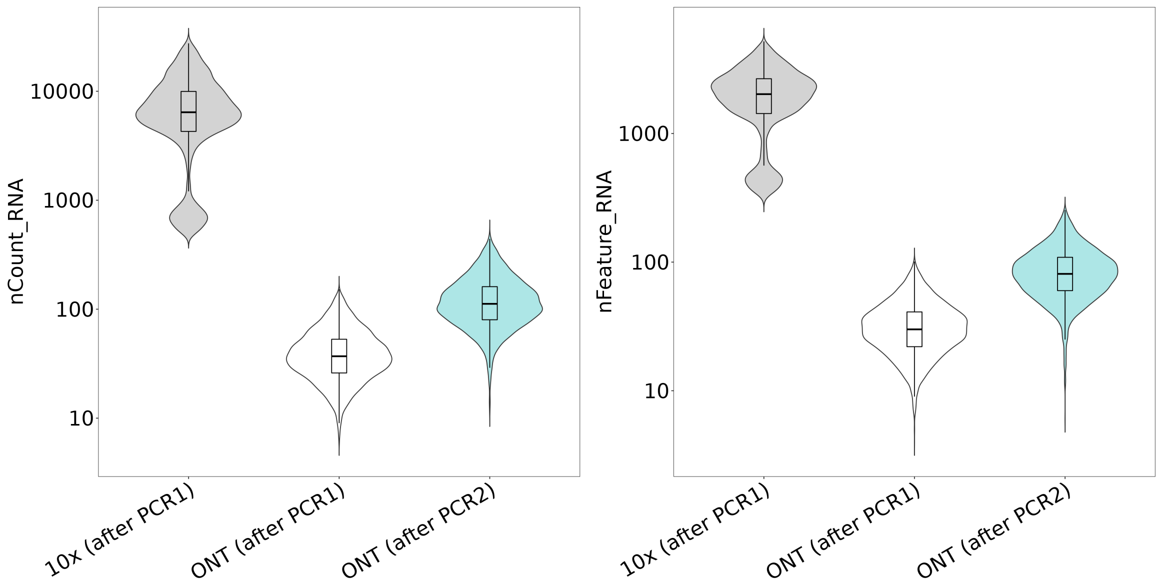


Figure 11. Effect of PCR1 and PCR2 on yield of transcripts and genes.
The violin plots show numbers and distributions of sAML1 transcripts (left) and genes (right) identified in the 10x dataset and ONT datasets after PCR1 and PCR2.



Thus, PCR2 represents a safe strategy to increase the amount of barcoded full-length cDNA available for target enrichment and ONT library preparation.

4.1.3 Identification of 10x CBs in the ONT dataset

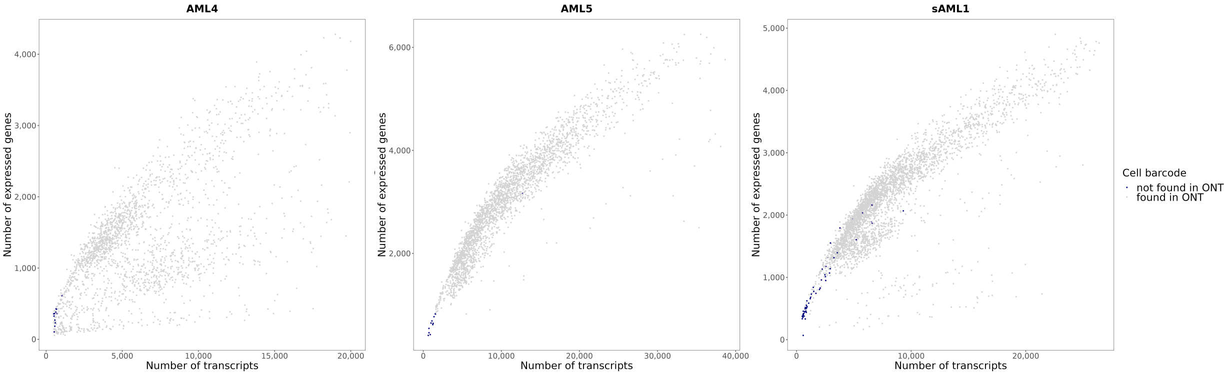
We then investigated the performance of identification in ONT data of CB matching the 10x CBs. For each of the three analyzed samples, we retrieved >98% of the 10x CB in the ONT dataset (Table 14).

Table 14. Numbers of 10x CBs identified in ONT datasets.

Sample	CB in 10x	CB common to 10x and ONT	CB in 10x only	CB in ONT only
AML4	2099	2087	12	13
AML5	2638	2627	11	28
sAML1	3465	3410	55	182

Remarkably, cells in the 10x dataset for which we could not find the corresponding CB in ONT data (on average 0.85% across the three samples) typically showed poor quality, i.e. low counts of genes and transcripts, as shown in Figure 12.

Figure 12. Quality of CBs identified in 10x data and missing in ONT data.
Quality of CBs identified in 10x data is assessed based on number of expressed transcripts and genes. CBs colored in blue are found in the 10x dataset but not in the ONT dataset.

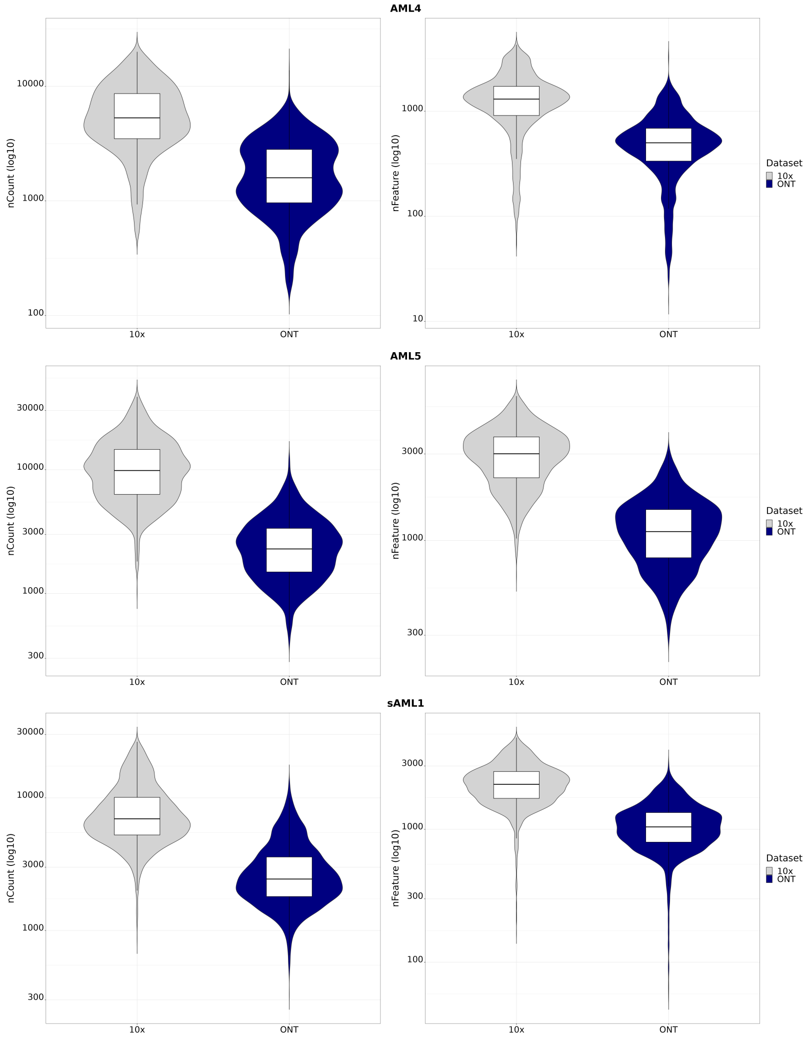


4.1.4 ONT data are fully representative of the transcriptional heterogeneity of 10x data

To ensure that 10x and non-enriched ONT datasets were fully comparable, we assessed the correspondence of transcriptional features and heterogeneity between the two sequencing platforms at both gene- and cell-level.

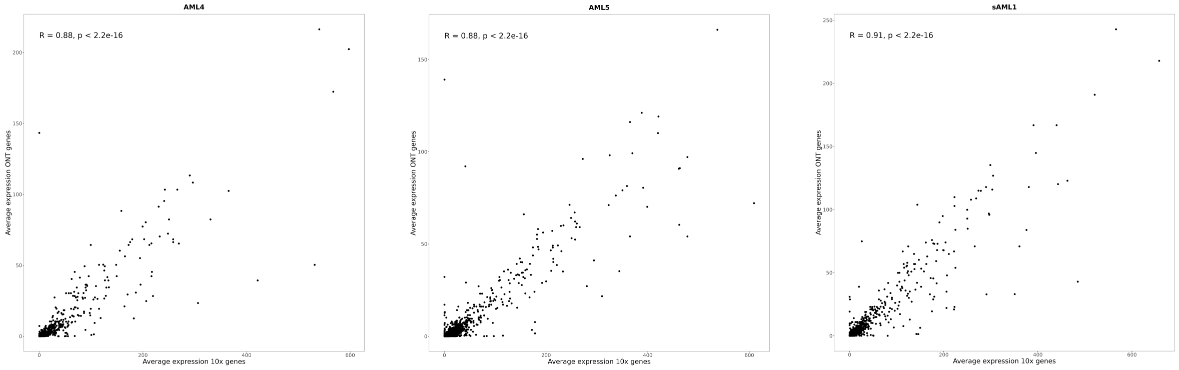
First, we investigated whether the two sequencing methods were comparable in terms of number of identified transcripts and genes on the same cells. As shown in Figure 13, ONT sequencing of non-enriched full-length cDNA identified less transcripts and genes as compared to 10x. This result is in accordance with previous studies(143) and might depend on read depth at matched CBs.

Figure 13. Comparison of 10x and ONT sequencing platforms by yield of transcripts and genes.
The violin plots show numbers and distribution of transcripts (left) and genes (right) for CBs identified in both the 10x and ONT dataset.



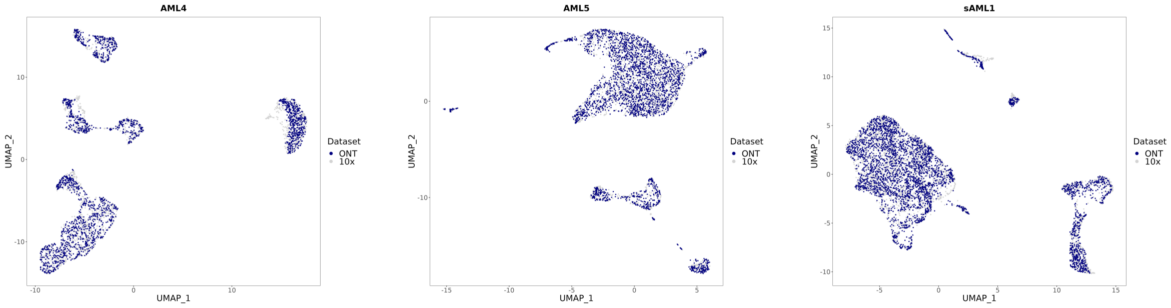
To test the consistency of expression levels for individual genes, we examined the relationship in the two datasets of average UMI counts *per* gene, considering genes expressed in >3 cells and shared between the 10x and ONT gene-count matrices. For all of the three samples, expression levels showed a strong linear correlation (Spearman correlation r 0.88, 0.88 and 0.91, p value < 2.2e-16 in all cases) (Figure 14).

Figure 14. Gene-by-gene correlation in 10x and ONT data.
Spearman correlation highlights the relationship of average UMI counts *per* gene, for cells and genes shared between the 10x and ONT gene-count matrices.



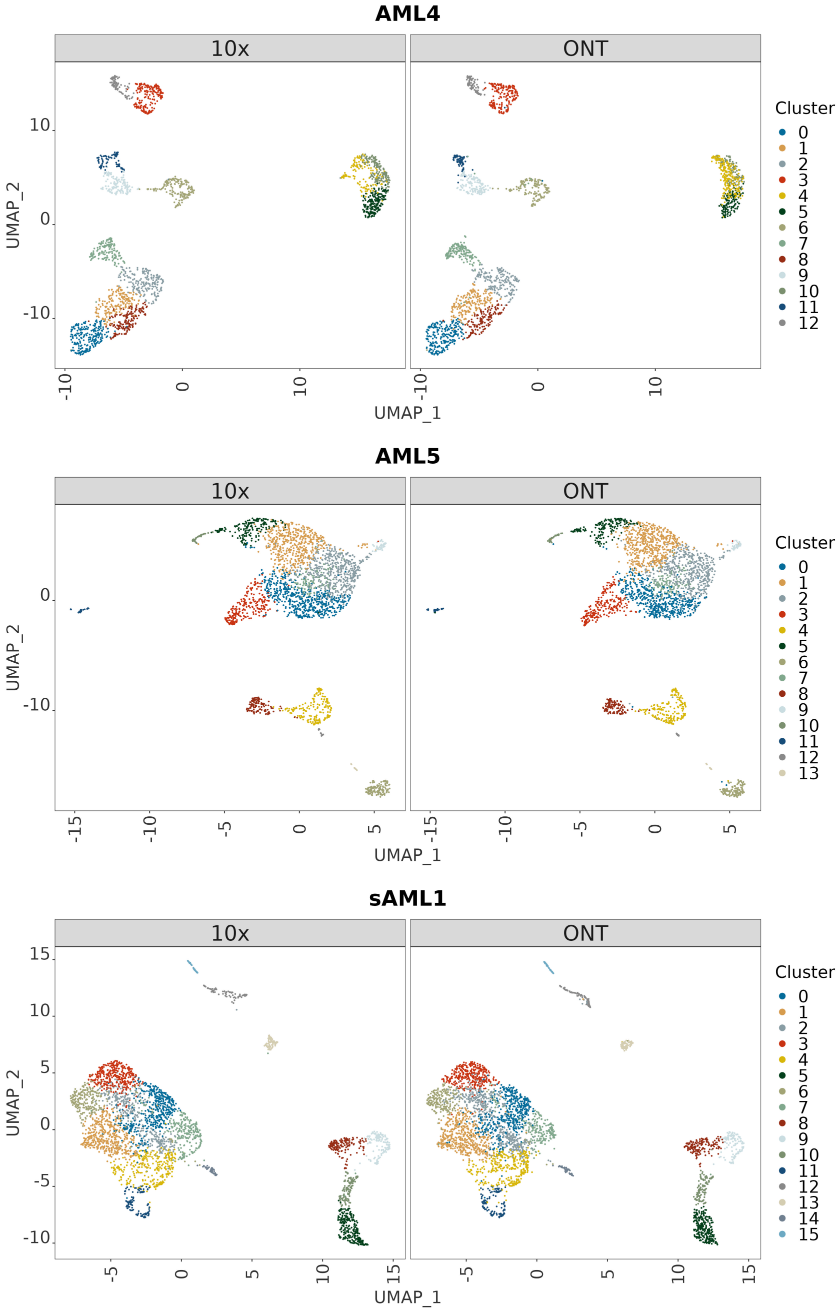
Finally, we compared the transcriptional identity of 10x and ONT cells directly, by visualization of expression profiles-driven similarities between cells in the UMAP space. To this end, we integrated and normalized/scaled gene-count matrices from the two datasets using the *SCTransform* function of Seurat, and then performed dimensionality reduction by PCA. Topology distribution of cells was largely overlapping across datasets, suggesting that the two technologies are consistent in terms of overall expression profiles, thus allowing comparable representation of the underlying transcriptional heterogeneity (Figure 15).

Figure 15. UMAP of integrated 10x and ONT scRNA data.
Cells scored with 10x and ONT are integrated in the same transcriptional space and colored by the corresponding sequencing platform.



To assess whether the two sequencing methods could capture most of the transcriptional variability of the AML samples, we performed clustering by expression similarity on cells from the integrated dataset using the *FindNeighbors* and *FindClusters* Seurat functions to compute k-nearest neighbor and run the Louvain algorithm for modularity optimization (on the first 15 principal components and with resolution parameter of 0.8). Then, we checked the similarity between cluster identity of the same CB according to the different sequencing platform. Namely, to measure the proportions of agreements between the two partitions, we computed the adjusted Rand index, which resulted consistently high (0.74, 0.77 and 0.73 for AML4, AML5 and sAML1, respectively). The high rate of correspondence was also evident upon visual inspection of the UMAP (Figure 16).

Figure 16. UMAP of integrated 10x and ONT scRNA data upon Louvain clustering.
Cells scored with 10x and ONT are integrated in the same transcriptional space and colored by transcriptional cluster assignment after Louvain clustering.



Overall, despite the lower number of transcripts and genes scored in the ONT dataset, we conclude that 10x and ONT sequencing provide comparable whole-transcriptome information both in terms of gene-level quality and cell-level transcriptional heterogeneity, which is essential to then reliably integrate distinct features of the approaches.

4.2. Mutation analysis at bulk and single-cell level

4.2.1 Mutational profile by WES

Somatic variant calling on WES datasets (median coverage ~100x) of sample-matched germline and leukemia sequences identified 23, 22 and 24 coding mutations for AML4, AML5 and sAML1, respectively. We then selected non-synonymous variants with >0.1 variant allelic fraction (VAF) (11, 15 and 16, respectively) for the design of mutation-specific target-enrichment probes. Expectedly, each AML case carried both driver and non-driver mutations, as based on comparison with established lists of recurrent driver mutations in AML(150,151). Drivers and non-drivers mutations are given sequentially in Table 15 for each sample, in decreasing order by VAF. All mutations were single nucleotide variants, but for two frameshifts, one insertion (*ASXL1*) and one deletion (*RUNX1*). All of the three AMLs showed mutations of the SRSF2 gene, which encodes a member of the serine/arginine (SR)-rich family of pre-mRNA splicing factors and is frequently mutated in AMLs(78). In keeping with previous studies, all the three *SRSF2*-mutated AMLs showed co-occurrence of variants in genes regulating DNA methylation and chromatin modification (*DNMT3A*, *IDH1*, *ASXL1, EZH2*), as well as members of the cohesin complex (*STAG2*). Mutations in genes belonging to signaling and kinase pathways were only represented in sAML1 (*FLT3* and *CSF3R*), with relatively lower VAF as expected for this functional category(36).

Table 15. Non-synonymous variants selected for target enrichment and single-cell analysis.

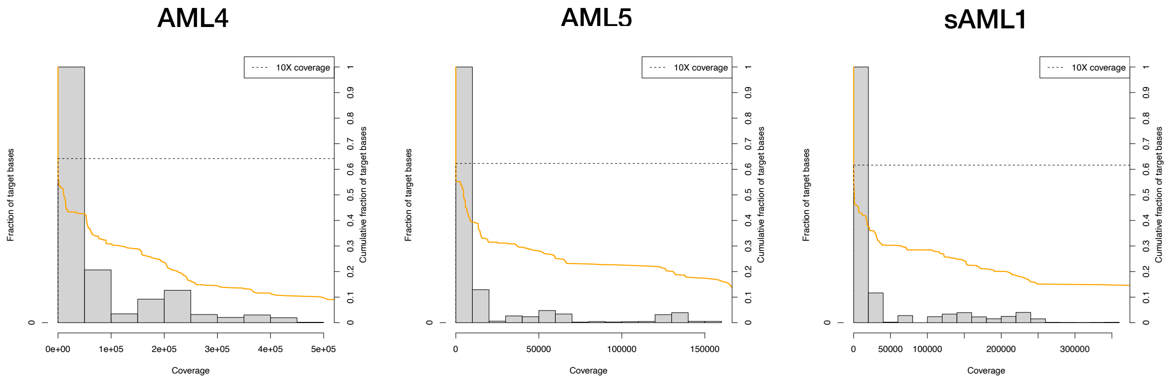
Driver/non-driver classification based on references 150-151.

Sample	Gene	Description	Variant Classification	Genome Change	cDNA Change	Protein Change	VAF	Known AML driver
AML4	STAG2	stromal antigen 2	Nonsense, SNV	g.chrX:124037566C>T	c.328C>T	p.R110*	0.716	yes
	SRSF2	serine and arginine rich splicing factor 2	Missense, SNV	g.chr17:76736877G>A	c.284C>T	p.P95L	0.398	yes
	DNMT3A	DNA methyltransferase 3 alpha	Missense, SNV	g.chr2:25244580C>A	c.1627G>T	p.G543C	0.361	yes
	IDH1	isocitrate dehydrogenase (NADP(+)) 1, cytosolic	Missense, SNV	g.chr2:208248389G>A	c.394C>T	p.R132C	0.333	yes
	CEBPA	CCAAT/enhancer binding protein alpha	Missense, SNV	g.chr19:33301454T>C	c.961A>G	p.N321D	0.205	yes
	MED12	mediator complex subunit 12	Nonsense, SNV	g.chrX:71137246C>T	c.5611C>T	p.R1871*	0.823	-
	ASXL3	additional sex combs like 3, transcriptional	Missense, SNV	g.chr18:33745114G>C	c.5266G>C	p.G1756R	0.431	-
	PCSK2	proprotein convertase subtilisin/kexin type 2	Missense, SNV	g.chr20:17481769G>A	c.1616G>A	p.R539H	0.431	-
	FRYL	FRY like transcription coactivator	Missense, SNV	g.chr4:48510879T>A	c.8251A>T	p.M2751L	0.396	-
	ABL1	ABL proto-oncogene 1, non-receptor tyrosine	Missense, SNV	g.chr9:130884970A>G	c.2737A>G	p.R913G	0.383	-
	CCDC158	coiled-coil domain containing 158	Missense, SNV	g.chr4:76369485G>A	c.1288C>T	p.R430C	0.173	-
AML5	RUNX1	runt related transcription factor 1	Missense, SNV	g.chr21:34880581T>C	c.403A>G	p.R135G	0.819	yes
	ASXL1	additional sex combs like 1, transcriptional	Frame_Shift_Ins	g.chr20:32434638_32434639insG	c.1926_1927insG	p.G646fs	0.442	yes
	SRSF2	serine and arginine rich splicing factor 2	Missense, SNV	g.chr17:76736877G>C	c.284C>G	p.P95R	0.435	yes
	EZH2	enhancer of zeste 2 polycomb repressive	Nonsense, SNV	g.chr7:148826499G>A	c.862C>T	p.R288*	0.41	yes
	ADAMTS7	ADAM metallopeptidase with thrombospondin	Frame_Shift_Ins	g.chr15:78774759_78774760insC	c.1740_1741insG	p.R581fs	0.39	yes
	ZAN	zonadhesin (gene/ pseudogene)	Missense, SNV	g.chr7:100763889G>A	c.4070G>A	p.R1357Q	0.676	-
	HRH4	histamine receptor H4	Missense, SNV	g.chr18:24476804G>A	c.415G>A	p.V139I	0.46	-
	CACNA1E	calcium voltage-gated channel subunit alpha1 E	Missense, SNV	g.chr1:181758014C>G	c.4397C>G	p.T1466S	0.426	-
	SPATA18	spermatogenesis associated 18	Missense, SNV	g.chr4:52072099G>A	c.605G>A	p.R202Q	0.425	-
	PDZD7	PDZ domain containing 7	Missense, SNV	g.chr10:101023943C>G	c.352G>C	p.E118Q	0.405	-
	SLC35B2	solute carrier family 35 member B2	Missense, SNV	g.chr6:44255600G>C	c.250C>G	p.P84A	0.389	-
	PCSK4	proprotein convertase subtilisin/kexin type 4	Missense, SNV	g.chr19:1484039G>A	c.1157C>T	p.A386V	0.325	-
	CNTNAP3	contactin associated protein like 3	Missense, SNV	g.chr9:39078846A>G	c.3517T>C	p.F1173L	0.32	-
	ALPP	alkaline phosphatase, placental	Missense, SNV	g.chr2:232379003G>C	c.109G>C	p.E37Q	0.247	-
	OR2T8	olfactory receptor family 2 subfamily T member 8	Missense, SNV	g.chr1:247921277G>A	c.260G>A	p.S87N	0.113	-
sAML1	RUNX1	runt related transcription factor 1	Frame_Shift_Del	g.chr21:34880691delG	c.293delC	p.P98fs	0.688	yes
	SRSF2	serine and arginine rich splicing factor 2	Missense, SNV	g.chr17:76736877G>A	c.284C>T	p.P95L	0.369	yes
	IDH1	isocitrate dehydrogenase (NADP(+)) 1, cytosolic	Missense, SNV	g.chr2:208248389G>A	c.394C>T	p.R132C	0.322	yes
	FLT3	fms related tyrosine kinase 3	Missense, SNV	g.chr13:28018505C>A	c.2503G>T	p.D835Y	0.183	yes
	CSF3R	colony stimulating factor 3 receptor	Nonsense, SNV	g.chr1:36466647G>A	c.2302C>T	p.Q768*	0.176	yes
	MACF1	microtubule-actin crosslinking factor 1	Missense, SNV	g.chr1:39337255T>C	c.10154T>C	p.I3385T	0.372	-
	TTF2	transcription termination factor 2	Missense, SNV	g.chr1:117097417C>T	c.3253C>T	p.L1085F	0.371	-
	WDR63	WD repeat domain 63	Missense, SNV	g.chr1:85081244A>T	c.114A>T	p.E38D	0.351	-
	HDAC7	histone deacetylase 7	Missense, SNV	g.chr12:47797107T>C	c.613A>G	p.K205E	0.338	-
	MALRD1	MAM and LDL receptor class A domain	Missense, SNV	g.chr10:19595262G>A	c.5749G>A	p.V1917I	0.337	-
	AADAT	aminoadipate aminotransferase	Missense, SNV	g.chr4:170069193C>T	c.758G>A	p.G253E	0.333	-
	ADAMTS16	ADAM metallopeptidase with thrombospondin type 1 motif 16	Missense, SNV	g.chr5:5182053A>T	c.511A>T	p.I171L	0.326	-
	SETD2	SET domain containing 2	Missense, SNV	g.chr3:47120267G>T	c.4369C>A	p.P1457T	0.276	-
	USH1C	USH1 protein network component harmonin	Missense, SNV	g.chr11:17531180C>T	c.361G>A	p.G121S	0.189	-
	PCDHB7	protocadherin beta 7	Missense, SNV	g.chr5:141174891G>A	c.2056G>A	p.V686I	0.162	-
	DGKI	diacylglycerol kinase iota	Missense, SNV	g.chr7:137469601G>A	c.2446C>T	p.P816S	0.118	-

4.2.2 Performance of mutation-specific target enrichment

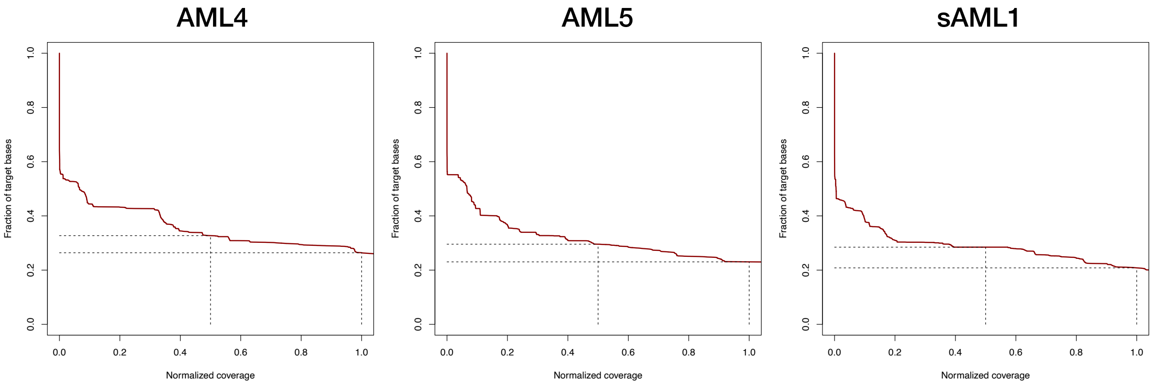
Prior to mutation analysis of ONT data, we investigated the performance of target enrichment by measuring capture specificity and sequence coverage of the targeted mutations. Data showed a consistently high capture specificity (measured as the fraction of aligned reads overlapping any target region) of 72.5%, 80.5% and 83.7% for AML4, AML5 and sAML1, respectively. Sequence coverages (i.e., numbers of reads for each base that is sequenced and/or located in a target region) are shown in Figure 17. Overall, the vast majority of targeted bases showed a coverage of at least 10X.

Figure 17. Coverage analysis of mutation-specific target enrichment.
The histograms show coverage distribution and cumulative density of target base coverage after mutation-specific target enrichment. The orange lines indicate the cumulative fraction of targeted bases (as of the right axis) with coverage of at least the value indicated by the x axis, while the dashed lines highlight the cumulative fraction of target bases with at least 10X coverage.



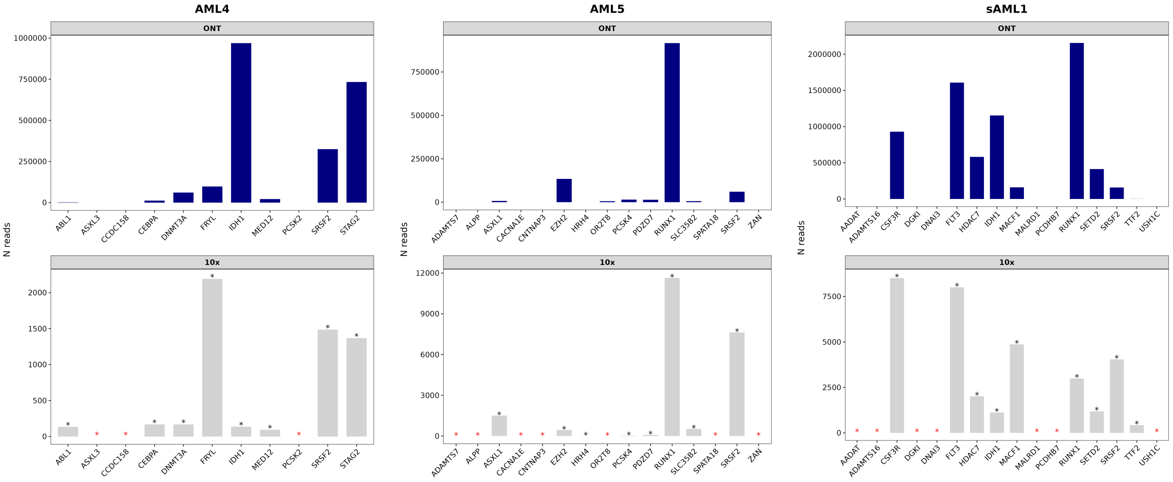
Since absolute numbers of sequenced reads differed across samples, possibly biasing the evaluation and comparison of coverage uniformity, we normalized the *per*-base coverages by the average coverage over all targeted bases (Figure 18). The x-axis in the figure is truncated at 1, which corresponds to the average normalized coverage. After normalization, all samples showed mostly poor variation in the fraction of targeted bases achieving at least the average or or at least half the average coverage (dashed lines), indicating mostly-homogeneous coverage uniformity.

Figure 18. Coverage uniformity of mutation-specific target enrichment across AML samples.
Per-base coverages are normalized by the average coverage over all targeted bases.



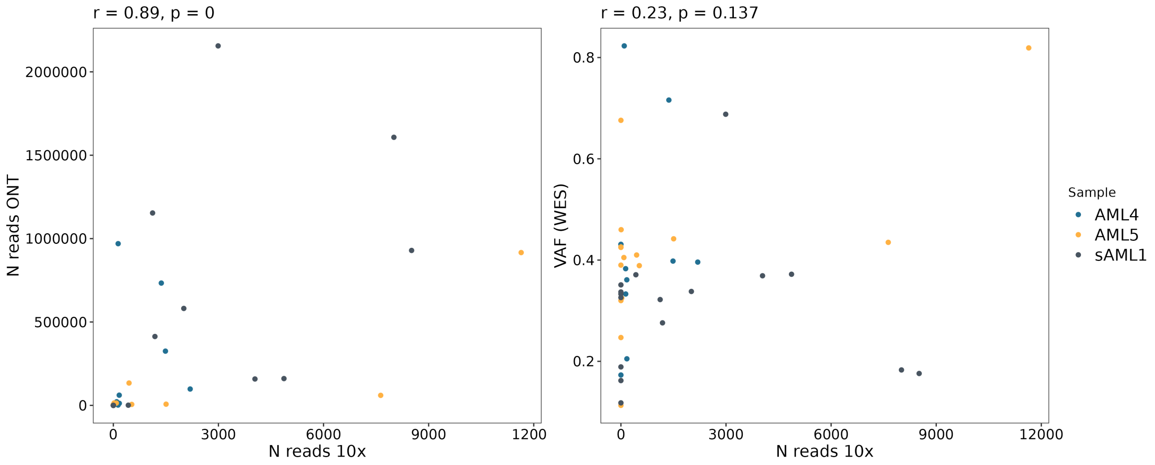
Analyses of the representation of individual target regions in ONT data, however, showed widely varying coverage across different targets (Figure 19, top panel). In particular, correlation analysis revealed a strong and significant association between target-specific coverage in the ONT dataset and gene expression levels in the 10x expression dataset (Spearman’s rank coefficient 0.89, p<0.001) (Figure 20, left panel). As shown in Figure 19 (bottom panel, bars marked by black stars), all targets in genes with detectable expression levels also showed significant coverage (8 out of 11 for AML4; 8 out of 15 for AML5; 9 out of 16 for sAML1). About 40% of the targets, however, were located in genes with no expression and no representation in ONT data (Figure 19, bottom panel, bars marked by a red star). Intriguingly, most of these gene variants showed relatively high VAFs (>30-40%) suggesting that they were acquired early during leukemogenesis. The same genes, however, have not been reported as drivers in AMLs (see Table 15), suggesting that expression may not correlate with genetic penetrance in passenger mutations.

Figure 19. Coverage across different target genes by sequencing platform.
The barplots show coverage (i.e., read depth) of targeted genes by ONT (top) and 10x (bottom). Black stars highlight genes expressed in 10x and recovered in ONT. Red stars highlight genes not expressed in 10x and not recovered in ONT.



Consistently, we did not find any relationship between the WES VAFs of mutated genes and differences across their expression levels (Spearman's rank coefficient 0.23, p = 0.137) (Figure 20, right panel).

Figure 20. Correlation between 10x expression of target genes, ONT expression and VAF.
Spearman correlation highlights the relationship between number of 10x reads covering target genes and number of ONT reads covering target genes (left) and WES VAF (right).



4.2.3 Mutation mapping

ONT reads for variants with detectable expression levels were filtered based on BQ ≥10 (BQ is a PHRED-scaled quality score, ranging 0-60) and MQ ≥20 (MQ measures the probability of the read being incorrectly aligned, ranging 0-40), resulting into a set of high-quality reads reduced that comprises an average of 88% of all variants (range 80.2-98.1%) (Figure 21, 22 and 23). As detailed in the Materials and Methods section (paragraph 3.9.2), we then classified filtered reads on the basis of the known variant position (from WES data) as mutated, wild-type or mismatch. The percentage of mismatch reads, which represents an indirect measure of the sequencing error-rate at each variant position, was on average less than 5% of all filtered reads (range 0.75 - 13.3%), thus reassuring for the reliability of our ONT-based approach (Figure 21, 22 and 23). Of note, we could successfully map also variants different from single nucleotide substitutions, namely one frameshift deletion in *RUNX1* and one frameshift insertion in *ASXL1*.

Figure 21. Output of mutation mapping on ONT data (sample AML4).
Number of ONT reads after mutation mapping (grey bar), after filtering according to base and mapping quality (blue bar) and proportion of mutated, wild-type and mismatch reads for each variant.

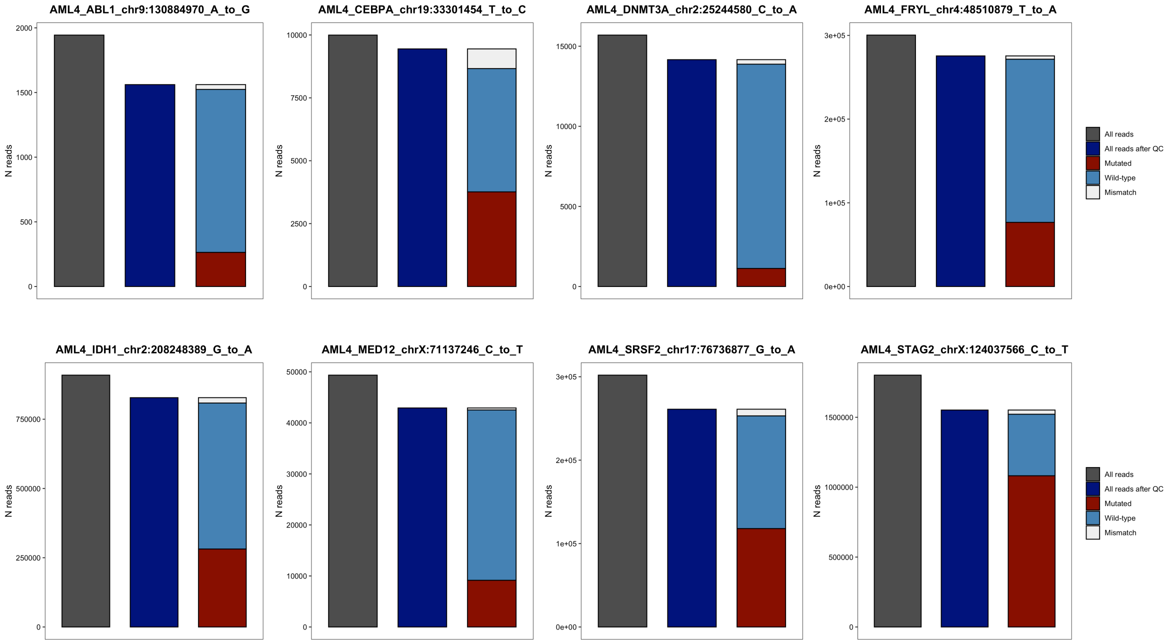


Figure 22. Output of mutation mapping on ONT data (sample AML5).
As in Figure 21.

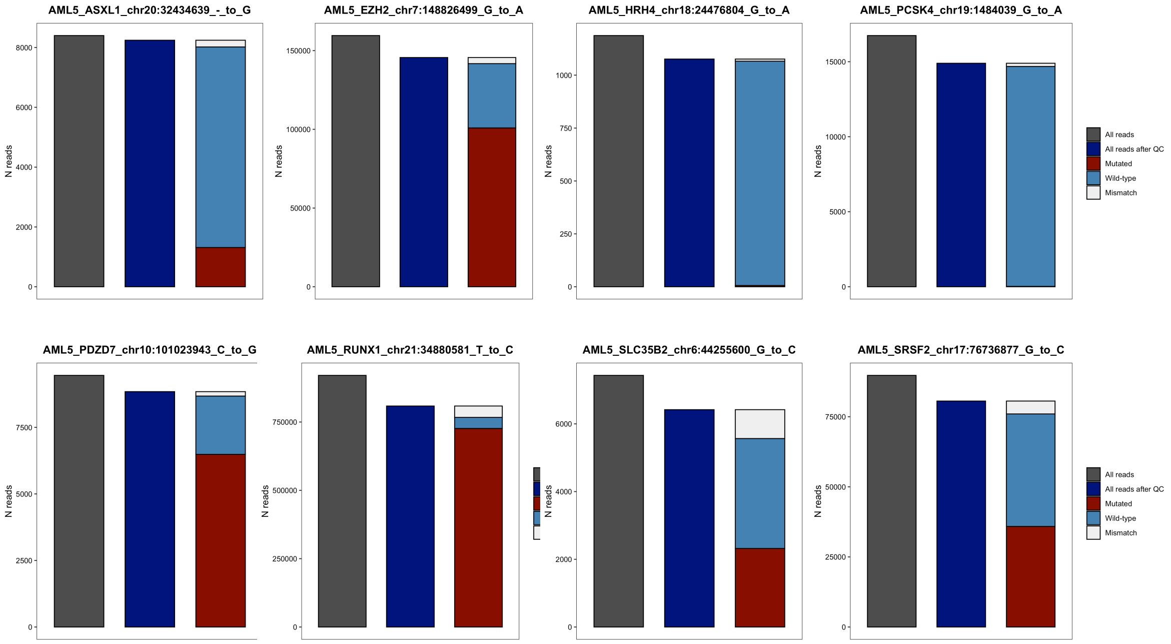
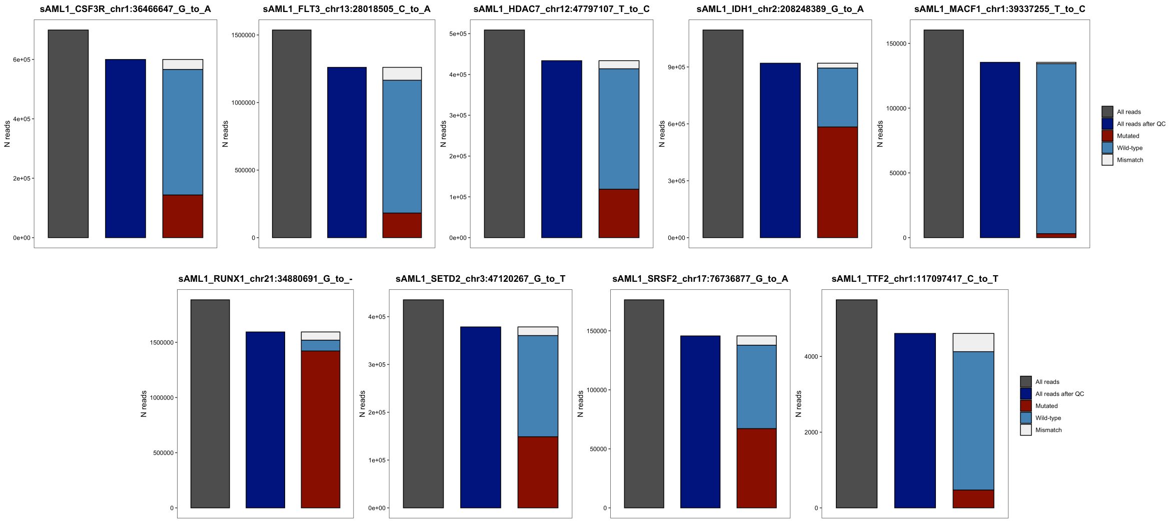


Figure 23. Output of mutation mapping on ONT data (sample sAML1).
As in Figure 21.



To further assess the accuracy of mutation mapping, we measured the frequency of sequencing errors at and around each variant position. To this end, we constructed a position frequency matrix (PFM) by scoring the occurrence of any nucleotide (or deletions) at each position in a region of 10 bps around each given SNV (21 bp in total). The PFM is visualized by bars for each scored nucleotide (or deletion; black bar) (Figure 24, 25 and 26) and allows visualization of the consensus sequence at and around each variant position and evaluation of the local occurrence of sequencing errors in the variant region. In all cases, the expected base (including the variant base) occurred at dramatically higher frequency as compared to any other, suggesting a very low frequency of local sequencing errors.

Figure 24. Variant position frequency matrix (sample AML4).
The bars represent the frequency of the scored nucleotides. The x-axis indicates the expected base. Nucleotides at the position of the variant of interest are highlighted in the light-blue box. Nucleotides at 10 positions afterwards and downwards are also presented.

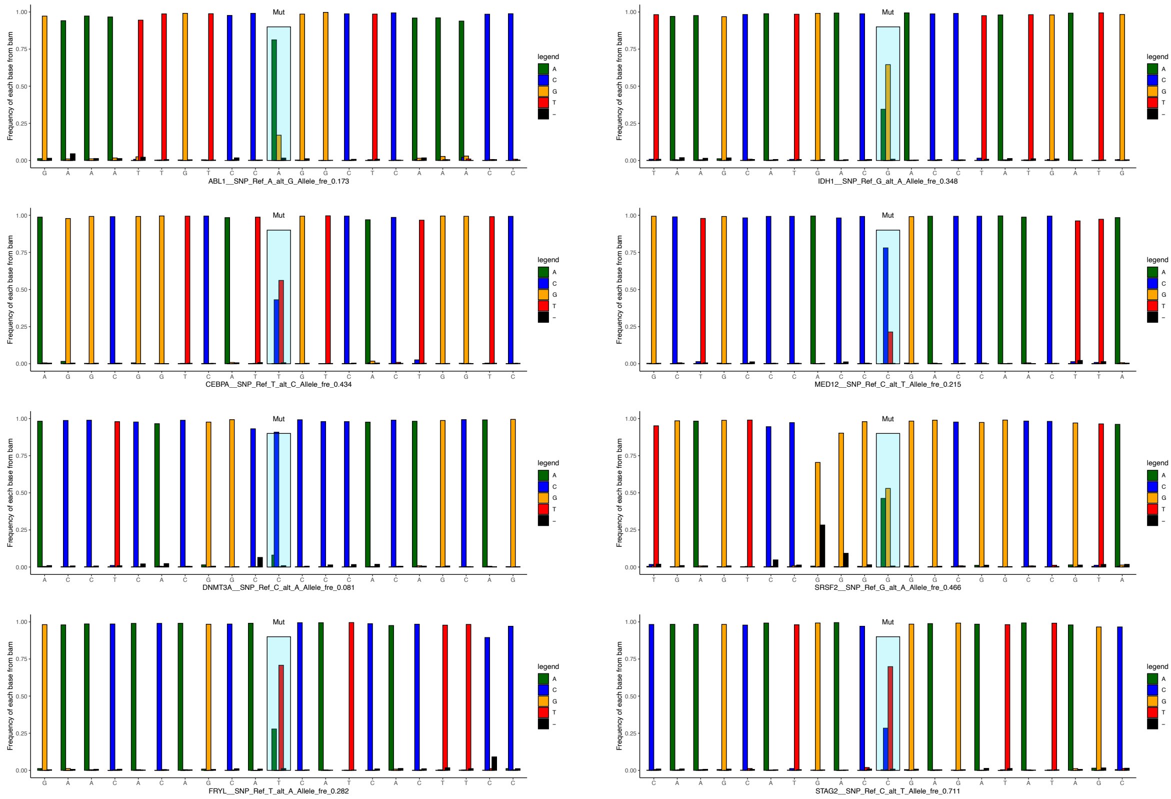


Figure 25. Variant position frequency matrix (sample AML5).

As in Figure 24.

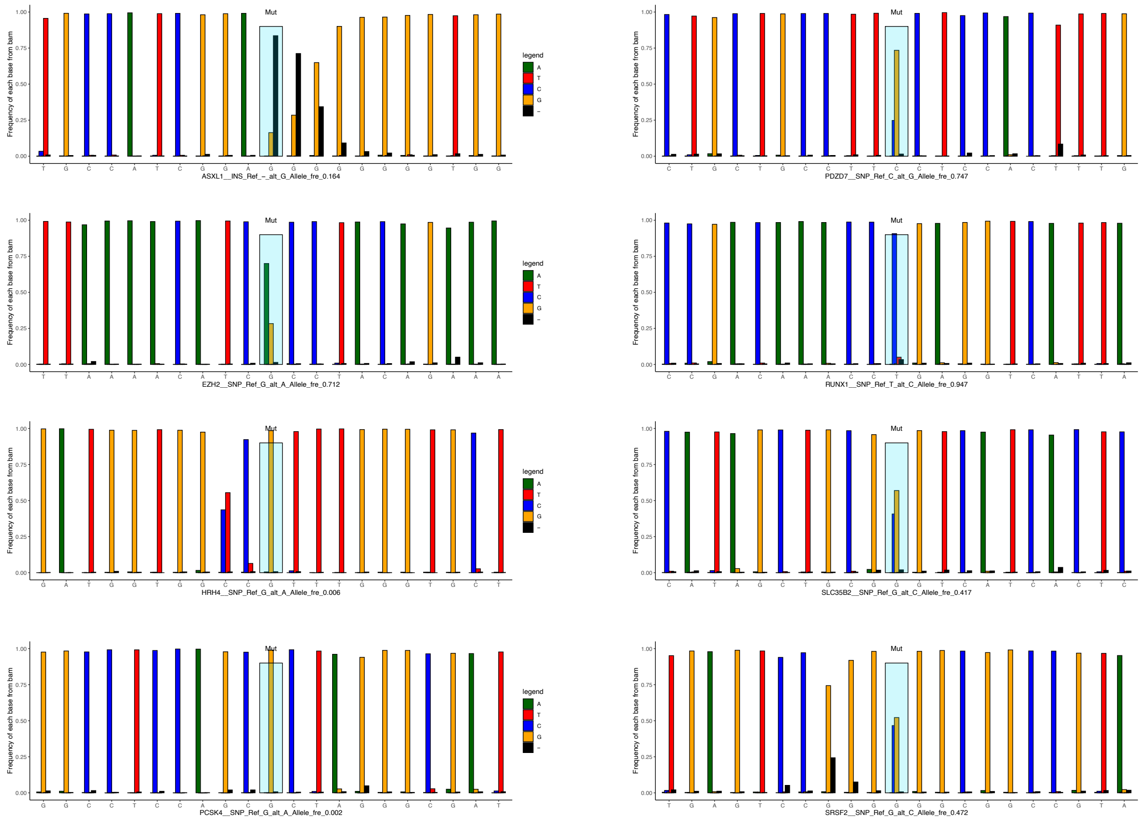
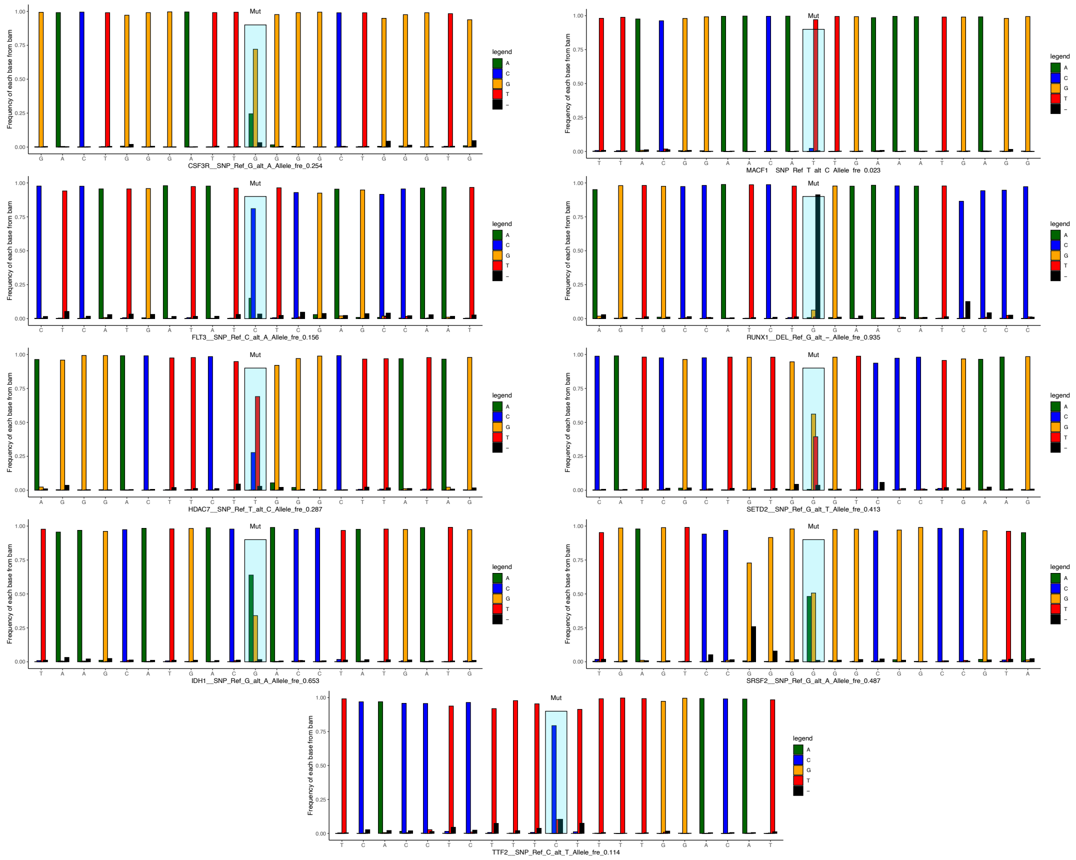


Figure 26. Variant position frequency matrix (sample sAML1).

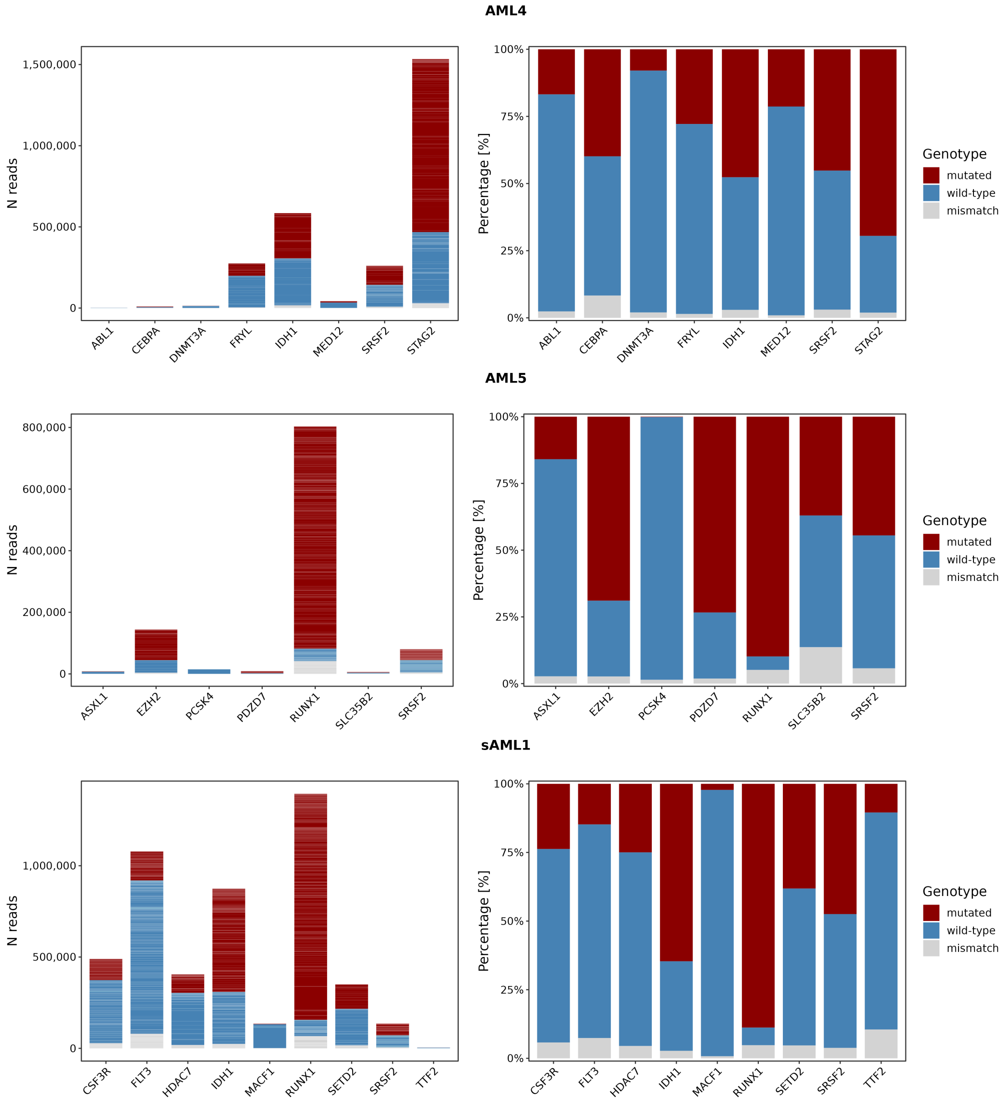
As in Figure 24.



4.2.4 Performance of single-cell genotype imputation

To impute genotypes at single-cell level, we first demultiplexed ONT reads for each variant to corresponding CBs. This process allowed to retrieve 1,415, 2,332 and 3,085 cells from the AML4, AML5 and sAML1 datasets, respectively, corresponding to 60.8%, 88.8% and 90.4% of CB shared between the 10x and ONT datasets for each sample. During this step, we discarded the AML5 variant in *HRH4* due to its representation in <5 cells. As expected, proportions of mutated, wild-type and mismatch reads at gene level (Figure 27) were comparable to those observed in the same dataset prior to demultiplexing (as reported in Figures 21-23).

Figure 27. ONT read counts by gene mutation and genotype category after demultiplexing by CBs.
The bars represent raw counts (left) and percentages (right) of reads covering each target gene mutation, and are colored by genotype category.



The proportion of mutated, wild-type and mismatch reads at single-cell level, instead, were significantly more heterogenous (highlighted in the top panels of Figure 28, 29 and 30, which show the distribution of number and genotype category of reads covering each cell for each gene variant). We then computed accuracy and confidence scores for genotype imputation for each of the analysed variants at single-cell level, by submitting the de-multiplexed ONT dataset to our single-cell genotyping algorithm (see Materials and Methods section, paragraph 3.9.3). Bottom panels in Figure 28, 29 and 30 show the outcomes of cell genotype imputation using ≥0.90 and ≥0.75 as thresholds of accuracy and confidence, respectively. Statistics of genotyping results are summarized in Table 16. Overall, SCM-seq allowed genotyping of 8 variants in AML4, 5 in AML5 and 9 in sAML1 out of 11, 15 and 16 targeted variants, respectively. Two additional AML5 variants (*ASXL1* and *PCSK4*) were discarded for their low cellular prevalence (≤10 cells covered by mutated or wild-type reads).

The percentage of cells in the 10x dataset showing genotype information for at least one gene variant was fairly high (60.5%, 78.9% and 83% in the three samples, respectively). Genotype imputation was especially successful for cells covered by higher numbers of reads, regardless of the presence of mismatches. Most cells with failed genotype assignment were indeed cells covered by just one or very few reads (Figure 28, 29 and 30, bottom panels).

Figure 28. Cell-wise read counts by genotype category and genotyping outcome (sample AML4).

The bars represent the distribution of log10-normalized read counts covering each cell by genotype category for each variant (top) and outcome of genotyping (bottom).

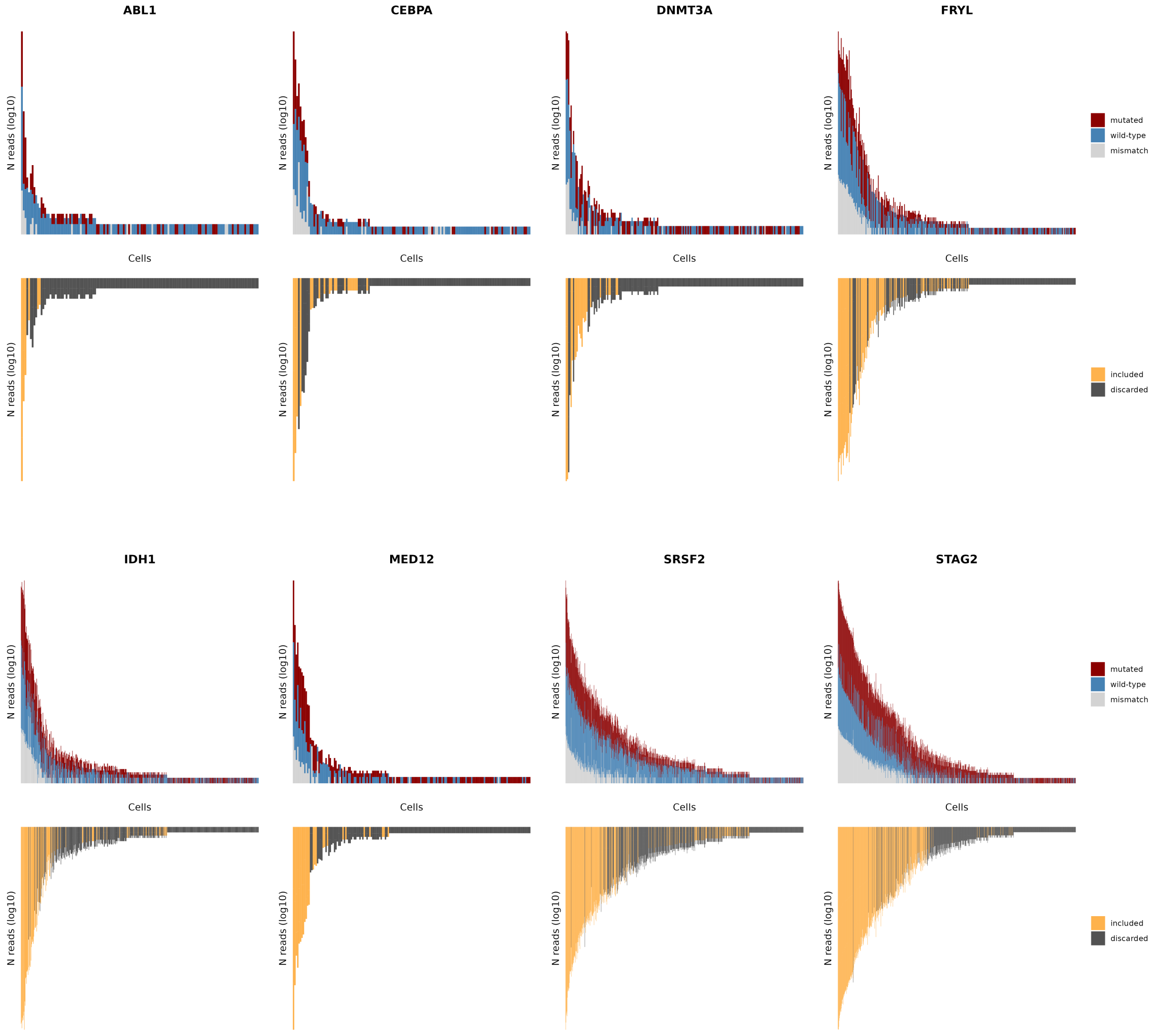


Figure 29. Cell-wise read counts by genotype category and genotyping outcome (sample AML5).

As in Figure 28.

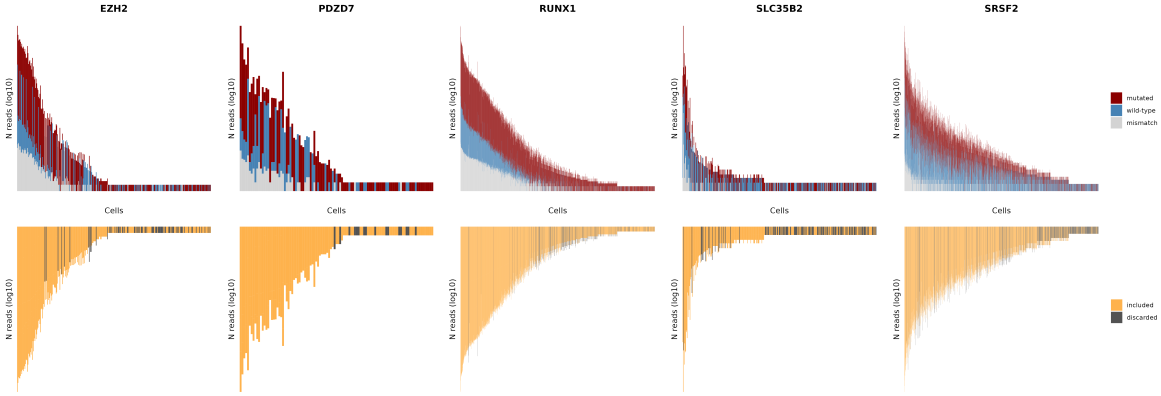


Figure 30. Cell-wise read counts by genotype category and genotyping outcome (sample sAML1).
As in Figure 28.

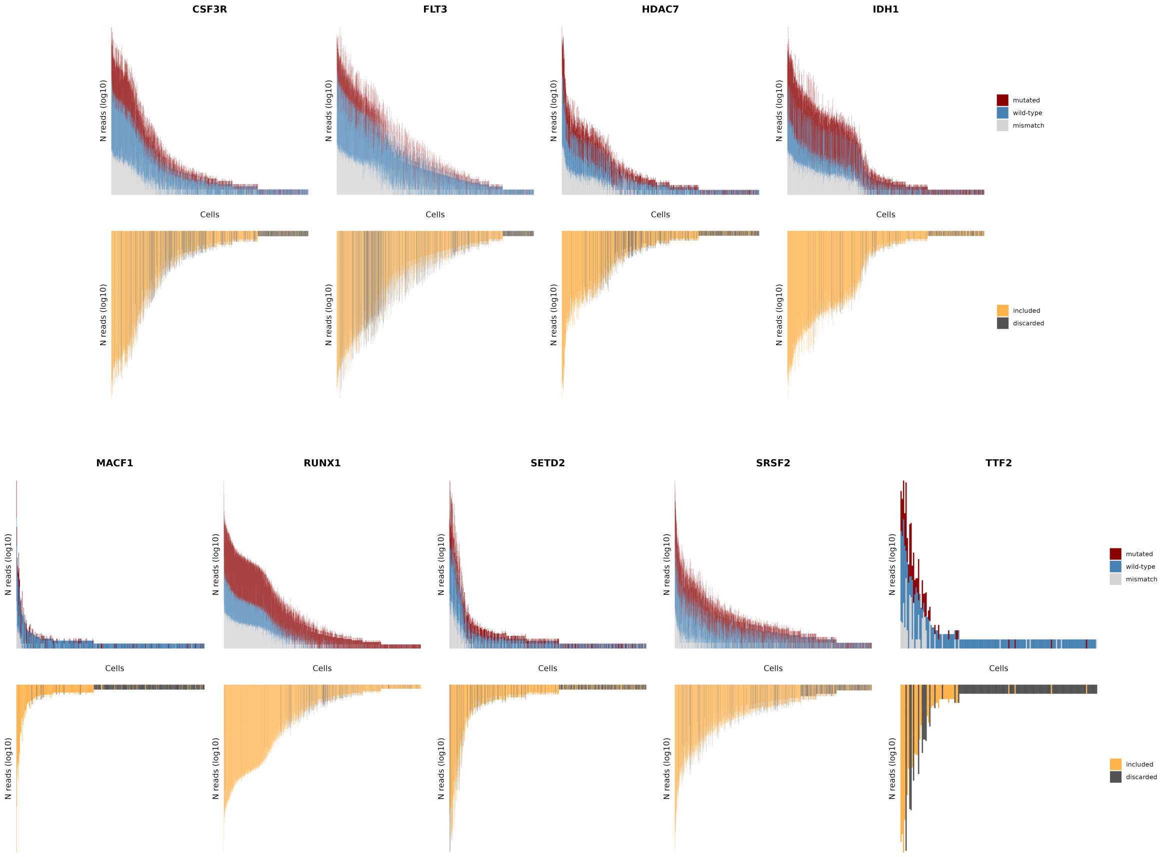


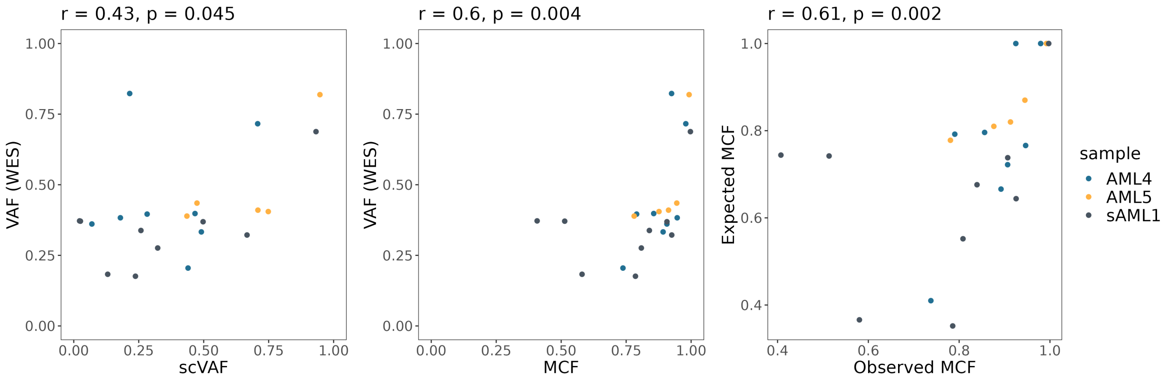
Table 16. Statistical summary of single-cell genotype imputation by SCM-seq.

	AML4	AML5	sAML1
Cells from 10x dataset, n	2099	2638	3465
Genotyped cells, n (%)	1270 (60.5)	2083 (78.9)	2876 (83)
Total mutant cells, n (% of genotyped cells)	1195 (94)	2026 (97.2)	2693 (93.6)
Mutations *per* cell, median (range)	1 (0-7)	2 (0-4)	2 (0-8)
Mutant cells *per* variant, median (range)	152 (45-766)	210 (78-1438)	862 (10-1827)
Wild-type cells *per* variant, median (range)	20 (11-84)	16 (3-96)	134 (5-656)

Most genotyped cells were mutant cells (94%, 97.2% and 93.6% of genotyped cells for the three samples), with a median of 152, 210 and 862 mutant cells *per* variant (Table 16). Wild-type cells were far less represented, with a median of 20, 16 and 134 cells *per* variant. While underrepresentation of wild-type cells is expected within a tumor population, it may also reflect the inherent definition of wild-type cells, which can be called only in the absence of concomitant mutated reads and the already mentioned sparseness of sequencing depth (i.e., presence of cells with low coverage).

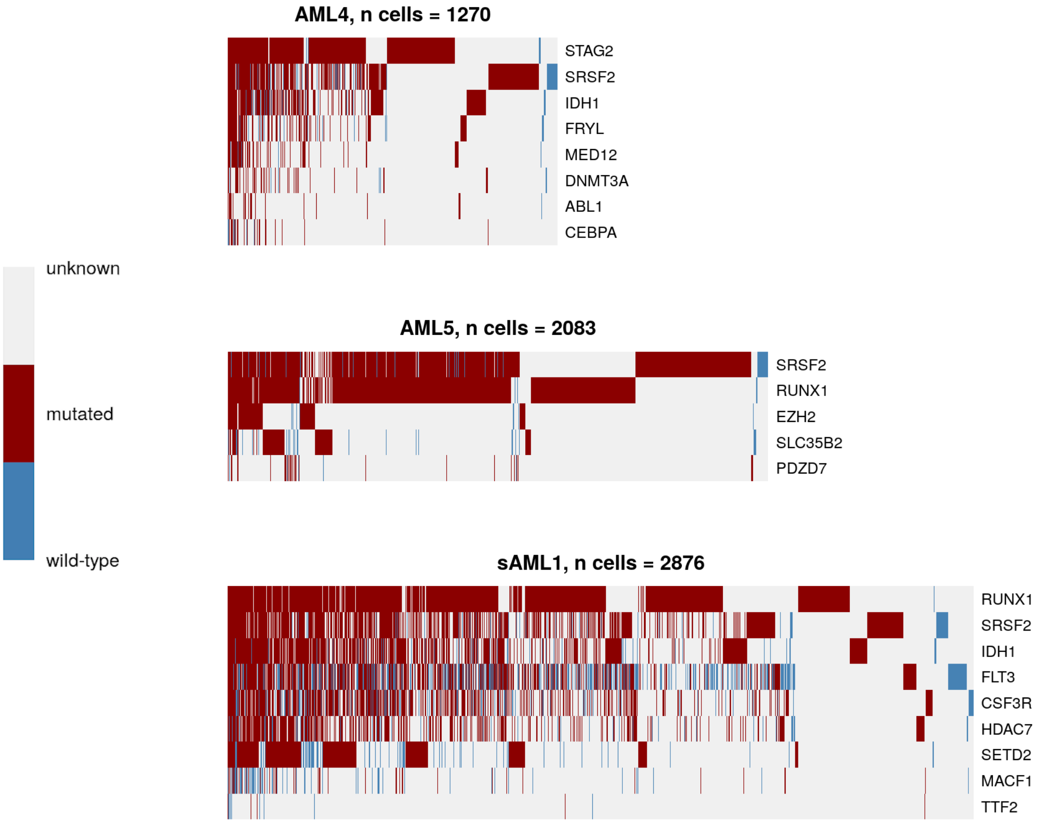
We then investigated the validity of genotyping imputation and mutation detection sensitivity by SCM-seq using read- or cell-based metrics, and compared results with VAFs calculated on WES data. First, as a read-based metric, we used the “single-cell Variant Allele Frequency” (scVAF) as described in Petti et al. We calculated scVAF as the ratio between total numbers of reads supporting a given variant and total numbers of non-mismatch reads covering the variant position, across the total number of cells that passed genotype imputation thresholds. Second, as a cell-based metric, we calculated the ‘mutant cell fraction’ (MCF) as the ratio between total numbers of cells imputed as mutated and total numbers of genotyped cells. Notably, both scVAF and MCF showed a positive correlation with WES VAF, which was stronger for MCF (scVAF: Spearman correlation r = 0.43, p = 0.045; MCF; Spearman correlation r = 0.6, p = 0.004) (Figure 31). Finally, we assessed the relationship of observed MCF to expected MCF, which corresponds to the theoretical number of mutated cells based on WES VAF (i.e., twice the WES VAF for heterozygous mutations). Again, we found a good correlation between the two (Spearman correlation r = 0.61, p = 0.002) (Figure 31).

Figure 31. Read- and cell-based metrics of genotype validity and mutation detection sensitivity by SCM-seq.
Spearman correlation highlights the relationship of scVAF with WES VAF (left), MCF with WES VAF (middle), and observed MCF with expected MCF (right).



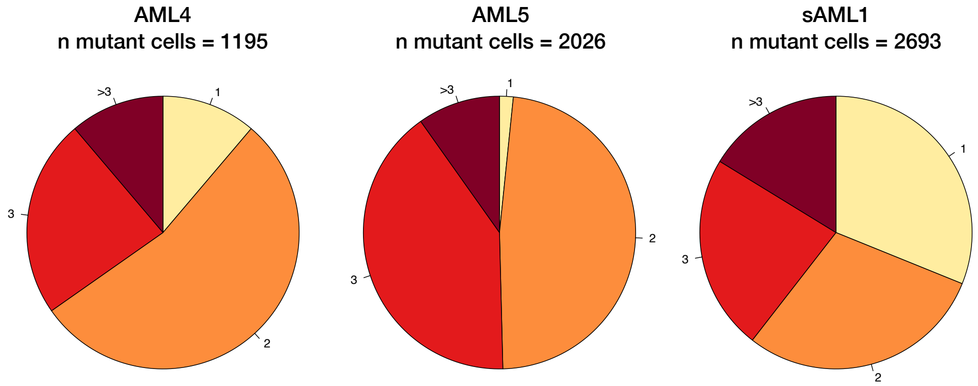
Analyses of mutation co-occurrence at single-cell level revealed a high degree of genetic complexity in all three samples (Figure 32).

Figure 32. Single-cell variant genotypes and co-occurrences.
The heatmaps show genotypes of each single cell (columns) for each gene variant (rows). Mutations are ordered by decreasing frequency.



Focusing on cells with at least one mutation, we could observe some cells with very high numbers of mutations (up to 7, 4 and 8 in AML4, AML5 and sAML1, respectively); more than half of cells carried at least two mutations, while around 25% showed three mutations and ≥10% more than three (Figure 33).

Figure 33. Distribution of numbers of mutations *per* cell.
The pie-charts show the proportion of cells bearing 1, 2, 3 or >3 mutations over the total of mutated cells.



Finally, we investigated whether genotyping performance is limited by expression-dropouts, given the strong correlation we observed between coverage of targeted variants and expression of the corresponding gene (Figure 20, paragraph 4.2.2). Notably, the mean log-expression (as derived from the 10x dataset) of cells with genotype information was systematically and significantly higher than that of cells without any genotype information for all variants (Figure 34, 35 and 36).

Figure 34. Relationship between variant expression and genotype outcome (sample AML4).
The violin plots show the mean log-normalized expression level of each targeted variant by genotype outcome. Two-sided Mann Whitney U test. ns = non significant, * = p<0.05, ** = p<0.01, *** = p<0.001.

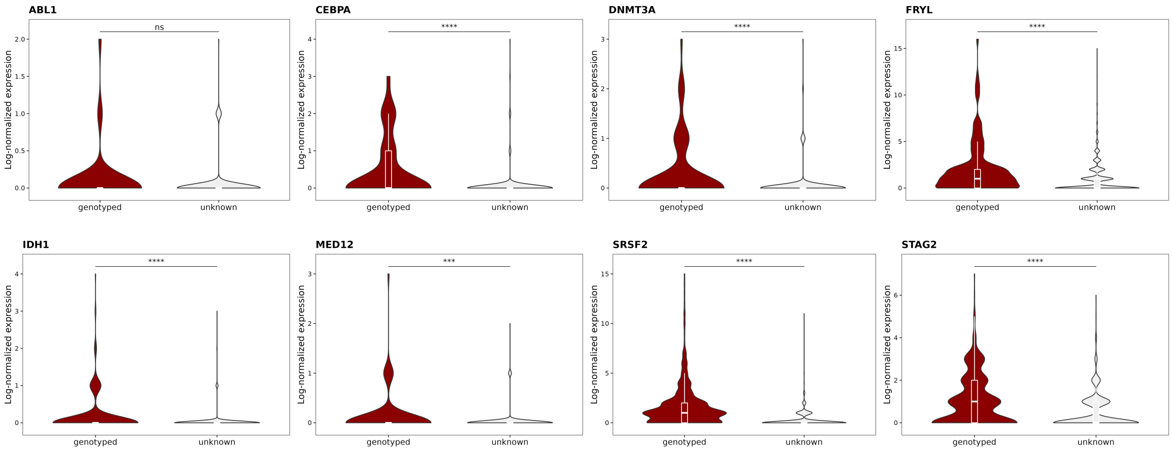


Figure 35. Relationship between variant expression and genotype outcome (sample AML5).
As in Figure 34.

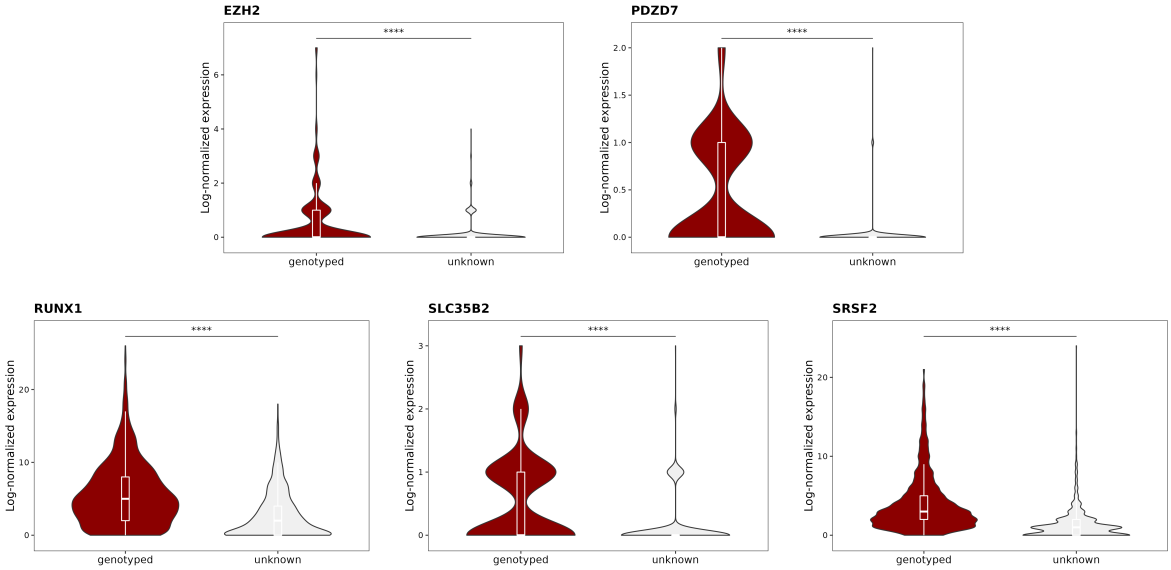
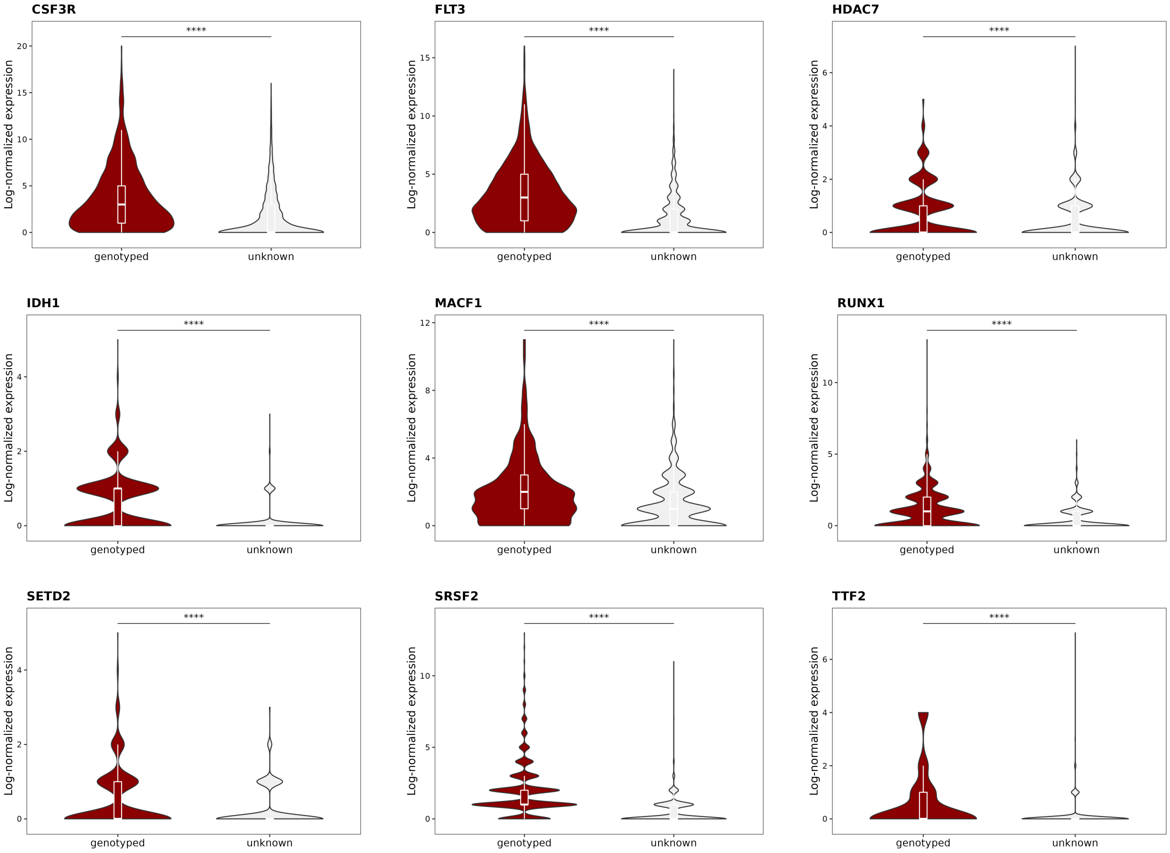


Figure 36. Relationship between variant expression and genotype outcome (sample sAML1).
As in Figure 34.



In summary, SCM-seq shows high mutation detection sensitivity both at read- and cell-level, preserving mutations co-occurrences and providing a mean to stratify groups of cells based on their genetic complexity (i.e., number and combination of mutations *per* cell). Single-cell genotyping performance was dependent on the expression level of the mutated gene, and limited by the sparseness of sequencing depth for a subset of cells. Although for any given variant wild-type cells were clearly underrepresented, for selected variants we were able to genotype both mutant and wild-type cells, which is the premise to investigate genotype-phenotype interactions.

…continued